Queueing Analysis of a Butterfly Network for Comparing Network Coding to Classical Routing

Parimal Parag, *Student Member, IEEE*, and Jean-Francois Chamberland, *Senior Member, IEEE*

***Abstract*—Network coding has gained significant attention in recent years as a means to improve throughput, especially in multicast scenarios. These capacity gains are achieved by combining packets algebraically at various points in the network, thereby alleviating local congestion at the nodes. The benefits of network coding are greatest when the network is heavily utilized or, equivalently, when the sources are saturated so that there is data to send at every scheduling opportunity. Yet, when a network supports delay-sensitive applications, traffic is often bursty and congestion becomes undesirable. The lighter loads typical of real-time traffic with variable sources tend to reduce the returns of network coding. This work seeks to identify the potential benefits of network coding in the context of delay-sensitive applications. As a secondary objective, this paper also studies the cost of establishing network coding in wireless environments. For a network topology to be suitable for coding, links need to possess a proper structure. The cost of establishing this structure may require excessive radio resources in terms of bandwidth and transmit power. Bursty traffic together with structural cost tend to decrease the potential benefits of network coding. This paper describes how, for real-time applications over wireless networks, there exist network topologies for which it may be best not to establish a network structure tailored to network coding.**

***Index Terms*—Butterfly network, communication system, delay, quality of service (QoS), network coding, routing, tail asymptotics, tandem queues, wireless networks, wireless systems.**

I. Introduction

NETWORK coding is a novel paradigm that has received much attention in the literature recently [1]–[5]. It has the potential to improve the throughput and robustness of future communication networks. These performance gains are achieved by relaxing the restriction that data belonging to different information flows should remain separated. Indeed, network coding is a transmission strategy where packets are combined algebraically at intermediate nodes in the network. It can be viewed as an extension of traditional routing. In certain circumstances, network coding helps improve overall throughput; and it is known to achieve the min-cut flow in multicast scenarios [6].

The research enthusiasm generated by network coding can be explained, partly, by the ever expanding demand for Internet access and fast connectivity. Not only is network coding mathematically elegant, but it also seeks to improve network performance at a time when the number of data applications is rising furiously. The growing demand for network connectivity is felt both at the core of the Internet and at its periphery, where wireless systems are increasingly employed to provide flexibility to mobile users. One class of data connections that is rapidly gaining prominence on the Internet is the traffic generated by real-time applications. Delay-sensitive services including voice over Internet protocol (VoIP), video conferencing, gaming, and electronic commerce are now commonly used by vanguardists on both wired and wireless devices. Future communication infrastructures are expected to carry much larger volumes of data with varying quality of service (QoS) requirements. As such, this paper seeks to provide preliminary answers to two important questions related to delay-sensitive traffic and the efficient utilization of network coding.

First, are the potential benefits of network coding as substantial in the context of delay-sensitive applications? It seems intuitively clear that the gains of network coding are maximal when the links in the network are fully utilized. However, the bursty nature of many data sources and the service quality required of most real-time applications may force a network to operate much below its maximum throughput. This phenomenon is captured by the concept of effective bandwidth, which identifies the data rate needed by a source to fulfill its service requirement [7], [8]. In general, the effective bandwidth of a source can be much larger than the average throughput it produces. The bursty traffic generated by delay-sensitive applications combined with the gains associated with statistical multiplexing act to decrease the benefits of network coding. Therefore, it is not clear how much we gain by applying network coding in a communication system subject to QoS constraints. In this paper, we provide quantitative results on the benefits of network coding for a simple butterfly network in the context of delay-sensitive applications.

Another pertinent observation about network coding is that it often requires a structured network topology. Coding benefits are optimum when the data rates of the various links are integer multiples of one another. In a wireless environment, physical-layer resources can be allocated progressively to the different nodes. To maximize the coding gain, these resources

Manuscript received June 25, 2008; revised September 04, 2009. Current version published March 17, 2010. This work was supported in part by the National Science Foundation (NSF) under Grants 0747363 and 0830696. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Toronto, ON, Canada, July 2008.

The authors are with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: parimal@neo.tamu.edu; chmbrlnd@tamu.edu).

Communicated by A. Nosratinia, Associate Editor for Communication Networks.

Color versions of Figures 2–7 in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2010.2040862

must be assigned to create a suitable topology. While this enables efficient coding, there may be a nonnegligible cost associated with creating such a structure. In other words, in a wireless environment, the performance of a system with network coding should be compared to the operation of the equivalent classic-routing system, with physical resources allocated optimally in both cases. This leads us naturally to the second question we seek to address. When is it relevant to create a topology suitable for network coding in a wireless environment?

These two important questions are not only related through the rising popularity of real-time applications and network coding, but also by their answers necessitating the development of analogous mathematical tools. This similarity motivates our joint treatment of these related topics. More specifically, we investigate the impact of network coding on the queueing behavior of wireless communication systems.

We consider a simple scenario where two varying rate sources communicate to multiple destinations through the notorious butterfly network, shown in Fig. 1. Every node is equipped with a data buffer where packets are stored prior to transmission. We analyze the performance of this system, and compute its achievable rate region when the network operates under stringent service constraints. Due to the time-varying nature of typical arrival and service processes, it is difficult to provide deterministic delay guarantees for such systems. Accordingly, we adopt a popular statistical QoS criterion that captures the asymptotic decay rate in buffer occupancy

$$\theta = -\lim_{x\to\infty} \frac{\ln \Pr\{L > x\}}{x} \tag{1}$$

where L has the equilibrium distribution of the buffer at the transmitter. Parameter θ reflects the perceived quality of the corresponding communication link: a larger θ implies a lower probability of violating a queue-length restriction and a tighter QoS constraint. This performance criterion is closely tied to large-deviations theory, and it forms a basis for the concept of effective bandwidth which has been studied extensively in the past [9]–[14]. Given a specific arrival process, the effective bandwidth characterizes the minimum constant service rate required for a communication system to meet its QoS requirements. Parameter θ is also related to the dual concept of effective capacity popularized by Guerin *et al.* [15], de Veciana *et al.* [16] and Wu and Negi [17]. Unlike wired networks, wireless links frequently feature time-varying service rates [18]. The effective capacity characterizes the maximum constant arrival rate that a wireless system can support, given a minimum buffer occupancy decay rate θ_0. When the decay rate θ_0 approaches zero, the effective capacity converges to the maximum throughput supported by the wireless channel.

To study the performance of a communication system subject to a buffer occupancy constraint akin to (1), we need to characterize the queueing performance of the network. In the mathematical framework under consideration, independent sources sharing a same link can be studied separately. This is one of the appealing properties of an analysis based on large deviations. The main challenge, as we will see, is to characterize the performance of the tandem network shown in Fig. 2. This network consists of two successive nodes where the output of the first node acts as an input to the second queue.

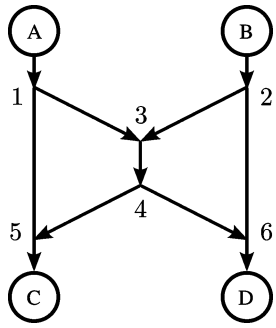

Fig. 1. Directed butterfly network.

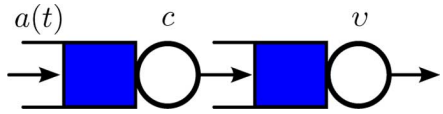

Fig. 2. Network with tandem queues.

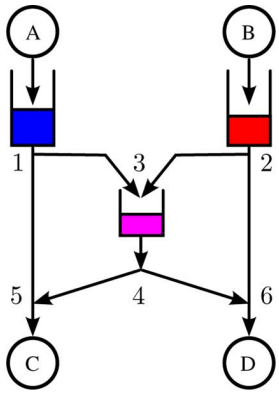

Fig. 3. Butterfly network of interest with corresponding buffers.

A. Contributions

Fig. 3 shows the butterfly network we want to study. For a multicast scenario, where stochastic sources A and B wish to communicate to destinations C and D, node 3 has the opportunity to employ packet combining. We consider two distinct versions of this simple butterfly network. First, we analyze a noise-limited network with constant and identical link capacities. This configuration is suitable for network coding and is the basis for our initial queueing comparison. Then, we examine a wireless network under a broadcast paradigm. In the latter scenario, we assume that physical resources can be allocated freely among the various nodes to create nonidentical links, thereby enabling optimal operation within each configuration.

To compare the queueing performance of network coding versus classic routing, we characterize the achievable rate regions for both these cases under a QoS guarantee on the tail-asymptotics of the buffer-content distributions. Not too surprisingly, network coding outperforms classic routing for a network with identical link capacities. Although statistical multiplexing had the potential to offset some of the coding gain,

TABLE I
COMPARING OUR WORK WITH THE LITERATURE

	[19]	[20]	[21]	[22]	[23]	[24]	[25]	[26]	[27]	[28]	[29]
Scheme	×U	B	U	M	B	×U	×U	×U	×U	×U	M, ×U
Arrivals	S	F	B	B	B	S	P	P	P	S	S
Service	C	S	B+V	IID	IID	V	C	C	C	C	V
Policy	WC	–	FCFS	FCFS	FCFS	FCFS	O	FCFS	O	FCFS	–
Aspect	D	C	D	C,D	D,Q	Q	Q	C	Q	Q	Q
QoS	PD	–	MD	S,D	MQ	S	S	MD	S	S	S
Network	L	D	P	D	D	A	2W	2W	2W	T	A

classic routing remains a distant second to network coding for all QoS requirements. More interesting results come from the analysis of wireless butterfly network. Combining packets at an intermediate wireless relay does not necessarily yield performance gains, and may even be detrimental in some cases. This behavior depends on the topology of the butterfly network and the physical locations of the nodes. This peculiarity follows from the fact that network coding needs symmetric links between the sources and their destinations for maximal coding gains to be realized. If the link capacities are not identical, then packet combining entails delay and inefficiencies. These results are detailed below, and they may be employed to provide guidelines on when to form a network suitable for combining packets algebraically at intermediate nodes. Analysis is limited to the simple butterfly network under consideration. However, the featured approach can be generalized to topologies with constant service rates, using the same queueing methods employed in this paper. In fact, our methodology for networks with two queues in tandem can be applied to several queues in cascade. Still, an exact analysis would get increasingly complex for large networks and these techniques may not lead to tractable expressions. In more complex topologies, it may be necessary to use approximation methods, which is a different topic altogether.

We would like to emphasize that our contribution is twofold. We provide a quantitative analysis of the tradeoff between the cost of establishing a structure suitable for network coding and the ensuing returns associated with algebraic packet combining. Second, we offer an alternate proof for deriving the tail-exponent of the second buffer in a tandem queue.

B. Relevant Work

There have been many recent contributions on queueing behavior when network coding is employed at the transmitter. We compare and contrast our work with the available literature below and summarize this discussion in Table I.

The random linear combining of data packets is considered in [19]–[23]. The authors focus on coding delay in [20], whereas decoding delay is studied in the remaining contributions. In [19], QoS is defined in terms of packet drop probability, and multiple flows are considered over an arbitrary network. In [20], the authors compare the performance of network coding versus scheduling for broadcast and multiple unicast scenarios; their work is based on average delay performance. In [21] and [22], the authors explore the throughput-delay tradeoff with and without network coding. In [21], the coding scheme adapts to the underlying traffic conditions. Stability and delay performance of a multicast erasure channel with stochastic arrivals are studied in [22]. In [23], the authors propose a coding and queue management algorithm. Note that the random linear combining of packet transmissions is in effect a coding scheme which trades off delay and throughput over a single flow. In our work, we study the achievable rate region over a butterfly network when one employs either network coding or classic routing at the intermediate node, a distinct framework altogether.

Previous contributions differ from the framework presented below in many more respects. First, we are comparing network coding to routing at a given node in a simple butterfly network. We consider a multicast scenario with two transmitters and two receivers. Second, we disregard coding/decoding delay and focus on the equilibrium queue-length buildup due to stochastic arrivals, limited service rate and feedback from the receiver. Third, we use preselected codes with fixed rates for reliable communication over the links. Furthermore, our QoS constraint on tail-asymptotics deviates significantly from average delay constraints, thereby offering better insights for delay-sensitive traffic. This methodology can be utilized to provide different QoS guarantees to flows with various requirements. In addition, we take into consideration the stochastic nature of arrivals, which may reduce the gains of network coding due to statistical multiplexing.

A two-way relay channel is considered in [24]–[27]. Note that two-way relays can be modeled as a butterfly network where sources are also destinations. However, this is not an equivalence relation. In the two-way relay model, direct links are absent and side information about the received data is available at the destination, an advantage that is not present in our network of interest. In [25], the authors characterize the end-to-end rate regions for MAC-XOR and PHY-XOR operations for two-way relay channel. They also present an opportunistic scheduling algorithm to show that the system can be stabilized for any bit-arrival rate pair within the Shannon rate region. In [26], the authors study the energy-delay tradeoff when network coding is used at the relay node. The energy is measured in terms of code rates and channel conditions. In [27], authors show that, for asymmetric traffic, one needs to perform time sharing between traditional slotted multihopping and network coding. In [28], the authors characterize the stability region for bursty traffic at multiple sources with and without network coding in the wireless network. In [24], the authors propose a framework to develop adaptive joint network coding and scheduling schemes. The authors show that, for asymmetric traffic, scheduling and network coding need to be combined to maximize supportable throughputs over this network. In [29], the authors propose a dynamic routing-scheduling-coding strategy based on queue-state information for serving multiple unicast sessions. The authors study queues over a butterfly network for a discrete-time packet

model, and they use fluid approximations to show the stability of the system for their proposed algorithm.

Again, we emphasize that the wireless butterfly network considered in our article is not equivalent to two-way relaying. We assume adaptive network coding in the sense that every time there is an opportunity to do network coding, the relay node combines packets algebraically. The servicing policy adopted throughout is taken to be first-come first-served. A fluid assumption allows us to bypass the scheduling problem. It is clear that, given a suitable network topology, the flexibility to switch between network coding and classical routing, when the need arises, is better than always using classical routing. What sets our work apart is accounting for the cost associated with making a network suitable for network coding. We study whether coding gains can offset the cost of facilitating a proper network structure. The rate regions achieved in [25] and [27]–[29] correspond to the cases where the queue is stable for stochastic arrivals. There is no explicit guarantee on the buffer distribution. In our work, we find the achievable rate region under a tail-asymptote requirement on the decay rate of the queue distribution. Therefore, the conclusions by the authors in the aforementioned papers would not necessarily apply in a framework akin to ours.

C. Organization

The remainder of this paper is organized as follows. We introduce the system model in Section II. We list pertinent results on the performance of tandem queues in Section III. These tools are used to analyze symmetric butterfly networks in Section IV. Wireless butterfly networks are studied in Section V. Key queueing results about the second buffer of a tandem network are established in Section VI. A complete characterization of the departure process at the output of a single queue with constant service and Markov-modulated arrivals is presented in Section VI-A. An expression for the equilibrium distribution of the buffer-occupancy is derived in Section VI-B. This enables us to obtain the corresponding asymptotic decay rate of buffer-occupancy in Section VII. In Section VIII, we determine the maximum achievable rate for the second buffer in the QoS constrained tandem network under consideration. Finally, we conclude with some relevant remarks and future directions in Section IX.

II. Problem Statement

We study a communication system where two independent users wish to send their messages to two common destinations over a butterfly network, as shown in Fig. 3. A multicast scenario is considered where independent sources A and B store their respective information in buffers at nodes 1 and 2, and must transmit their data to both destinations C and D. To facilitate this process, node 1 sends its packets to nodes 3 and 5. Similarly, node 2 forwards its packets to nodes 3 and 6. Node 3 can take two courses of action; either it stores the received packets from the sources in a queue and then forwards them individually to node 4, or it combines the packets algebraically before transmitting the data.

The first setting will be called the *classic routing* case. In this scenario, node 4 duplicates the received packets from node 3 and forwards copies to nodes 5 and 6. These destination nodes disregard redundant information (they could potentially take advantage of redundancy to improve the reliability of previously received messages, but this is beyond the scope of this article) and retain new data. For the second scenario, we consider the network coding scheme where node 3 adds the two streams of packets over *GF(2)* and relays the coded messages to node 4 [6]. The latter duplicates the received packets and transmits them to nodes 5 and 6. Node 5 can resolve the information received from node 2 by adding the packets obtained from node 1 to the corresponding packets received from node 4. In a similar fashion, node 6 can decode the information originating from node 1 by adding packets from node 2 to the corresponding packets from node 4. Service quality is captured by a global QoS constraint θ_0 on the system. That is, the asymptotic decay rate of buffer occupancy must be greater than or equal to θ_0 for all the queues in the system.

For the sake of analysis, we assume that packets are infinitely divisible and hence the arrival and service processes are fluid in nature. Thus, it becomes possible to define instantaneous arrival and service rates. Under this assumption, every node in the network is equipped with a single fluid queue served by an individual transmitter. We also take the buffers in the system to be arbitrary large. A similar approach applies to the finite buffer case, albeit with additional boundary conditions on the buffer occupancy.

A. Source Model

Many real-time traffic sources can be accurately represented by *on-off* models [30]. This motivates our assumption of arrivals being two-state Markov-modulated fluid processes. In addition, there is a vast amount of literature available on the queueing behaviors of Markov-modulated fluid processes for wire-line networks [8], [9], [31]. We postulate that sources A and B are independent, and that they both satisfy the following assumption.

Assumption 1: During an on period, the source emits packets at a constant peak rate into its buffer; it remains idle otherwise. Moreover, the on and off times are independent and exponentially distributed.

Mathematically, the sources are defined through their underlying Markov chains. Let $\{I_1(t) \in \{0,1\} : t \geq 0\}$ and $\{I_2(t) \in \{0,1\} : t \geq 0\}$ be independent two-state continuous-time Markov chains (CTMC) modulating *on-off* sources A and B, respectively. State zero represents the *off* state and state one denotes the *on* state. Suppose that the peak rate for the source at node $i \in \{1,2\}$ is taken as a_i. With a slight abuse of notation, we can write the arrival process at buffer i as

$$a(I_i(t)) = a_i \mathbf{1}_{\{I_i(t)=1\}}, \qquad i \in \{1,2\}$$

where $\mathbf{1}_{\{\cdot\}}$ represents the standard set indicator function. We denote the mean *off* and *on* times by λ_i^{-1} and μ_i^{-1}, respectively. The generator matrix for the modulating two-state Markov process can then be written as

$$Q_i = \begin{bmatrix} -\lambda_i & \lambda_i \\ \mu_i & -\mu_i \end{bmatrix}, \qquad i \in \{1,2\}.$$

B. Queueing Model

We denote the capacity of link i–j by c_{ij}. This capacity effectively limits the offered service rate at node i for transmission to node j. In particular, if there exists a link between nodes i and j and the buffer associated with node i is nonempty, then node i can transmit to node j at a maximum rate c_{ij}. For simplicity, we assume that $c_{34} = c_{45} = c_{46} = c_3$. The offered service rates on links 4–5 and 4–6 are then equal to the maximum arrival rate at node 4. As such, node 4 does not need to store data. It only facilitates the duplication and the forwarding of its received packets to nodes 5 and 6. In other words, the buffer associated with node 4 is always empty.

Node 1 sends the same information to both nodes 3 and 5, and therefore retains data in its buffer until both receiving nodes have acquired the corresponding packet. Accordingly, the service rate at node 1 is $c_1 = \min\{c_{13}, c_{15}\}$. Similarly, the service offered to the buffer at node 2 is $c_2 = \min\{c_{23}, c_{26}\}$. Altogether, nodes 1, 2, and 3 transmit packets at rates c_1, c_2, and c_3, respectively, whenever their own buffers are nonempty. Observe that, by construction, congestion can only occur at these three nodes. We can therefore safely assume that there are no queues at the other nodes. We have depicted the fluid model of interest in Fig. 3 for the butterfly network under consideration. We represent the fluid level in the buffer at node i and time t by $L_i(t)$. The stochastic evolution of $L_i(t)$ depends on whether one opts for network coding or classic routing.

C. Network Coding

For network coding, packets originating from links 1–3 and 2–3 are combined algebraically over *GF(2)* and then stored in the buffer at node 3. From a fluid perspective, this is equivalent to both flows entering buffer 3 oblivious of each other. Buffer 3 can be serviced at a maximum rate c_3. However, to prevent decoding delays at the destinations, the service rates offered at nodes 1, 2, and 3 are made equal to $\hat{c}_1 = \min\{c_1, c_3\}, \hat{c}_2 = \min\{c_2, c_3\}$ and $\check{c}_3 = \max\{\hat{c}_1, \hat{c}_2\}$. In this scenario, there is no congestion at node 3 and hence $L_3(t) = 0$ for all times t. Furthermore, for the nontrivial case where $a_i > \hat{c}_i$, we can write the stochastic evolution of $L_i(t)$ as

$$\frac{d}{dt}L_i(t) = (a(I_i(t)) - \hat{c}_i)\mathbf{1}_{\{I_i(t)=1\}} - \hat{c}_i\mathbf{1}_{\{I_i(t)=0, L_i(t)>0\}} \quad (2)$$

where $i \in \{1,2\}$. When the buffer at node 1 is nonempty, the net rate of fluid input is $a(I_1(t)) - \hat{c}_1$; this is called the drift rate. For $i \in \{1,2\}$, we can define a drift matrix $D_i = \mathrm{diag}(-\hat{c}_i, a_i - \hat{c}_i)$ for the buffer at node i, whose diagonal entries are the drift rates corresponding to the state of the arrival process. In matrix form, we have

$$D_i = \begin{bmatrix} -\hat{c}_i & 0 \\ 0 & a_i - \hat{c}_i \end{bmatrix}, \qquad i \in \{1,2\}.$$

D. Classic Routing

For the case of classic routing, the service rate offered at node $i \in \{1,2\}$ is c_i. The evolution of $L_i(t)$ for the nontrivial case of $a_i > c_i$ is then governed by

$$\frac{d}{dt}L_i(t) = (a(I_i(t)) - c_i)\mathbf{1}_{\{I_i(t)=1\}} - c_i\mathbf{1}_{\{I_i(t)=0, L_i(t)>0\}} \quad (3)$$

where $i \in \{1,2\}$. It will be shown in the later sections that the departure process at the output of buffer i is a two-state *on-off* process modulated by a countable-state Markov process $K_i(t)$. The departure process at node i can be represented as

$$c(K_i(t)) = c_i\mathbf{1}_{\{K_i(t)\neq 0\}}, \qquad i \in \{1,2\}.$$

The buffer at node 3 is fed by the aggregation of these two independent arrival processes, and it is serviced at a constant rate c_3. To initiate the analysis of this more complicated scenario, we study the simple case where the resources at buffer 3 are split between the flows of sources A and B. Consider two parallel buffers at node 3 with positive constant service rates ν and $c_3 - \nu$, respectively. We assume that the flow from node 1 goes to the first parallel buffer; and the flow from node 2, to the second one. The aggregate fluid in both the buffers will be greater than or equal to the fluid level of a single-buffer system with incoming data from nodes 1 and 2 and service rate c_3. Thus, the decay rate of buffer occupancy for a system with a single buffer at node 3 must be no less than the exponential decay rate of the corresponding system with parallel buffers and partitioned service. Yet, it can be shown that these asymptotic values are equal for independent flows and optimal splitting rate [11], [13], [14], [32]. The two independent flows can therefore be decoupled and studied separately. If the shared buffer is constrained by a requirement θ_0 on the decay rate of buffer occupancy, the queues in the decoupled system are constrained by the same parameter θ_0 as well. We denote by $L_{i3}(t)$ the fluid level in the buffer of the decoupled system holding data from node i. We can write the stochastic evolution of $L_{i3}(t)$ for the nontrivial case of $c_i > \max\{\nu, c_3 - \nu\}$, as

$$\begin{aligned}\frac{d}{dt}L_{13}(t) &= (c(K_1(t)) - \nu)\mathbf{1}_{\{K_1(t)\neq 0\}} - \nu\mathbf{1}_{\{K_1(t)=0, L_{13}(t)>0\}} \\ \frac{d}{dt}L_{23}(t) &= (c(K_2(t)) - c_3 + \nu)\mathbf{1}_{\{K_2(t)\neq 0\}} \\ &\quad - (c_3 - \nu)\mathbf{1}_{\{K_2(t)=0, L_{23}(t)>0\}}.\end{aligned}$$

III. Key Results

In this section, we list the mathematical results needed to compute the achievable rate regions for data multicast through the butterfly network, under specific QoS requirements. Let $\ell_1(t)$ be the amount of fluid at time t in a queue being fed by an *on-off* source satisfying Assumption 1, and serviced at a constant rate c. Let a denote the arrival rate into this buffer when the source is *on*. The mean *off* and *on* times of the source are denoted by λ^{-1} and μ^{-1}, respectively. Furthermore, the output of this queue (also called departure process) is fed into another arbitrary large reservoir. This second queue is being serviced at a constant rate v. The amount of fluid in the latter buffer at time t is denoted by $\ell_2(t)$.

We wish to find the maximum peak rate a, such that the QoS criterion of (1) is no less than θ_0 for both queues. Let θ_1 and θ_2 be the asymptotic buffer decay rates governing the first and second queues in the tandem network, i.e.,

$$\theta_j = -\lim_{x\to\infty} \frac{\ln \Pr\{\ell_j > x\}}{x} \quad (4)$$

where ℓ_j is the steady-state queue-length of buffer j. More specifically, we wish to identify set $\mathcal{A}$ (as a function of θ_0, c, and υ) defined by

$$\begin{aligned}\mathcal{A}(\theta_0, c, \upsilon) &= \{a \in \mathbb{R}^+ \colon \min\{\theta_1, \theta_2\} \geq \theta_0\} \\ &= \{a \in \mathbb{R}^+ \colon \theta_1 \geq \theta_0\} \cap \{a \in \mathbb{R}^+ \colon \theta_2 \geq \theta_0\}. \quad (5)\end{aligned}$$

If $a \leq c$, the first buffer always remains empty and the behavior of the tandem queue reduces to that of a system with a single queue. We therefore focus on the nontrivial case where $a > c$. Under this condition, Theorem 1 in Section VI asserts that if $a\lambda/(\lambda+\mu) < c < a$ then

$$\theta_1 = \frac{\mu}{a-c} - \frac{\lambda}{c}. \quad (6)$$

The lower bound on c ensures stability of the queue. This formula implicitly determines the maximum peak rate a such that a QoS constraint of θ_0 is satisfied at buffer 1. We define

$$\begin{aligned}\mathcal{A}_1(\theta_0, c) &= \{a \in \mathbb{R}^+ \colon \theta_1 \geq \theta_0\} \\ &= \{a \in \mathbb{R}^+ \colon a \leq \bar{a}_1(\theta_0, c)\} \quad (7)\end{aligned}$$

where we define $\bar{a}_1(\theta, c) = c + c\mu/(\lambda + c\theta)$. The second buffer always remains empty if $\upsilon \geq c$. Thus, we consider the situation where $\upsilon < c$. We identify the range of allowable rates such that buffer 2 satisfies QoS constraint θ_0 (see Section VIII) as

$$\begin{aligned}\mathcal{A}_2(\theta_0, c, \upsilon) &= \{a \in \mathbb{R}^+ \colon \theta_2 \geq \theta_0\} \\ &= \{a \in \mathbb{R}^+ \colon a \leq \bar{a}_2(\theta_0, c, \upsilon)\} \quad (8)\end{aligned}$$

where the function $\bar{a}_2(\theta, c, \upsilon)$ is given by

$$\bar{a}_2(\theta, c, \upsilon) = \begin{cases} \bar{a}_1(\theta, \upsilon), & 0 < \upsilon \leq \upsilon^* \\ \bar{a}_3(\theta, c, \upsilon), & \upsilon^* < \upsilon < c. \end{cases}$$

Here, $\bar{a}_1(\theta, \upsilon)$ is as defined above and $\bar{a}_3(\theta, c, \upsilon)$ is given by the expression

$$\bar{a}_3(\theta, c, \upsilon) = c + \frac{c\mu}{\lambda} \left(\frac{-1 + \sqrt{1 - \left((c-\upsilon)\frac{\theta}{\mu} - 1\right)\left((c-\upsilon)\frac{\theta}{\lambda} - 1\right)}}{(c-\upsilon)\frac{\theta}{\lambda} - 1} \right)^2$$

with parameter υ^* determined implicitly by

$$\frac{c}{\upsilon^*} - 1 = \frac{\theta \upsilon^* \mu}{\lambda\mu + (\lambda + \theta\upsilon^*)^2}.$$

Collecting these results, we obtain $\mathcal{A}(\theta_0, c, \upsilon) = \{a \in \mathbb{R}^+ \colon a \leq \min\{\bar{a}_1(\theta_0, c), \bar{a}_2(\theta_0, c, \upsilon)\}\}$. That is, a is admissible if and only if $a \in \mathcal{A}(\theta_0, c, \upsilon)$ with

$$a \leq \begin{cases} \bar{a}_1(\theta_0, \upsilon), & 0 < \upsilon \leq \upsilon^* \\ \min\{\bar{a}_1(\theta_0, c), \bar{a}_3(\theta_0, c, \upsilon)\}, & \upsilon^* < \upsilon < c. \end{cases} \quad (9)$$

An intuitive explanation for the behavior of this tandem network is as follows. Buildups in the system can occur at buffers 1 and 2. When the service rate of the second buffer is small ($\upsilon \leq \upsilon^*$), the behavior of the system is dominated solely by the action of the second queue. On the other hand, when $\upsilon > \upsilon^*$, deviations are caused by the combined behavior of the two queues with large buildups occurring in the first queue when $\bar{a}_1(\theta_0, c) \geq \bar{a}_3(\theta_0, c, \upsilon)$, and in the second queue otherwise. The more complicated expression corresponding to this latter case follows from the fact that, altered by the first buffer, the structure of the arrival process feeding the second buffer is more intricate. We plot the boundary of $\mathcal{A}(\theta_0, c, \upsilon)$ with increasing $\upsilon \in (0, c)$ in Fig. 4 for various values of QoS constraint θ_0. It is clear from the figure that the achievable rate region $\mathcal{A}(\theta_0, c, \nu)$ shrinks with QoS constraint θ_0, as one would expect. The system parameters used in this example are $\lambda^{-1} = 0.65$ s, $\mu^{-1} = 0.352$ s, and $c = 1$ Mb/s.

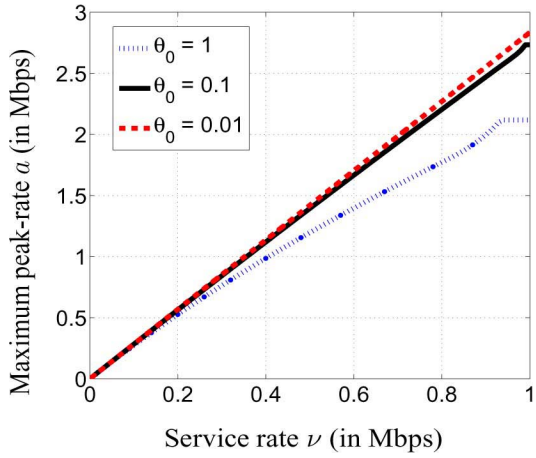

Fig. 4. Maximum supportable peak rate of the on-off source $a(\theta_0, \nu) = \sup \mathcal{A}(\theta_0, c, \nu)$ versus ν under various values of QoS requirement θ_0 for a tandem queue. The first queue is serviced by a fixed constant service rate $c = 1$ Mb/s and the second queue is serviced by a rate $\nu \in [0, c]$ Mb/s. Parameter θ_0 denotes the target asymptotic exponential decay rate of the tail buffer occupancy.

Before we apply these results to compute achievable rate regions for the butterfly network, we point out that there are at least three different ways of obtaining the tail exponents of tandem queues. First, there is the transform method studied in [33]–[35]. Under this approach, the authors find the Laplace–Stieltjes transform of the desired distributions [34], [35]; still, these transforms are not straightforward to invert. A much stronger result, the joint distribution for tandem queues, is obtained in [33]. However, for our purposes it suffices to know the marginal content distributions of the two buffers. Furthermore, the marginal distribution for the second buffer in [33] is not expressed in its convenient reduced form.

A second approach would be to use sample-path large deviations as in the seminal paper by Chang–Zajic [11], [13], [14]. One can use the Lindley recursion and inverses of counting processes (Galois connections) to construct a discrete-time embedded process that would be closely related to the continuous-time process at hand. Thus, one can employ this methodology to study stationary random variables such as buffer content. However, to use these results [11], [13], [14], one needs to verify the general mixing conditions for the sample-path large deviations of the departure process. These conditions are highly technical

and, to show they are satisfied, one requires expertise in probability theory and filtrations; this would lead to a more contrived exposition.

The third approach, which we adhere to in this paper, can also be found in previous literature [36]–[38]. One can study two queues separately by first characterizing the stationary departure process from the first queue, and using it as an arrival process for the second queue. This gives us an explicit distribution for the marginal content of the second buffer in a tandem network. That is, we provide an alternate proof for the tail-exponent of the second buffer in a tandem queue (though it is a specialized result, it gives us the desired form) utilizing the Anick–Mitra–Sondhi approach [36]–[38] together with a characterization of the departure process by Aalto [39], [40].

IV. Achievable Rate Regions

The results listed above can be employed to identify achievable rate regions for the butterfly network under consideration. With a slight abuse of notation, we let θ_i be the asymptotic decay rate of buffer-occupancy for the queue at nodes $i \in \{1,2,3\}$ in Fig. 3, for a pair of peak-arrival rates (a_1, a_2) at sources A and B. We need to find the set of all two-tuples (a_1, a_2) such that the global QoS constraint θ_0 is satisfied, i.e.,

$$\mathcal{R} = \left\{(a_1, a_2) \in \mathbb{R}^+ \times \mathbb{R}^+ \colon \min\{\theta_1, \theta_2, \theta_3\} \geq \theta_0\right\}.$$

1) Network Coding: As mentioned earlier, the effective service rates offered at nodes 1, 2, and 3 are $\hat{c}_1 = \min\{c_1, c_3\}$, $\hat{c}_2 = \min\{c_2, c_3\}$, and $\check{c}_3 = \max\{\hat{c}_1, \hat{c}_2\}$. This prevents undue decoding delays at the destinations. Using the notation of the previous section, we can write the achievable rate region $\mathcal{R}$ for this system as

$$\mathcal{R}_{\text{nc}} = \mathcal{A}_1^{(1)}(\theta_0, \hat{c}_1) \times \mathcal{A}_1^{(2)}(\theta_0, \hat{c}_2)$$

where $\mathcal{A}_1$ is the set defined in (7).

2) Classic Routing: For classic routing, consider two parallel buffers at node 3 with constant service rates ν and $c_3 - \nu$, respectively. Assume that the flow from node 1 goes to the first buffer; and the flow from node 2, to the second one. Again, we emphasize that the aggregate fluid in both the buffers will be greater than or equal to the level of fluid in a single-buffer system with combined arrivals from nodes 1 and 2 and serviced at rate c_3. Using the aforementioned splitting property, the two independent flows can be decoupled and studied separately. If the shared buffer is constrained by a QoS requirement θ_0, the queues in the decoupled system are subject to the same criterion. For a fixed $0 < \nu < c_3$, there exists a unique peak rate pair (a_1, a_2) such that the QoS constraint θ_0 is satisfied by the system if the achievable rate region is $\mathcal{A}(\theta_0, c_1, \nu) \times \mathcal{A}(\theta_0, c_2, c_3 - \nu) = [0, a_1] \times [0, a_2]$. Accordingly, the achievable rate region is equal to the union of the regions corresponding to all the possible values of ν. That is

$$\mathcal{R}_{\text{cr}} = \bigcup_{0 \leq \nu \leq c_3} \mathcal{A}^{(1)}(\theta_0, c_1, \nu) \times \mathcal{A}^{(2)}(\theta_0, c_2, c_3 - \nu) \quad (10)$$

where $\mathcal{A}$ is the achievable rate region of (5).

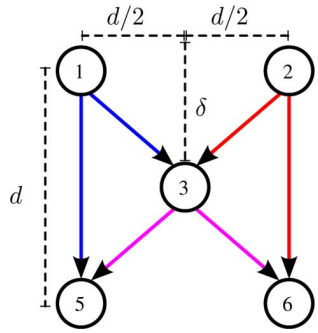

Fig. 5. Wireless butterfly network with two sources, two destinations and a relay node.

V. Wireless Butterfly Network

In this section, we study a wireless butterfly network under a broadcast paradigm. We assume that the system operates in frequency division multiplexing (FDM) mode. We should point out that FDM operation is not necessarily optimal in terms of achievable rates or delays. More complex schemes provide future avenues of research. It is not clear though that a similar analytical framework can be used while considering alternative multiple-access schemes.

Consider the multicast scenario where two sources wish to communicate with two destinations. An additional node that acts as a relay is present to facilitate communication over the network. All the nodes have an identical power budget P, and the total spectral bandwidth available to the system is limited. The source nodes produce independent *on-off* traffic, as in our previous setting. The total available spectral bandwidth W is divided to make three noninterfering frequency bands.

Node 1 broadcasts its messages to nodes 3 and 5, and node 2 does the same to nodes 3 and 6 (see Fig. 5). Node 3 sends its messages to nodes 5 and 6 simultaneously. The packets transmitted by node 3 can be either multiplexed messages from nodes 1 and 2, or algebraic sums thereof. Again, we call these modes of operation classic routing and network coding, respectively. We note that this setup is closer to the practical mesh networks being deployed today than it is to an information theoretic perspective seeking to identify fundamental limits of wireless systems.

For simplicity, we assume that all the transmission links are time-invariant. We consider two cases: the additive white Gaussian noise (AWGN) case, and a scenario where the wireless channels are subject to path loss. Again, we suppose that every node is equipped with an arbitrary large buffer to store data packets that are awaiting transmission through the wireless medium. We also assume that a simple link layer acknowledgment scheme is present, so that data can be flushed out of the corresponding buffer once reception is confirmed.

A. Channel Model

For an AWGN channel, the maximum rate at which error-free data transfer is possible is given by

$$W \log_2\left(1 + \frac{P}{N_0 W}\right) \quad (11)$$

where P represents the expected power of the signal, $N_0/2$ is the double-sided power spectral density of the noise process, and W is the spectral bandwidth. Recent developments in error-control coding allow operation near Shannon capacity with minimal error rates and small delays. Therefore, the channel capacity expression of (11) can be viewed as an optimistic approximation of code performance. We assume that codes are designed to operate at a fixed rate, which is the constant service rate offered by the channel. In the case where all the links are AWGN limited, we allocate equal spectral bandwidth to the three nodes, and hence they become of equal capacity. We denote the constant service rate offered by each connection as $c = (W/3)\log_2(1 + 3P/(N_0W))$, where $P/(N_0W)$ is the observed signal-to-noise ratio (SNR).

In the second case, we assume that the received power decays exponentially in distance with exponential factor α. That is, the ratio of the transmit power to the received power is κd^{α}. Given a spectral bandwidth allocation of ηW and a distance of d meters, the capacity of the corresponding connection becomes

$$c(d,\eta) = \eta W \log_2\left(1 + \frac{P}{\eta N_0 W \kappa d^{\alpha}}\right) \text{ b/second.} \qquad (12)$$

Once the spectral bandwidth allocation is completed, the capacity of each link stays fixed. We choose a code rate to operate close to capacity, and this rate becomes the maximum allowable constant service rate for the corresponding queue.

B. AWGN Links

We begin the wireless analysis with the scenario where node 3 utilizes network coding. In this case, the data rate to be transmitted out of node 3 is the maximum of the rates from nodes 1 and 2, which is c. Therefore, there is no congestion at this node and the achievable rate region is limited by the QoS constraint at nodes 1 and 2. Under a QoS constraint θ_0, the maximum possible rate received by destinations C and D have identical functional form and are equal to a_1 from source A and a_2 from source B, where

$$a_i = c + \frac{c\mu_i}{\lambda_i + c\theta_0}, \qquad i \in \{1,2\}.$$

That is, the achievable rate region for the source rate pair (a_1, a_2) is $\mathcal{R}_{\text{nc}} = \mathcal{A}_1^{(1)}(\theta_0, c) \times \mathcal{A}_1^{(2)}(\theta_0, c)$.

Next, we consider the situation where node 3 simply forwards packets from nodes 1 and 2 to the destinations. In this case, source rate pairs (a_1, a_2) are also limited by the congestion at node 3; and, for all $\nu \in (0, c)$, we have $\mathcal{A}(\theta_0, c, \nu) \subset \mathcal{A}_1(\theta_0, c)$. The total achievable rate region for classic routing thus becomes

$$\mathcal{R}_{\text{cr}} = \bigcup_{0 \le \nu \le c} \mathcal{A}^{(1)}(\theta_0, c, \nu) \times \mathcal{A}^{(2)}(\theta_0, c, c - \nu).$$

The results for a specific value of θ_0 are shown in Fig. 6. The system parameters selected for this numerical study appear in Table II. Additionally, we chose a mean received power equal to 100 mW.

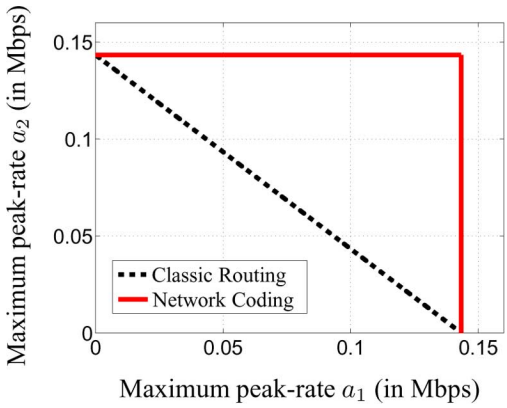

Fig. 6. Boundaries of achievable peak rate regions for on-off sources when (a) classical routing or (b) network coding is employed at intermediate node 3 for the QoS constrained communication over butterfly network of Fig. 5, where each link is additive white Gaussian noise limited. The asymptotic exponential decay rate of the buffer-occupancy is bounded below by $\theta_0 = 0.1$.

TABLE II
SYSTEM PARAMETERS

$N_0 = 10^{-6}$ W/Hz	Noise power spectral density
$W = 22$ MHz	Spectral bandwidth
$\lambda_1^{-1} = \lambda_2^{-1} = 650$ ms	Mean *off*-time
$\mu_1^{-1} = \mu_2^{-1} = 352$ ms	Mean *on*-time

C. Links With Path Loss

To illustrate the effects of path attenuation, we consider an example where the sources and destinations are located on the vertices of a perfect square of side length d. The relay node lies on the perpendicular bisector of the edges connecting the two sources at a distance δ from the top of the square. The distance from the two sources to the relay node being identical, we assume that a fraction $\eta/2$ of the total bandwidth is allocated to each source; and the remaining $(1-\eta)W$, to the relay node.

To maximize the gains of network coding, we need to make the link capacities identical. The capacity of the link between a source and the relay is $c(d_{13}, \eta/2)$, where $d_{13} = d_{23} = \sqrt{(d/2)^2 + \delta^2}$. Similarly, the capacity of the link between the relay and a destination can be written as $c(d_{35}, (1-\eta))$, where the distance from the relay to a destination is $d_{35} = d_{36} = \sqrt{(d/2)^2 + (d-\delta)^2}$. Since the service rate available to the source is limited by the minimum of the direct-link and relay-link capacities, we allocate bandwidth parameter η in the following fashion:

$$\begin{aligned}\min\left\{c\left(d_{13}, \frac{\eta}{2}\right), c\left(d, \frac{\eta}{2}\right)\right\} &= c\left(\max\{d_{13}, d\}, \frac{\eta}{2}\right) \\ &= c\left(d_{35}, (1-\eta)\right). \end{aligned} \qquad (13)$$

We denote this optimal bandwidth allocation parameter by η^*, which is evaluated numerically. We would like to point out that this η^* is unique, which follows from the monotonicity of $c(d,\eta)$ in η [see (12)].

Proposition 1: The optimal bandwidth allocation parameter η^* that satisfies (13) is unique.

Proof: It is easy to see that $c(d,\eta)$ is continuous and even differentiable in the range $(0,1]$. We will quickly show

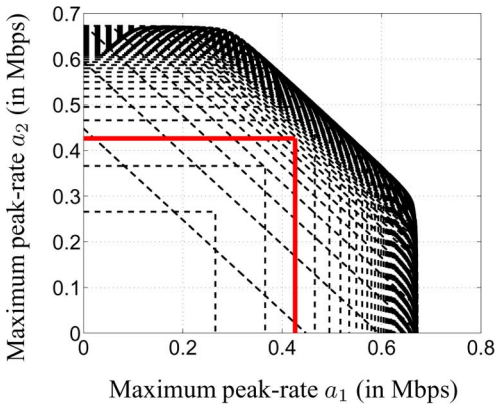

Fig. 7. Boundaries of achievable peak rate regions for on-off sources when (a) classical routing (denoted by dashed lines for values of fraction $\eta \in [0.01, 0.99]$) and (b) network coding (denoted by the solid line) are employed at the intermediate node 3 for the QoS constrained communication over butterfly network of Fig. 5 where each link is limited by path loss. The asymptotic exponential decay rate of the buffer occupancy is bounded below by $\theta_0 = 0.1$ and the relay node is at distance $\delta = 9$ m.

that $c(d,\eta)$ is strictly increasing in η. Then, it will follow that $c(d,(1-\eta))$ is continuous and decreasing in η and hence η^* is unique since $c(d_1,0) = 0 < c(d_2,1)$ for any finite d_1, d_2. To that end, it suffices to show that $\frac{\partial c(d,\eta)}{\partial \eta} \geq 0$ for all $\eta \in [0,1]$. From (12), it is straightforward to verify that $\frac{\partial c(d,\eta)}{\partial \eta}$ is decreasing in η; hence we only need to check that $\frac{\partial c(d,\eta)}{\partial \eta}|_{\eta=1} = \frac{W}{\ln 2}\left(\ln(1+k_2) - \frac{k_2}{1+k_2}\right)$ is greater than zero. Here, we are denoting $\frac{P}{N_0 W \kappa d^\alpha}$ by k_2. Clearly, if $f(k_2) = \ln(1+k_2) - \frac{k_2}{1+k_2} \geq 0$ for all $k_2 \geq 0$, we are done. This is easy to verify since $f(0) = 0$ and $f'(k_2) = \frac{k_2}{(1+k_2)^2} \geq 0$. □

For the classic routing case, each source broadcasts its packets to the relay node where information is stored. The relay then forwards the received messages to the destinations using a first-come-first-serve service policy. The links from the sources to the relay node have identical capacity $c(d_{13}, \eta/2) = c(d_{23}, \eta/2)$. Similarly, the link from the relay node to the destinations have capacity $c(d_{35}, (1-\eta))$. Using the same queueing performance analysis as before and for a fixed δ, we can obtain the achievable rate region under QoS constraint θ_0 for network coding and classic routing (for different values of η in the latter case). The two regions are shown in Fig. 7 for a transmit power of $P = 40$ W and the system parameters of Table II. In this example, we have taken $d = 15$ m and $\alpha = 1.8$ (typical values of α range from 1.6 to 1.8 for line-of-sight communication in buildings [41, p. 139, Table 4.2]). The region enclosed by the thick solid line represents the achievable rate region achieved by network coding. The thin dashed lines characterize the regions corresponding to classic routing for different values of η. Clearly, classic routing outperforms network coding in this case. This is a scenario where the cost of establishing a network topology suitable for network coding exceeds the benefits of packet combining.

VI. Queueing Results

Below, we list and derive queueing results related to the analysis of the tandem network introduced in Section III. Recall that the arrival process feeding the first queue is a Markov-modulated fluid process with *on* rate a. This queue is serviced at a constant rate c, whereas the second queue is served at a rate v. When $a \leq c$, the first buffer in the system remains empty at all times, and the analysis of the tandem network degenerates into a single-queue scenario. We therefore assume that $a > c$, which is the more interesting case. For this situation, the following theorem enables us to obtain the tail-asymptotics of the first buffer in a tandem network. The corresponding rate region is governed by (4) and (6).

Theorem 1 (Mitra [36], [42], [9]): Let $\ell_1(t)$ be the amount of fluid in an arbitrary large reservoir being fed by an on-off source satisfying Assumption 1, and serviced at a constant rate c. The off and on times of the source are exponentially distributed with means λ^{-1} and μ^{-1}, respectively. If the constant arrival rate a of the fluid in the on state is such that

$$\frac{a\lambda}{\lambda+\mu} < c < a$$

then the limit $\lim_{t\to\infty} \Pr\{\ell_1(t) > x\}$ exists for all $x \geq 0$. The corresponding asymptotic decay rate of buffer occupancy can be identified by looking at the largest negative eigenvalue ζ that satisfies the matrix equation $\zeta D\phi = Q\phi$. Here, Q is the generator matrix for the modulating Markov chain and D is the drift matrix corresponding to the arrival process. More precisely, we have

$$\theta_1 = -\lim_{x\to\infty} \frac{\ln \Pr\{\ell_1 > x\}}{x} = \frac{\mu}{a-c} - \frac{\lambda}{c}$$

where ℓ_1 is a random variable whose distribution coincides with the equilibrium distribution of the queue.

This is a standard result and, as such, we state it without proof. For alternate treatments of this theorem and other pertinent queueing results on the decay rate of buffer-occupancy under various conditions, see [7]–[11], [13], [14], [18], [31], and [37].

A. Departure Process of a Fluid Buffer

We next consider the case where there are two queues in tandem and the departure process of the first queue serves as the input to the second buffer. The departure process at the output of the first buffer is characterized using a theorem first proved by Aalto [39], [40].

Theorem 2 (Aalto): For the fluid queue described in Theorem 1, the departure process can be viewed as an on-off source where packets are emitted at a constant rate during an on period, and the source is idle otherwise. The off times are exponentially distributed with mean λ^{-1}; while the on times have the same distribution as the duration of a busy period in an M/M/1 queue with arrival rate $(1-c/a)\lambda$ and service rate $c\mu/a$. Furthermore,

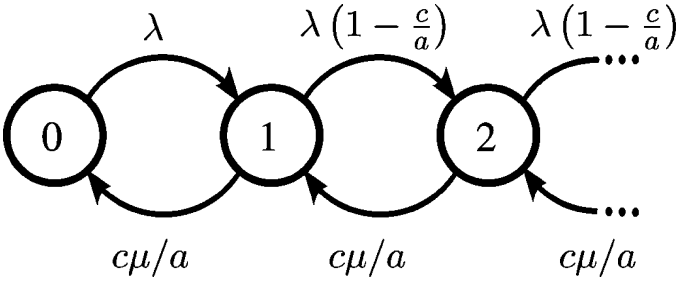

Fig. 8. Graphical representation of the modulating birth-death process.

the departure rate is c when the queue is nonempty. Mathematically, this departure process is modulated by a countable state birth-death process $\{K(t)\colon t \geq 0\}$ with given transition rates

$$\lambda_{0,1} = \lambda, \tag{14}$$

$$\lambda_{n-1,n} = \left(\frac{a-c}{a}\right)\lambda, \qquad n = 2,3,\ldots \tag{15}$$

$$\mu_{n,n-1} = \frac{c\mu}{a}, \qquad n \in \mathbb{N}. \tag{16}$$

The modulating birth-death process is depicted in Fig. 8. The departure process is off, when $K(t) = 0$ and it is on otherwise. That is, the departure rate is given by

$$c(K(t)) = \begin{cases} 0, & K(t) = 0 \\ c, & K(t) \in \mathbb{N}. \end{cases} \tag{17}$$

Having characterized the departure process of the queue, we next present an expression for the equilibrium distribution of $K(t)$. Let

$$p_n = \lim_{t\to\infty} \Pr(K(t) = n), \qquad n = 0,1,2,\ldots$$

and define $\rho = \left(\frac{a-c}{c}\right)\frac{\lambda}{\mu}$. Then, the stationary distribution of K is given by [43]

$$p_0 = \frac{(a-c)(1-\rho)}{a-c+c\rho}$$

$$p_n = \left(\frac{a}{a-c}\right)\rho^n p_0, \qquad n = 1,2,\ldots$$

B. Distribution of the Tandem Queue

For the tandem network described in Section III, we wish to find the equilibrium distribution of the second buffer. Note that we already have similar results for the first queue based on Theorem 1. We also have knowledge of the departure process from the first queue, as afforded by Theorem 2. The properties of the second queue can therefore be studied based on this departure process alone.

Consider a buffer that is being fed by a Markov modulated *on-off* source generating fluid at a rate $c(K(t))$, where $c(K(t))$ is the departure function of (17) and $K(t)$ is the modulating Markov process described in Theorem 2. The arrival rate in the queue is c when $K(t) \geq 1$, and it is zero otherwise. Fluid is removed from the queue at constant rate υ, provided that it is nonempty. Paralleling the approach of Virtamo and Norros [38], we derive necessary and sufficient conditions for a nontrivial stationary probability distribution to exist. Then, we find the spectrum of the operator that governs the equilibrium distribution of the buffer. This allows us to present an explicit expression for the distribution of the queue.

Clearly, if $\upsilon \geq c$, there is no congestion in this buffer. We therefore examine the case where $\upsilon < c$. Writing the stochastic evolution equation for the buffer of interest, we get

$$\frac{d}{dt}\ell_2(t) = (c-\upsilon)\mathbf{1}_{\{K(t)\neq 0\}} - \upsilon\mathbf{1}_{\{K(t)=0,\ell_2(t)>0\}}.$$

It can be shown that the stochastic process governing the evolution of this buffer $\{\ell_2(t)\colon t \geq 0\}$ is positive recurrent if and only if [31], [44]

$$\left(\frac{a-\upsilon}{\upsilon}\right)\frac{\lambda}{\mu} < 1.$$

That is, there exists a stationary probability distribution for the buffer provided that the stability condition mentioned above is satisfied. Applying standard queueing arguments, we can derive the Chapman–Kolmogorov equations for the probability distribution of stationary buffer occupancy ℓ_2 as follows. We use $\pi(x,n)$ to denote the probability that ℓ_2 exceeds x while the underlying birth-death process is in state n, i.e.,

$$\begin{aligned}\pi(x,n) &= \lim_{t\to\infty} \Pr(\ell_2(t) > x, K(t) = n) \\ &= \Pr(\ell_2 > x, K = n).\end{aligned}$$

The corresponding Chapman–Kolmogorov equations become

$$\begin{aligned}&\left(c\mathbf{1}_{\{n\geq 1\}} - \upsilon\right)\frac{d}{dx}\pi(x,n) \\ &\quad = \lambda_{n-1,n}\mathbf{1}_{\{n\geq 1\}}\pi(x,n-1) + \mu_{n+1,n}\pi(x,n+1) \\ &\qquad - \left(\lambda_{n,n+1} + \mu_{n,n-1}\mathbf{1}_{\{n\geq 1\}}\right)\pi(x,n)\end{aligned} \tag{18}$$

for $n = 0,1,2,\ldots$ The constants $\{\lambda_{n-1,n}, \mu_{n,n-1}\colon n \in \mathbb{Z}^+\}$ are the transition rates of the modulating process $K(t)$ defined in (14)–(16). Additionally, we employ the convention that $\mu_{0,-1} = \lambda_{-1,0} = 0$. With some abuse of notation, we define the sequence $\pi(x)$ as

$$\pi(x) = \left\{\pi(x,n)\colon n \in \mathbb{Z}^+\right\}.$$

Let $\mathbf{H}$ denote the Hilbert space of square summable sequences indexed by $\mathbb{Z}^+$, and let $\mathcal{B}(\mathbf{H})$ be the space of bounded linear operators from $\mathbf{H}$ to itself. For sequence $\mathbf{h} \in \mathbf{H}$, we define operators D and Q in $\mathcal{B}(\mathbf{H})$ by

$$(D\mathbf{h})_n = \left(c\mathbf{1}_{\{n\geq 1\}} - \upsilon\right)h_n, \tag{19}$$

$$\begin{aligned}(Q\mathbf{h})_n &= \lambda_{n-1,n}\mathbf{1}_{\{n\geq 1\}}h_{n-1} - \lambda_{n,n+1}h_n \\ &\quad + \frac{c\mu}{a}\left(h_{n+1} - h_n\mathbf{1}_{\{n\geq 1\}}\right)\end{aligned} \tag{20}$$

where $\{\lambda_{n-1,n}\}$ are defined in (14)–(16). It is straightforward to check continuity of these two operators. In particular, for any $\mathbf{h} \in \mathbf{H}$, we have

$$\begin{aligned}\|D(\mathbf{h})\| &\leq \max\{\upsilon, c-\upsilon\}\|\mathbf{h}\| \\ \|Q(\mathbf{h})\| &\leq 3\left(\lambda + \frac{c\mu}{a}\right)\|\mathbf{h}\|.\end{aligned}$$

For the aforementioned system, we can rewrite (18) in a compact form in terms of the sequence $\pi(x)$ and the bounded linear transformations Q and D

$$D\frac{d}{dx}\pi(x) = Q\pi(x). \tag{21}$$

The transformation D can be expressed in terms of the identity operator I and the standard projection operator $\mathbf{e}_0^*$. For any $\mathbf{h} \in \mathbf{H}$, we have $I\mathbf{h} = \mathbf{h}$ and $\mathbf{e}_0^*\mathbf{h} = h_0$. We can then write

$$D = (c-\upsilon)\tilde{D} = (c-\upsilon)\left(I - \frac{c}{c-\upsilon}\mathbf{e}_0\mathbf{e}_0^*\right).$$

The linear transformation Q can be described in matrix form as

$$Q = \frac{c\mu}{a}E\tilde{Q}E^{-1}$$

where we have defined the operators $E = \operatorname{diag}\left(\sqrt{\frac{a-c}{a}}, \rho^{1/2}, \rho, \rho^{3/2}, \ldots\right)$ and

$$\tilde{Q} = \begin{bmatrix} \frac{-a\rho}{a-c} & \sqrt{\frac{a\rho}{a-c}} & 0 & 0 & \cdots \\ \sqrt{\frac{a\rho}{a-c}} & -(1+\rho) & \sqrt{\rho} & 0 & \cdots \\ 0 & \sqrt{\rho} & -(1+\rho) & \sqrt{\rho} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

We emphasize that $\tilde{Q}$ is a symmetric matrix, which makes it very convenient for the spectral analysis carried out below. Note also that E, E^{-1}, and D are diagonal and hence commutative operators.

It is easy to see that $\tilde{D}$ and $\tilde{Q}$ belong to $\mathcal{B}(\mathbf{H})$, and that $\tilde{D}$ is invertible. It follows that $\tilde{D}^{-1}\tilde{Q} \in \mathcal{B}(\mathbf{H})$. Since $\tilde{D}^{-1} = \left[I - \frac{c}{\upsilon}\mathbf{e}_0\mathbf{e}_0^*\right]$, we can represent $\tilde{D}^{-1}\tilde{Q}$ using a countable state matrix as

$$\begin{bmatrix} \frac{-a\rho}{a-c}\left(\frac{\upsilon-c}{\upsilon}\right) & \sqrt{\frac{a\rho}{a-c}}\left(\frac{\upsilon-c}{\upsilon}\right) & 0 & 0 & \cdots \\ \sqrt{\frac{a\rho}{a-c}} & -(1+\rho) & \sqrt{\rho} & 0 & \cdots \\ 0 & \sqrt{\rho} & -(1+\rho) & \sqrt{\rho} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{22}$$

It should also be apparent from the Chapman–Kolmogorov equation of (21) that

$$\pi(x) = E\exp\left(x\frac{c\mu}{a(c-\upsilon)}\tilde{D}^{-1}\tilde{Q}\right)E^{-1}\pi(0). \tag{23}$$

To evaluate $\pi(x)$, we need to identify boundary condition $\pi(0)$ and find a spectral representation for the operator

$$\exp\left(x\frac{c\mu}{a(c-\upsilon)}\tilde{D}^{-1}\tilde{Q}\right). \tag{24}$$

The equilibrium distribution $\pi(0)$ can be obtained from the natural boundary conditions on the buffer occupancy for a stable system; this is accomplished later. Rather, we begin by deriving an expression for (24) through a two-step approach: first we compute the spectrum of $\tilde{D}^{-1}\tilde{Q}$, and then we obtain an equivalent representation for I in terms of the associated eigenfunctions.

1) Spectrum of Bounded Linear Operator: In general, finding the spectrum of a noncompact bounded linear operator is a difficult task. However, in the present case, the relevant operator is a compact perturbation of a linear combination of standard shift operators. It is well known in the literature that the continuous spectrum of a self-adjoint operator remains invariant under compact perturbations [45], [46]. We present relevant results formally in the following theorem, which will be used to obtain a spectral representation for the operator of (24).

Let us denote the spectrum of an operator A by

$$\sigma(A) = \{\zeta \in \mathbb{C}\colon (A - \zeta I) \text{ is not invertible in } \mathcal{B}(\mathbf{H})\}.$$

This spectrum is composed of two parts. First, there is the discrete spectrum of A, also called the set of eigenvalues of A, which is defined as $\sigma_d(A) = \{\zeta \in \mathbb{C}\colon (A-\zeta I) \text{ is not injective}\}$. That is, ζ is an eigenvalue of A if and only if there exists a sequence $\mathbf{h} \in \mathbf{H}$ such that $(A - \zeta I)\mathbf{h} = \mathbf{0}$. On the other hand, the values of $\zeta \in \sigma(A)$ for which the operator $(A - \zeta I)$ is injective but not surjective belong to the continuous spectrum of A. If there exists $\mathbf{h} \in \mathbf{H}$ such that $(A - \zeta I)^{-1}\mathbf{h} \notin \mathbf{H}$, yet the range of operator $(A - \zeta I)^{-1}$ is dense in $\mathbf{H}$, then $\zeta \in \sigma_c(A)$.

Theorem 3: Let operators $D, Q \in \mathcal{B}(\mathbf{H})$ be as defined in (19)–(20). Then, the continuous spectrum of $D^{-1}Q$ is

$$\sigma_c(D^{-1}Q) = \left[-\frac{c\mu}{a(c-\upsilon)}\left(1 + \sqrt{\left(\frac{a-c}{c}\right)\frac{\lambda}{\mu}}\right)^2, -\frac{c\mu}{a(c-\upsilon)}\left(1 - \sqrt{\left(\frac{a-c}{c}\right)\frac{\lambda}{\mu}}\right)^2\right] \tag{25}$$

and $D^{-1}Q$ has a discrete eigenvalue $\zeta_0 = 0$. Furthermore, if

$$\frac{a/c - 1}{a/\upsilon - 1} < \sqrt{\left(\frac{a-c}{c}\right)\frac{\lambda}{\mu}}$$

then there is an additional eigenvalue located at $\zeta_1 = -\frac{\mu}{a-\upsilon} + \frac{\lambda}{\upsilon}$. In this latter case, we have $\zeta_1 > \sup \sigma_c(D^{-1}Q)$.

Let $\mathbf{w} \in l^\infty(\mathbb{Z}^+)$ be such that, for some ζ, we have

$$\zeta D\mathbf{w} = Q\mathbf{w}. \tag{26}$$

Then, ζ belongs to the spectrum of the linear operator $D^{-1}Q$. Substituting $\xi = \frac{a(c-\upsilon)}{c\mu}\zeta$ and $\tilde{\mathbf{w}} = E^{-1}\mathbf{w}$, we obtain

$$\xi\tilde{D}\tilde{\mathbf{w}} = \tilde{Q}\tilde{\mathbf{w}}. \tag{27}$$

There is a one-to-one correspondence between the eigenvalues and eigenfunctions of (26) and (27). Therefore, it suffices to show that spectrum of $\tilde{D}^{-1}\tilde{Q}$ has a continuous part $\left[-\left(1+\sqrt{\rho}\right)^2, -\left(1-\sqrt{\rho}\right)^2\right]$ and an eigenvalue $\xi_0 = 0$. In addition, we need to show that operator $\tilde{D}^{-1}\tilde{Q}$ also has an additional eigenvalue $\xi_1 = -(1-c')\left(1 - \frac{\rho}{c'}\right)$ when $\sqrt{\rho} > c'$. Here we have implicitly defined $c' = \frac{a/c-1}{a/\upsilon-1}$. We show in Appendix I that this is indeed true. We choose to solve (27) due to its symmetric structure and greater simplicity.

2) Spectral Representation of Identity: For further analysis, we introduce variable y and constants p and q in the following way:

$$y = y(\xi) = \frac{1+\rho+\xi}{2\sqrt{\rho}}, \quad p = \frac{2}{1-c'}, \quad q = -\frac{c'+\rho}{(1-c')\sqrt{\rho}}.$$

We emphasize that there is a one-to-one correspondence between y and ξ; we can therefore express y as a function of ξ, and ξ as a function of y unambiguously. Hence, we use $y(\xi)$ and $\xi(y)$ interchangeably depending on the context. We define $y_0 = y(\xi_0)$ and $y_1 = y(\xi_1)$ for eigenvalues ξ_0 and ξ_1. In addition, we assume without loss of generality that $\tilde{w}_0 = 1$ for eigensequence $\tilde{\mathbf{w}}(y(\xi))$ corresponding to $\xi \in \sigma(\tilde{D}^{-1}\tilde{Q})$. For convenience, we make the dependence of $\tilde{\mathbf{w}}$ on y (and hence ξ) explicit hereafter. For eigenvalues ξ_0 and ξ_1, we gather from Appendix I that the corresponding eigensequences are

$$\tilde{\mathbf{w}}(y_0) = \sqrt{\frac{a}{a-c}}\left(\sqrt{\frac{a-c}{a}}, \sqrt{\rho}, \rho, \dots\right)$$
$$\tilde{\mathbf{w}}(y_1) = \sqrt{\frac{a}{a-c}}\left(\sqrt{\frac{a-c}{a}}, \frac{c'}{\sqrt{\rho}}, \left(\frac{c'}{\sqrt{\rho}}\right)^2, \dots\right).$$

This leads to their z-transforms being

$$\tilde{W}(y_0, z) = \left(1 - \sqrt{\frac{a}{a-c}}\right) + \sqrt{\frac{a}{a-c}}\left(\frac{1}{1-z\sqrt{\rho}}\right)$$
$$\tilde{W}(y_1, z) = \left(1 - \sqrt{\frac{a}{a-c}}\right) + \sqrt{\frac{a}{a-c}}\left(\frac{1}{1-zc'/\sqrt{\rho}}\right).$$

We can also find $\tilde{W}(y_i, z)$ by substituting $\alpha(\xi_i) = i \in \{0, 1\}$ in (34). Furthermore, we can write the z-transform of the eigensequence corresponding to $\xi \in \sigma_c(\tilde{D}^{-1}\tilde{Q})$ by substituting y, p, and q in (33) to obtain

$$\tilde{W}(y, z) = \sum_{k \in \mathbb{Z}^+} \tilde{w}_k(y) z^k = \left(1 - \sqrt{\frac{a}{a-c}}\right) + \sqrt{\frac{a}{a-c}}\left(\frac{1-(py+q)z}{1-2yz+z^2}\right).$$

We notice that $\tilde{W}\left(y, \sqrt{\rho}\right) = 1 + \sqrt{\frac{a}{a-c}}\frac{c'}{(1-c')}$ is independent of y for $y \neq y_0$, and $\tilde{W}\left(y_0, \sqrt{\rho}\right) = 1 + \sqrt{\frac{a}{a-c}}\frac{\rho}{(1-\rho)}$.

For any eigensequence $\tilde{\mathbf{w}}$ and corresponding $\xi \in \sigma(\tilde{D}^{-1}\tilde{Q})$, we have $\tilde{Q}\tilde{\mathbf{w}} = \xi\tilde{D}\tilde{\mathbf{w}}$, which in turn implies

$$\tilde{\mathbf{w}}^*\tilde{Q} = \xi\tilde{\mathbf{w}}^*\tilde{D}. \tag{28}$$

Moreover, for $i \in \{0, 1\}$ and $y \in [-1, 1]$, we have $\xi(y_i) \neq \xi(y)$. Note that $\tilde{\mathbf{w}}(y) \notin \mathbf{H}$; however, $\tilde{\mathbf{w}}(y) \in l^\infty(\mathbb{Z}^+)$ and $\tilde{\mathbf{w}}(y_i) \in l^1(\mathbb{Z}^+)$ (see Appendix I). Therefore, the usual inner product on $\mathbf{H}$ (sum of pointwise product of sequences) is well defined for $\tilde{\mathbf{w}}(y)$ and $\tilde{\mathbf{w}}(y_i)$. Hence, using (28) and the eigenrelationship, we get $\tilde{\mathbf{w}}^*(y_i)\xi(y)\tilde{D}\tilde{\mathbf{w}}(y) = \xi(y_i)\tilde{\mathbf{w}}^*(y_i)\tilde{D}\tilde{\mathbf{w}}(y)$, which in turn yields

$$(\xi(y) - \xi(y_i))\langle\tilde{D}\tilde{\mathbf{w}}(y), \tilde{\mathbf{w}}(y_i)\rangle = 0.$$

Thus, we obtain

$$\langle\tilde{D}\tilde{\mathbf{w}}(y), \tilde{\mathbf{w}}(y_i)\rangle = 0. \tag{29}$$

It is equally straightforward to see that

$$\langle\tilde{D}\tilde{\mathbf{w}}(y_1), \tilde{\mathbf{w}}(y_0)\rangle = 0. \tag{30}$$

Below, we use these orthogonal properties to derive the desired spectral representation of the identity operator.

Theorem 4: Let s_0, s_1 denote the weights corresponding to eigenvalues ξ_0 and ξ_1, and let $s(y)$ be the weighting function associated with the continuous spectrum of $\tilde{D}^{-1}\tilde{Q}$. Define these quantities as follows:

$$s_0 = (\tilde{\mathbf{w}}(y_0)^*\tilde{D}\tilde{\mathbf{w}}(y_0))^{-1} = -\left(\frac{a-c}{a}\right)(1-\rho)\left(\frac{1-c'}{c'-\rho}\right)$$
$$s_1 = \mathbf{1}_{\{c'<\sqrt{\rho}\}}\left(\tilde{\mathbf{w}}(y_1)^*\tilde{D}\tilde{w}(y_1)\right)^{-1} = \mathbf{1}_{\{c'<\sqrt{\rho}\}}\left(\frac{a-c}{a}\right)\left(\frac{\rho}{c'} - c'\right)\left(\frac{1-c'}{c'-\rho}\right)$$
$$s(y) = \frac{2}{\pi}\left(\frac{a-c}{a}\right)\frac{\sqrt{1-y^2}}{1-y^2+((p-1)y+q)^2}.$$

Then, the identity operator I can be expressed in terms of $\tilde{D}$ as

$$I = \left(s_0\tilde{\mathbf{w}}(y_0)\tilde{\mathbf{w}}(y_0)^* + s_1\tilde{\mathbf{w}}(y_1)\tilde{\mathbf{w}}(y_1)^* + \int_{-1}^{1} s(y)\tilde{\mathbf{w}}(y)\tilde{\mathbf{w}}(y)^* dy\right)\tilde{D}. \tag{31}$$

A proof for this result is contained in Appendix II.

We are now ready to characterize the equilibrium distribution of the buffer overflow probability for the second buffer in a tandem network.

Theorem 5: For the system described in Theorem 1, let the departure process of the first queue serve as an input to the second buffer. The latter queue is assumed to be served at constant rate υ and its occupancy is denoted by $\ell_2(t)$. If

$$\left(\frac{a}{\upsilon} - 1\right)\frac{\lambda}{\mu} < 1$$

then we can express the equilibrium probability distribution of $\ell_2(t)$ exceeding a threshold x as

$$\lim_{t\to\infty} \Pr(\ell_2(t) > x) = K'\left(s_1 e^{\zeta_1 x} + \int_{-1}^{1} s(y)e^{\zeta(y)x} dy\right)$$

where K' is a constant.

This result is proved in Appendix III.

VII. Tail Asymptotics for Buffer Occupancy

In this section, we characterize the exponential decay rate of the complementary cumulative distribution function (also referred to as tail-asymptote) of the equilibrium buffer-occupancy random variable of the second buffer in a tandem queue. We emphasize that finding this tail-asymptote is essentially the same

as obtaining the effective bandwidth of the departure process discussed in Section VI-A (see also [39] and [40]). From the buffer distribution derived above, we can obtain the dominant exponential decay rate. An alternate way of finding the effective bandwidth of the departure process would be to first compute the moment generating function of the busy period of the fluid queue [47]–[49], and then use the method proposed in [50]. There is also literature available on finding tail-asymptote of the departure process in a discrete-time queue [11], [13], [14].

Recall that $\zeta(y) < \zeta_1 < 0$ for all $y \in [-1,1]$, whenever $\sqrt{\rho} > c'$. That is, if the discrete eigenvalue ζ_1 exists, then it is larger than the supremum of the continuous spectrum. We characterize the tail-asymptote in the following theorem.

Theorem 6: For $\ell_2(t)$ described in Theorem 5, the exponential decay rate associated with the steady-state probability of the buffer exceeding a threshold is given by θ_2, where

$$\begin{aligned}\theta_2 &= -\lim_{x\to\infty}\frac{\ln \Pr(\ell_2 > x)}{x} \\ &= \begin{cases} \frac{\mu}{a-\upsilon} - \frac{\lambda}{\upsilon} & \sqrt{\rho} > c' \\ \frac{c\mu}{a(c-\upsilon)}(1-\sqrt{\rho})^2 & \sqrt{\rho} \le c.' \end{cases} \end{aligned} \tag{32}$$

Proof: We use the fact that $s(y)$ is nonnegative, bounded and integrable. For $\sqrt{\rho} > c'$, the desired result follows from

$$K' s_1 e^{\zeta_1 x} \le \Pr(\ell_2 > x) \le K' e^{\zeta_1 x}\left(s_1 + \int_{-1}^{1} s(y)dy\right).$$

On the other hand, for the case where $\sqrt{\rho} \le c'$ and for some $\epsilon \in (0,1)$, we have

$$\begin{aligned} K' e^{\zeta(1-\epsilon)x}\int_{1-\epsilon}^{1} s(y)dy &\le \Pr(\ell_2 > x) \\ &\le K' e^{\zeta(1)x}\int_{-1}^{1} s(y)dy. \end{aligned}$$

Taking logarithms on both sides, dividing by x, and taking limits, we get

$$\begin{aligned}\zeta(1-\epsilon) &\le \liminf_{x\to\infty}\frac{\ln \Pr(\ell_2 > x)}{x} \\ &\le \limsup_{x\to\infty}\frac{\ln \Pr(\ell_2 > x)}{x} \le \zeta(1).\end{aligned}$$

Finally, letting $\epsilon \to 0$ and using the continuity of $\zeta(\cdot)$, we obtain the desired result. □

VIII. Maximum Achievable Rate for Departure Process

In Theorem 6, we found the tail-asymptote of the second buffer in a tandem queue as described in Theorem 1. In this section, we find the achievable rate region $\mathcal{A}_2(\theta_0, c, \nu)$ for the tandem queue considered in Section III. While establishing this region, several cases must be considered. These cases are not individually difficult, but they are collectively tedious. Therefore, the proof appears in Appendix IV. We have tried to make the presentation as clear and concise as possible.

Theorem 7: For exponential-decay rate θ_2 described in Theorem 6, we define

$$\begin{aligned}\mathcal{A}_2(\theta_0, c, \upsilon) &= \{a \in \mathbb{R}^+ \colon \theta_2 \ge \theta_0\} \\ &= \{a \in \mathbb{R}^+ \colon a \le \bar{a}_2(\theta_0, c, \upsilon)\}\end{aligned}$$

where the function $\bar{a}_2(\theta, c, \upsilon)$ is given by

$$\bar{a}_2(\theta, c, \upsilon) = \begin{cases} \bar{a}_1(\theta, \upsilon), & 0 < \upsilon \le \upsilon^* \\ \bar{a}_3(\theta, c, \upsilon), & \upsilon^* < \upsilon < c. \end{cases}$$

The first component $\bar{a}_1(\theta, \upsilon)$ is equal to

$$\bar{a}_1(\theta, \upsilon) = \nu\left(1 + \frac{\mu}{\lambda + \nu\theta}\right)$$

and the second component $\bar{a}_3(\theta, c, \upsilon)$ is given by the expression

$$\begin{aligned}&\bar{a}_3(\theta, c, \upsilon) \\ &= c + \frac{c\mu}{\lambda}\left(\frac{-1 + \sqrt{1 - \left((c-\upsilon)\frac{\theta}{\mu} - 1\right)\left((c-\upsilon)\frac{\theta}{\lambda} - 1\right)}}{(c-\upsilon)\frac{\theta}{\lambda} - 1}\right)^2\end{aligned}$$

with parameter υ^* determined implicitly by

$$\frac{c}{\upsilon^*} - 1 = \frac{\theta\upsilon^*\mu}{\lambda\mu + (\lambda + \theta\upsilon^*)^2}.$$

IX. Conclusion

We compared network coding to classic routing for a QoS constrained communication system, and computed the achievable rate regions for both scenarios. For an AWGN model with identical link capacities, network coding significantly outperforms classic routing. This essentially implies that the benefits associated with network coding are far more important than the multiplexing gains achieved by routing for symmetric networks. However, we obtained more interesting results for the wireless butterfly network. In this case, allocating resources to form a network topology suitable for packet combining at intermediate nodes does not always offer gains and may even be detrimental at times. These results depend on the topology of the butterfly network. For network coding to be useful, we need symmetric direct links. It turns out that it is often better to route packets rather than trying to establish a direct link to the destination using excessive amounts of physical resources. A possible avenue of future research is to study networks with varying service rates.

Appendix I
Proof of Theorem 3

From (22), it follows that

$$\begin{aligned}\xi\tilde{w}_0 &= -\frac{a\rho}{a-c}\left(\frac{\upsilon - c}{\upsilon}\right)\tilde{w}_0 + \sqrt{\frac{a\rho}{a-c}}\left(\frac{\upsilon - c}{\upsilon}\right)\tilde{w}_1 \\ \xi\tilde{w}_1 &= \sqrt{\frac{a\rho}{a-c}}\tilde{w}_0 - (1+\rho)\tilde{w}_1 + \sqrt{\rho}\tilde{w}_2 \\ \xi\tilde{w}_n &= \sqrt{\rho}\tilde{w}_{n-1} - (1+\rho)\tilde{w}_n + \sqrt{\rho}\tilde{w}_{n+1}, \qquad n \ge 2.\end{aligned}$$

Taking the z-transform, $\tilde{W}(z) = \sum_{n=0}^{\infty} \tilde{w}_n z^n$, we get

$$\tilde{W}(z) = \left(1 - \sqrt{\frac{a}{a-c}}\right)\tilde{w}_0 + \sqrt{\frac{a}{a-c}}\left[\frac{1 - \left(1 + \xi\frac{c/a - c/\upsilon}{1 - c/\upsilon}\right)\frac{z}{\sqrt{\rho}}}{1 - \frac{1+\rho+\xi}{\sqrt{\rho}}z + z^2}\right]\tilde{w}_0. \quad (33)$$

We define $\gamma_0(\xi), \gamma_1(\xi)$ to be the roots of the characteristic polynomial $\gamma^2 - \frac{(1+\rho+\xi)}{\sqrt{\rho}}\gamma + 1$.

To solve for $\tilde{\mathbf{w}}$ from its z-transform, we break the problem into two separate cases. First, assume that $|\gamma_0(\xi)| \neq 1$. Then $\gamma_0(\xi), \gamma_1(\xi)$ are two different roots, and the z-transform can be written as

$$\tilde{W}(z) = \left(1 - \sqrt{\frac{a}{a-c}}\right)\tilde{w}_0 + \sqrt{\frac{a}{a-c}}\left(\frac{1-\alpha(\xi)}{1 - z\gamma_0(\xi)} + \frac{\alpha(\xi)}{1 - z\gamma_1(\xi)}\right)\tilde{w}_0 \quad (34)$$

where we have implicitly defined

$$\alpha(\xi) = \frac{1}{2}\left(1 + \frac{\rho - 1 - \frac{1+c'}{1-c'}\xi}{\sqrt{(1+\rho+\xi)^2 - 4\rho}}\right).$$

From this decomposition, we gather that

$$\tilde{w}_n(\xi) = \sqrt{\frac{a}{a-c}}\Big((1-\alpha(\xi))\gamma_0(\xi)^n + \alpha(\xi)\gamma_1(\xi)^n\Big)\tilde{w}_0$$

where $n \in \mathbb{N}$. Note that $\tilde{\mathbf{w}} \in \mathbf{H}$ only when $\alpha(\xi)$ is 0 or 1 since $\gamma_0(\xi)\gamma_1(\xi) = 1$ (see [38]). The corresponding eigenvalues are $\xi_0 = 0$ and $\xi_1 = -(1-c')\left(1 - \frac{\rho}{c'}\right)$. Since $\gamma_0(0) = \sqrt{\rho}$, zero is always an eigenvalue of (27). However, $\gamma_1(\xi_1) = \frac{c'}{\sqrt{\rho}}$ which implies that ξ_1 is an eigenvalue only when $c' < \sqrt{\rho}$.

On the other hand, we claim that ξ belongs to the continuous spectrum if and only if $|\gamma_1(\xi)| = 1$; this is equivalent to $\xi \in \left[-\left(1+\sqrt{\rho}\right)^2, -\left(1-\sqrt{\rho}\right)^2\right]$. To prove the claim, we note that $\tilde{D}^{-1}\tilde{Q}$ can be equivalently expressed in terms of the right shift operator S, the left shift operator T, a compact perturbation K, and the identity operator I. In particular, we can write

$$\tilde{D}^{-1}\tilde{Q} = \sqrt{\rho}\left(S + T - \frac{1+\rho}{\sqrt{\rho}}I + K\right) \quad (35)$$

where $S = \sum_{n\in\mathbb{Z}^+} \mathbf{e}_{n+1}\mathbf{e}_n^*$, $T = \sum_{n\in\mathbb{Z}^+} \mathbf{e}_n\mathbf{e}_{n+1}^*$, and

$$K = \left(\frac{1+\rho}{\sqrt{\rho}} - \sqrt{\rho}\left(\frac{a\upsilon - ac}{a\upsilon - c\upsilon}\right)\right)\mathbf{e}_0\mathbf{e}_0^* - \left(1 - \sqrt{\frac{a}{a-c}}\right)\mathbf{e}_1\mathbf{e}_0^* - \left(1 - \sqrt{\frac{a}{a-c}}\left(1 - \frac{c}{\upsilon}\right)\right)\mathbf{e}_0\mathbf{e}_1^*.$$

It is immediate that $S+T$ is self-adjoint with real continuous spectrum $\sigma_c(S+T)$, and $\|S\| = \|T\| = 1$. Also, it is a well-known result by Weyl that the continuous spectrum (plus limit points of point spectrum if any) of a self-adjoint operator remains unchanged under compact perturbations [45], [46]. This fact, along with (35), gives us

$$\sigma_c\left(\tilde{D}^{-1}\tilde{Q}\right) = \sqrt{\rho}\sigma_c\left(S + T - \frac{1+\rho}{\sqrt{\rho}}I\right) = \sqrt{\rho}\sigma_c(S+T) - (1+\rho).$$

Next, we show that $\sigma_c(S+T) = [-2,2]$. Since $\|S+T\| \leq 2$, it is clear that $\sigma_c(S+T) \subset [-2,2]$ [51, Prop. 7.19]. To establish set equality, we use [51, Prop. 7.39] which states that

$$\sigma_c(S+T) = \left\{\lambda\colon \inf_{\|\mathbf{h}\|=1} \|(S+T-\lambda I)\mathbf{h}\| = 0\right\}.$$

We choose the following sequence $\{\mathbf{h}^{(N)}(\phi)\colon N \in \mathbb{Z}^+\}$ parametrized by $\phi \in [0,1]$

$$\mathbf{h}^{(N)}(\phi) = \begin{cases} \frac{1}{\sqrt{N}}e^{j2\pi\phi n}, & n \in \{0,1,\ldots,N-1\} \\ 0, & \text{otherwise.} \end{cases}$$

Clearly, $\|\mathbf{h}^{(N)}\| = 1$. Furthermore, taking $\lambda = 2\cos(2\pi\phi)$, we get

$$\left\|(S+T-\lambda I)\mathbf{h}^{(N)}\right\| = \sqrt{\frac{3}{N}}.$$

This shows that $[-2,2] \subset \sigma_c(S+T)$. As a consequence, we have

$$\sigma_c\left(\tilde{D}^{-1}\tilde{Q}\right) = \left[-\left(1+\sqrt{\rho}\right)^2, -\left(1-\sqrt{\rho}\right)^2\right]$$

and hence (25) follows. We now show that $\zeta_1 > \sup\sigma_c(D^{-1}Q)$. It suffices to show that $\xi_1 > \sup\sigma_c(\tilde{D}^{-1}\tilde{Q})$. Substituting the expressions for $\xi_1 = -(1-c')(1-\frac{\rho}{c'})$ and $\sup\sigma_c(\tilde{D}^{-1}\tilde{Q}) = -(1-\sqrt{\rho})^2$, and canceling common terms on both sides, we need to show that

$$\frac{c'}{\sqrt{\rho}} + \frac{\sqrt{\rho}}{c'} > 2.$$

This holds because $\gamma_1(\xi_1) = \frac{c'}{\sqrt{\rho}} < 1$. This completes the proof.

Appendix II
Proof of Theorem 4

It is easy to compute s_0 and s_1 using orthogonality by post-multiplying both sides of (31) by $\tilde{\mathbf{w}}(y_i)$. The proper weights are then obtained through (29) and (30). Getting $s(y)$ is slightly more involved. First, we right multiply (31) by $\tilde{D}^{-1}$, and then take the double z-transform of both sides. After rearranging terms, we deduce that it is equivalent to show that

$$\frac{1}{1-z_1z_2} - \frac{c}{\upsilon} - s_0\tilde{W}(y_0,z_1)\tilde{W}(y_0,z_2) - s_1\tilde{W}(y_1,z_1)\tilde{W}(y_1,z_2) = \int_{-1}^{1} s(y)\tilde{W}(y,z_1)\tilde{W}(y,z_2)dy \quad (36)$$

where we know from the definition of $\tilde{D}$ that

$$\sum_{k=0}^{\infty}\sum_{l=0}^{\infty} z_1^k z_2^l \left[\tilde{D}^{-1}\right]_{kl} = \frac{1}{1-z_1z_2} - \frac{c}{\upsilon}.$$

It can be shown that the right-hand side of (36) is the contour integral of a complex integrand over the unit circle [52]. To prove this, we denote the right-hand side of (36) by $H(z_1,z_2)$ and substitute $y=\cos\theta$ to get

$$\begin{aligned}
&H(z_1,z_2)\\
&\quad= \frac{2}{\pi}\int_0^{\pi}\left(1-\frac{c}{a}\right)\frac{\sin^2\theta}{\sin^2\theta+((p-1)\cos\theta+q))^2}\\
&\qquad\times\tilde{W}(\cos\theta,z_1)\,\tilde{W}(\cos\theta,z_2)\,d\theta\\
&\quad= \frac{1}{\pi}\int_0^{2\pi}\left(1-\frac{c}{a}\right)\mathrm{Re}\left[\frac{\sin\theta}{\sin\theta-i\left((p-1)\cos\theta+q)\right)}\right]\\
&\qquad\times\tilde{W}(\cos\theta,z_1)\,\tilde{W}(\cos\theta,z_2)\,d\theta\\
&\quad= \frac{1}{\pi}\int_0^{2\pi}\left(1-\frac{c}{a}\right)\frac{\sin\theta}{\sin\theta-i\left((p-1)\cos\theta+q)\right)}\\
&\qquad\times\tilde{W}(\cos\theta,z_1)\,\tilde{W}(\cos\theta,z_2)\,d\theta.
\end{aligned}$$

The last equality follows from $\tilde{W}(\cos\theta,z)$ and $\mathrm{Im}\left(\frac{\sin\theta}{\sin\theta-i((p-1)\cos\theta+q))}\right)$ being respectively, even and odd in θ. Substituting $t=e^{j\theta}$ in the equation above, we get

$$H(z_1,z_2) = \frac{1}{i\pi}\oint\frac{dt}{t}\tilde{S}(t)\tilde{W}\left(\frac{t^2+1}{2t},z_1\right)\tilde{W}\left(\frac{t^2+1}{2t},z_2\right) \tag{37}$$

where we have used

$$\tilde{S}(t) = \left(1-\frac{c}{a}\right)\frac{t^2-1}{p\left(t-\sqrt{\rho}\right)\left(t-\frac{c'}{\sqrt{\rho}}\right)}$$

$$\begin{aligned}
\tilde{W}\left(\frac{t^2+1}{2t},z\right) &= \left(1-\sqrt{\frac{a}{a-c}}\right)\\
&\quad+\sqrt{\frac{a}{a-c}}\left(\frac{\left(p(t^2+1)+2qt\right)z-2t}{2(t-z)(zt-1)}\right).
\end{aligned}$$

We need to show that $H(z_1,z_2)$ is identical to the left-hand side of (36). To do so, we employ complex integration, residue theory, and the fact that

$$\begin{aligned}
\tilde{W}\left(\frac{t^2+1}{2t},z\right)\Big|_{t=\sqrt{\rho}} &= \tilde{W}(y_0,z)\\
\tilde{W}\left(\frac{t^2+1}{2t},z\right)\Big|_{t=\frac{c'}{\sqrt{\rho}}} &= \tilde{W}(y_1,z).
\end{aligned}$$

We find the residues pertinent to the integrand as

$$\begin{aligned}
\mathrm{Res}\left[\frac{1}{t}\tilde{S}(t)\tilde{W}\left(\frac{t^2+1}{2t},z\right),t=z\right] &= \frac{1}{2}\sqrt{\frac{a-c}{a}}\\
\mathrm{Res}\left[\frac{1}{t}\tilde{S}(t),t=\sqrt{\rho}\right] &= -\frac{s_0}{2}\\
\mathrm{Res}\left[\frac{1}{t}\tilde{S}(t),t=\frac{c'}{\sqrt{\rho}}\right] &= -\frac{s_1}{2}.
\end{aligned}$$

From the residue theorem, we obtain

$$\begin{aligned}
&H(z_1,z_2)\\
&\quad= \lim_{t\to 0}\left(2\tilde{S}(0)\tilde{W}\left(\frac{t^2+1}{2t},z_1\right)\tilde{W}\left(\frac{t^2+1}{2t},z_2\right)\right)\\
&\qquad+\sqrt{\frac{a-c}{a}}\left(\tilde{W}\left(\frac{z_1^2+1}{2z_1},z_2\right)+\tilde{W}\left(\frac{z_2^2+1}{2z_2},z_1\right)\right)\\
&\qquad- s_0\tilde{W}(y_0,z_1)\tilde{W}(y_0,z_2) - s_1\tilde{W}(y_1,z_1)\tilde{W}(y_1,z_2).
\end{aligned}$$

We can verify through algebraic manipulation that

$$\begin{aligned}
&\lim_{t\to 0}\left(2\tilde{S}(0)\tilde{W}\left(\frac{t^2+1}{2t},z_1\right)\tilde{W}\left(\frac{t^2+1}{2t},z_2\right)\right)\\
&\quad+\sqrt{\frac{a-c}{a}}\left(\tilde{W}\left(\frac{z_1^2+1}{2z_1},z_2\right)+\tilde{W}\left(\frac{z_2^2+1}{2z_2},z_1\right)\right)\\
&\qquad= \frac{1}{1-z_1z_2}-\frac{c}{\upsilon}
\end{aligned}$$

and, as such, the desired result follows. That is, (36) holds and hence the identity expression of (31) is valid.

Appendix III
Proof of Theorem 5

Using the Chapman–Kolmogorov equation given in (21) and the identity expression of (31), we can rewrite (23) as

$$\begin{aligned}
\pi(x) &= E\exp\left(x\frac{c\mu}{a(c-\upsilon)}\tilde{D}^{-1}\tilde{Q}\right)\\
&\quad\times\left(s_0\tilde{\mathbf{w}}(y_0)\tilde{\mathbf{w}}(y_0)^* + s_1\tilde{\mathbf{w}}(y_1)\tilde{\mathbf{w}}(y_1)^*\right.\\
&\qquad\left.+\int_{-1}^{1}s(y)\tilde{\mathbf{w}}(y)\tilde{\mathbf{w}}(y)^*dy\right)\tilde{D}E^{-1}\pi(0).
\end{aligned}$$

For this system to be stable, we need $\lim_{x\to\infty}\pi(x)=0$ and $\pi(0,n)=p_n$ for $n\in\mathbb{N}$, since at steady state the probability of the buffer being empty is zero for the states where the input rate c exceeds the service rate υ. These boundary conditions imply that $\tilde{\mathbf{w}}(y_0)^*\tilde{D}E^{-1}\pi(0)=0$, which can be employed to obtain $\pi(0,0)$

$$\pi(0,0)\sqrt{\frac{a}{a-c}}\left(\frac{\upsilon}{c-\upsilon}\right) = \frac{ap_0}{a-c}\left(\tilde{W}(y_0,\sqrt{\rho})-1\right).$$

Since $\tilde{D}E^{-1}\pi(0)$ is almost a geometric sequence, the expression $\tilde{\mathbf{w}}(y)^*\tilde{D}E^{-1}\pi(0)$ is closely related to the z-transform of $\tilde{\mathbf{w}}(y)$. In view of the discussion at the beginning of Section VI-B2, it is not too surprising to find that $\tilde{\mathbf{w}}(y)^*\tilde{D}E^{-1}\pi(0)$ is constant for all $y\in\{y_1\}\cup[-1,1]$; and it is equal to

$$\begin{aligned}
&\tilde{\mathbf{w}}(y)^*\tilde{D}E^{-1}\pi(0)\\
&\quad= -\pi(0,0)\sqrt{\frac{a}{a-c}}\left(\frac{\upsilon}{c-\upsilon}\right)+\frac{ap_0}{a-c}\left(\tilde{W}(y,\sqrt{\rho})-1\right)\\
&\quad= p_0\left(\frac{a}{a-c}\right)^{\frac{3}{2}}\left(\frac{c'}{1-c'}-\frac{\rho}{1-\rho}\right) = -\frac{p_0}{s_0}\sqrt{\frac{a}{a-c}} > 0.
\end{aligned}$$

This, along with the fact that $\zeta = \frac{c\mu}{a(c-\nu)}\xi$, implies

$$\pi(x) = -\frac{p_0}{s_0}\sqrt{\frac{a}{a-c}}E\bigg(e^{\zeta_1 x}s_1\tilde{\mathbf{w}}(y_1) + \int_{-1}^{1} e^{\zeta(y)x}s(y)\tilde{\mathbf{w}}(y)dy\bigg).$$

We can get an expression for the probability of the buffer exceeding a fixed threshold x, using the relationship $\Pr(\ell_2 > x) = \sum_{n=0}^{\infty}\pi(x,n) = \langle\pi(x), \mathbf{1}\rangle$. Noting that

$$\langle E\tilde{\mathbf{w}}(y), \mathbf{1}\rangle = \sqrt{\frac{a-c}{a}} + \sqrt{\frac{a}{a-c}}\left(\frac{c'}{1-c'}\right) > 0$$

for all $y \in \{y_1\} \cup [-1, 1]$, we obtain

$$\Pr(\ell_2 > x) = -\frac{p_0}{s_0}\left(1 + \frac{ac'}{(a-c)(1-c')}\right) \times \left(s_1 e^{\zeta_1 x} + \int_{-1}^{1} s(y)e^{\zeta(y)x}dy\right).$$

This is the desired expression.

Appendix IV
Proof of Theorem 7

If the second buffer in the tandem queue described in Theorem 5 has a QoS constraint θ_0 on the asymptotic decay rate of buffer-occupancy, then we must have

$$\theta_0 \le \theta_2.$$

This condition enables us to determine the set $\mathcal{A}_2(\theta_0, c, \nu)$ of *on*-time arrival rates that can be supported by this queue under QoS constraint θ_0, and for a given service rate ν.

First, we gather that, for $\nu_1 \le \nu_2, \mathcal{A}_2(\theta_0, c, \nu_1) \subseteq \mathcal{A}_2(\theta_0, c, \nu_2)$. This fact follows directly from stochastic majorization of the buffer-content processes. Second, if the constant service rate ν of the second queue is greater than or equal to the constant service rate c of the first queue, then the second queue always remains empty (equivalently, $\theta_2 = \infty$). Under such circumstances, the second queue does not limit peak rate a. We therefore focus on the case where $\nu \in [0, c)$. If $a \in [0, \nu)$, then both queues stay empty with $\theta_1 = \theta_2 = \infty$; thus, we have $[0, \nu) \subseteq \mathcal{A}_2(\theta_0, c, \nu)$ for all $\theta_0 \ge 0$ and all $0 \le \nu < c$.

Let us define $\bar{a}_1(\theta_0, \nu) = \nu\left(1 + \frac{\mu}{\lambda + \theta_0\nu}\right)$. Note that $\bar{a}_1(\theta_0, \nu)$ is increasing in ν. Let ν_c be the service rate such that $\bar{a}_1(\theta_0, \nu_c) = c$. For $\nu \in [0, \nu_c]$, we can show that $[0, \bar{a}_1(\theta_0, \nu)] \subseteq \mathcal{A}_2(\theta_0, c, \nu)$. When $a \in [\nu, \bar{a}_1(\theta_0, \nu)] \subseteq [\nu, c]$, the first queue remains empty and the arrival processes at the first and second buffers are pathwise identical. Therefore, by Theorem 1, we get $\theta_2 = \frac{\mu}{a-\nu} - \frac{\lambda}{\nu}$. Since $a \le \bar{a}_1(\theta_0, \nu)$, it follows that $\theta_2 \ge \theta_0$ and hence $[0, \bar{a}_1(\theta_0, \nu)] \subseteq \mathcal{A}_2(\theta_0, c, \nu)$. It is clear that for any finite θ_0, $\nu_c < c$. We can explicitly write ν_c as

$$\nu_c = \begin{cases} \frac{1}{2}\left(c - \frac{\lambda+\mu}{\theta_0}\right)\left(1 + \sqrt{1 + \frac{4c\lambda\theta_0}{(c\theta_0 - \lambda - \mu)^2}}\right), & c \ge \frac{\lambda+\mu}{\theta_0} \\ \frac{1}{2}\left(c - \frac{\lambda+\mu}{\theta_0}\right)\left(1 - \sqrt{1 + \frac{4c\lambda\theta_0}{(c\theta_0 - \lambda - \mu)^2}}\right), & c < \frac{\lambda+\mu}{\theta_0}. \end{cases}$$

Next, we consider the case where $a \ge c$. In this scenario, the arrival rate at the second queue is $c(K(t))$, as described in Section VI-A. In the previous section, we obtained the exponential decay rate θ_2 governing the queue distribution for this system

$$\theta_2 = \begin{cases} \frac{\mu}{a-\nu} - \frac{\lambda}{\nu}, & \frac{a/c-1}{a/\nu-1} < \sqrt{\rho} \\ \frac{c\mu}{a(c-\nu)}(1-\sqrt{\rho})^2, & \frac{a/c-1}{a/\nu-1} \ge \sqrt{\rho}. \end{cases}$$

We note that $\bar{a}_1(\theta_0, \nu)$ denotes the maximum supportable rate when the discrete eigenvalue ζ_1 of $D^{-1}Q$ governs the decay rate of the buffer overflow probability, i.e., when $c' = \frac{a/c-1}{a/\nu-1} < \sqrt{\rho}$. Furthermore, the condition $\sqrt{\rho} > c'$ is equivalent to the quadratic expression $\rho - \frac{\nu}{c}\sqrt{\rho} + \frac{\lambda}{\mu}\left(1 - \frac{\nu}{c}\right) > 0$. This equation specifies upper and lower bounds on the existence of ζ_1. That is, ζ_1 exists only when

$$\sqrt{\rho} \notin \left[\frac{\nu}{2c}\left(1 - \sqrt{1 - \frac{4\lambda c}{\mu\nu}\left(\frac{c}{\nu} - 1\right)}\right), \frac{\nu}{2c}\left(1 + \sqrt{1 - \frac{4\lambda c}{\mu\nu}\left(\frac{c}{\nu} - 1\right)}\right)\right].$$

We call the endpoints of the interval $\sqrt{\rho_l}$ and $\sqrt{\rho_u}$, respectively. We have $\rho = \frac{\lambda}{\mu}\left(\frac{a}{c} - 1\right)$, which gives upper and lower bounds $a_{u1}(\nu) = c\left(1 + \frac{\mu}{\lambda}\rho_u\right)$ and $a_{l1}(\nu) = c\left(1 + \frac{\mu}{\lambda}\rho_l\right)$ in terms of ρ_u and ρ_l, respectively. Thus, when $a \in \mathcal{A}_2(\theta_0, c, \nu)$ lies between these two bounds, the supremum of the continuous spectrum dominates the tail-asymptotics and the discrete eigenvalue ζ_1 disappears. We can write $a_{u1}(\nu)$ and $a_{l1}(\nu)$ explicitly

$$a_{u1}(\nu) = \nu\left(1 + \frac{\nu\mu}{2c\lambda}\left(1 + \sqrt{1 - 4\left(\frac{c}{\nu} - 1\right)\frac{c\lambda}{\nu\mu}}\right)\right) \quad (38)$$

$$a_{l1}(\nu) = \nu\left(1 + \frac{\nu\mu}{2c\lambda}\left(1 - \sqrt{1 - 4\left(\frac{c}{\nu} - 1\right)\frac{c\lambda}{\nu\mu}}\right)\right). \quad (39)$$

For the real interval $[a_{l1}, a_{u1}]$ to exist, the discriminant in (38) and (39) must be nonnegative. Defining

$$\nu' = \frac{2c}{1 + \sqrt{1 + \frac{\mu}{\lambda}}}$$

we deduce that a_{u1} and a_{l1} are well-defined real numbers provided that $\nu > \nu'$. That is, discrete eigenvalue ζ_1 exists for all $\nu \le \nu'$. In this case, the achievable rate region is limited by $\bar{a}_1(\theta_0, \nu)$. Note that $\nu' < c$. Clearly

$$1 - 4\left(\frac{c}{\nu} - 1\right)\frac{c\lambda}{\nu\mu} = (1 + \frac{\lambda}{\mu}) - \frac{\lambda}{\mu}\left(\frac{2c}{\nu} - 1\right)^2$$

is an increasing function that maps $\nu \in [\nu', c]$ to $[0, 1]$. Therefore, $\sqrt{\rho_u}$ is an increasing function of ν. Furthermore, because $\sqrt{\rho_u}\sqrt{\rho_l} = \frac{\lambda}{\mu}\left(1 - \frac{\nu}{c}\right)$ is a decreasing function of ν, it follows that $\sqrt{\rho_l}$ is a decreasing function of ν in $[\nu', c]$. Since ρ_u and ρ_l

are both nonnegative in this interval, it follows that a_{u1} monotonically increases and ranges over $\left[\nu'\left(1+\frac{\nu'\mu}{2c\lambda}\right), c\left(1+\frac{\mu}{\lambda}\right)\right]$, whereas a_{l1} decreases and ranges over $\left[\nu'\left(1+\frac{\nu'\mu}{2c\lambda}\right), c\right]$.

When $a \in [a_{l1}(\nu), a_{u1}(\nu)]$, the continuous spectrum dominates the tail asymptotics. This implies

$$\left((c-\nu)\frac{\theta_0}{\lambda}-1\right)\rho+2\sqrt{\rho}+\left((c-\nu)\frac{\theta_0}{\mu}-1\right)\leq 0.$$

Therefore, since $\sqrt{\rho}\geq 0$, we know that $\sqrt{\rho}$ belongs to the interval

$$\left[0, \frac{-1+\sqrt{1-\left((c-\nu)\frac{\theta_0}{\lambda}-1\right)\left((c-\nu)\frac{\theta_0}{\mu}-1\right)}}{(c-\nu)\frac{\theta_0}{\lambda}-1}\right]$$

when $(c-\nu)\frac{\theta_0}{\lambda}>1$; and it belongs to the set

$$\left[0, \frac{1-\sqrt{1-\left((c-\nu)\frac{\theta_0}{\lambda}-1\right)\left((c-\nu)\frac{\theta_0}{\mu}-1\right)}}{1-(c-\nu)\frac{\theta_0}{\lambda}}\right] \bigcup \left[\frac{1+\sqrt{1-\left((c-\nu)\frac{\theta_0}{\lambda}-1\right)\left((c-\nu)\frac{\theta_0}{\mu}-1\right)}}{1-(c-\nu)\frac{\theta_0}{\lambda}}, \infty\right)$$

when $(c-\nu)\frac{\theta_0}{\lambda}<1$. We emphasize that we require $(c-\nu)\frac{\theta_0}{\mu}\leq 1$ in the first case for a nonnegative $\sqrt{\rho}$ to exist. Furthermore, we need $\nu\in[c-\frac{\lambda+\mu}{\theta_0}, c]$ for a real $\sqrt{\rho}$ to exist. For the second queue to be stable, we must have $a<\nu\left(1+\frac{\nu}{\lambda}\right)$, which implies $a\leq \bar{a}_3(\theta_0,c,\nu)$ where

$$\begin{aligned}\bar{a}_3(\theta_0,c,\nu) &= \sup\left\{a\leq c\left(1+\frac{\mu}{\lambda}\right): \right. \\ &\qquad \left. \theta_0\leq \frac{c\mu}{a(c-\nu)}\left(1-\sqrt{\left(\frac{a-c}{c}\right)\frac{\lambda}{\mu}}\right)^2\right\}, \\ &= c+\frac{c\mu}{\lambda}\left(\frac{\sqrt{1-\left((c-\nu)\frac{\theta_0}{\mu}-1\right)\left((c-\nu)\frac{\theta_0}{\lambda}-1\right)}-1}{(c-\nu)\frac{\theta_0}{\lambda}-1}\right)^2.\end{aligned}$$

It is clear from the explicit form of $\bar{a}_3(\theta_0,c,\nu)$ that it is increasing in ν. We can write the achievable rate region $\mathcal{A}_2(\theta_0,c,\nu)$ in terms of $\bar{a}_1$, a_{u1}, a_{l1} and $\bar{a}_3$ as

$$\begin{aligned}\mathcal{A}_2(\theta_0,c,\nu) = &\left\{a\in\mathbb{R}^+/[a_{l1}(\nu),a_{u1}(\nu)]\colon a\leq \bar{a}_1(\theta_0,\nu)\right\} \\ &\cup\left\{a\in[a_{l1}(\nu),a_{u1}(\nu)]\colon a\leq \bar{a}_3(\theta_0,c,\nu)\right\}.\end{aligned}$$

Let ν^* be the value of ν where $\bar{a}_1(\theta_0,\nu)$ intersects a_{l1} or a_{u1}. In other words, if we substitute $\bar{a}_1$ for a in the expression for ρ, the value of ν that equates $\sqrt{\rho}$ and c' is ν^*. It can be seen that $\bar{a}_3(\theta_0,c,\nu^*)=\bar{a}_1(\theta_0,\nu^*)$ by substituting $\lambda=\frac{\mu(a/c-1)}{(a/\nu^*-1)^2}$ and $\sqrt{\rho}=\frac{a/c-1}{a/\nu^*-1}$ to obtain

$$\theta_0=\frac{\bar{a}_1\mu(c-\nu^*)}{c(\bar{a}_1-\nu^*)^2}=\frac{\bar{a}_3\mu(c-\nu^*)}{c(\bar{a}_3-\nu^*)^2}$$

where $a=\bar{a}_1(\theta_0,\nu^*)=\bar{a}_3(\theta_0,c,\nu^*)$. The value ν^* can be written as the positive root of the following equation in the interval $[0,c]$:

$$\theta_0(c-\nu)=\frac{(\theta_0\nu)^2\mu}{\lambda\mu+(\lambda+\theta_0\nu)^2}.$$

The left-hand side is continuous and monotonically decreasing for $\nu\in[0,c]$ and ranges over $[c\theta_0,0]$, whereas the right-hand side is continuous and monotonically increasing in the same interval and ranges over $\left[0,\frac{(c\theta_0)^2\mu}{\lambda\mu+(\lambda+c\theta_0)^2}\right]$. It is clear from the continuity and the monotonicity of these functions that there exists a unique real $\nu^*\in[0,c]$ for all values of λ, μ, c, $\theta_0>0$. Furthermore, since $\nu>\nu'$ for a_{u1} and a_{l1} to exist, we have $\nu^*\geq\nu'$. In addition, since $\left((c-\nu)\frac{\theta_0}{\lambda}-1\right)\left((c-\nu)\frac{\theta_0}{\mu}-1\right)$ is decreasing in ν, we can prove that it is less than or equal to 1 for all $\nu\in[\nu^*,c]$ by showing it to be less than or equal to 1 at $\nu=\nu^*$. This is equivalent to showing $(c-\nu^*)\theta_0\in[0,\lambda+\mu]$. Therefore, we only need to show that equation

$$\frac{(\theta_0\nu^*)^2\mu}{\lambda\mu+(\lambda+\theta_0\nu^*)^2}\leq\lambda+\mu$$

is valid, which is equivalent to $\lambda(\theta_0\nu^*+\lambda+\mu)^2\geq 0$ and obviously true. Hence, the result holds. We know from Theorem 3 that discrete eigenvalue ζ_1 is always greater than the supremum of the continuous spectrum of $D^{-1}Q$, whenever it exists. Therefore

$$\frac{\mu}{a\left(1-\frac{\nu}{c}\right)}\left(1-\sqrt{\left(\frac{a}{c}-1\right)\frac{\lambda}{\mu}}\right)^2\geq\frac{\mu}{a-\nu}-\frac{\lambda}{\nu}$$

which in turn implies $c\left(1+\frac{\mu}{\lambda}\right)\geq\bar{a}_3(\theta_0,c,\nu)\geq\bar{a}_1(\theta_0,\nu)$.

If $\bar{a}_1$ intersects a_{u1} and $\bar{a}_3$ at ν^*, then $\bar{a}_1>a_{u1}>a_{l1}$ for all $\nu<\nu^*$. Since, a_{l1} is decreasing, we have $\bar{a}_1>a_{l1}$ for $\nu\in[\nu^*,c]$ as well. Also, since $\bar{a}_1$ and a_{u1} intersect uniquely at ν^* in $[\nu',c]$ and are both increasing, $\bar{a}_1<a_{u1}$ within this interval. Thus, we have

$$\begin{aligned}\mathcal{A}_2(\theta_0,c,\nu) = &\{a\in\mathbb{R}^+\colon a\leq\bar{a}_1(\theta_0,\nu),\nu<\nu^*\} \\ &\cup\{a\in\mathbb{R}^+\colon a\leq\bar{a}_3(\theta_0,c,\nu),\nu\geq\nu^*\}.\end{aligned}$$

Otherwise, when $\bar{a}_1$ intersects a_{l1} and $\bar{a}_3$ at ν^*, $\bar{a}_3<a_{l1}<a_{u1}$ for all $\nu<\nu^*$ because $\bar{a}_3$ is increasing in ν. Since a_{l1} is decreasing, $\bar{a}_1>a_{l1}$ for $\nu\in(\nu^*,c]$. In this interval, $\bar{a}_1$ and a_{u1} do not intersect and are both increasing with $\bar{a}_1<a_{u1}$. We conclude that

$$\begin{aligned}\mathcal{A}_2(\theta_0,c,\nu) = &\{a\in\mathbb{R}^+\colon a\leq\bar{a}_1(\theta_0,\nu),\nu<\nu^*\} \\ &\cup\{a\in\mathbb{R}^+\colon a\leq\bar{a}_3(\theta_0,c,\nu),\nu\geq\nu^*\}.\end{aligned}$$

From the discussion above, it is clear that we have $a_{l1}(\nu) < \bar{a}_1(\theta_0,\nu) < a_{u1}(\nu)$ for $\nu > \nu^*$. Also, it was shown that $\bar{a}_3(\theta_0,c,\nu^*) = \bar{a}_1(\theta_0,\nu^*)$, and therefore we conclude that $\mathcal{A}_2(\theta_0,c,\nu) = [0,\bar{a}_2(\theta_0,c,\nu)]$ where the maximum achievable rate $\bar{a}_2(\theta_0,c,\nu)$ is continuous in ν. This rate can therefore be characterized completely for all $\nu \in [0,c]$ as

$$\bar{a}_2(\theta_0,c,\nu) = \begin{cases} \bar{a}_1(\theta_0,\nu) & \nu \in [0,\nu^*] \\ \bar{a}_3(\theta_0,c,\nu) & \nu \in [\nu^*,c]. \end{cases}$$

This is the desired result.

ACKNOWLEDGMENT

The authors would like to thank Dr. S. Aalto for providing them a preprint of his paper. The authors would also like to thank Dr. T. Schlumprecht, Dr. N. Sivakumar, and Dr. G. Berkolaiko for helpful comments and suggestions. They would also like to thank Dr. A. Sprintson for his comments on network coding. Finally, they would like to thank the anonymous reviewers for their helpful comments and suggestions, which have significantly improved the quality and presentation of this paper.

REFERENCES

[1] R. Ahlswede and I. Csiszár, "To get a bit of information may be as hard as to get full information," *IEEE Trans. Inf. Theory*, vol. IT-27, no. 4, pp. 398–408, Jul. 1981.

[2] R. Koetter and M. Médard, "An algebraic approach to network coding," *IEEE/ACM Trans. Netw.*, vol. 11, pp. 782–795, Oct. 2003.

[3] S. Jaggi, P. Sanders, P. A. Chou, M. Effros, S. Egner, K. Jain, and L. M. G. M. Tolhuizen, "Polynomial time algorithms for multicast network code construction," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1973–1982, Jun. 2005.

[4] T. C. Ho, Y.-H. Chang, and K. J. Han, "On constructive network coding for multiple unicasts," in *Proc. 44th Allerton Conf. Commun. Control Comput.*, Sep. 2006.

[5] A. Khreishah, C.-C. Wang, and N. B. Shroff, "Optimization based rate control for communication networks with inter-session network coding," in *Proc. 27th IEEE Conf. Comput. Commun.*, Apr. 2008, pp. 81–85.

[6] R. Ahlswede, N. Cai, R. Li S.-Y, and R. W. Yeung, "Network information flow," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1204–1216, Jul. 2000.

[7] F. P. Kelly, "Effective bandwidths at multi-class queues," *Queueing Syst.*, vol. 9, pp. 5–15, 1991.

[8] G. Kesidis, J. C. Walrand, and C.-S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Trans. Netw.*, vol. 1, pp. 424–428, Aug. 1993.

[9] A. I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Netw.*, vol. 1, pp. 329–343, Jun. 1993.

[10] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Trans. Autom. Control*, vol. 39, no. 5, pp. 913–931, May 1994.

[11] C.-S. Chang, "Sample path large deviations and intree networks," *Queueing Syst.*, vol. 20, no. 1-2, pp. 7–36, Mar. 1995.

[12] W. Whitt, "Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues," *Telecommun. Syst.*, vol. 2, no. 1, pp. 71–107, Dec. 1993.

[13] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high speed digital networks," *IEEE J. Sel. Areas Commun.*, vol. 13, pp. 1091–1100, Aug. 1995.

[14] C.-S. Chang and T. Zajic, "Effective bandwidths of departure processes from queues with time varying capacities," in *Proc. INFOCOM*, 1995, pp. 1001–1009.

[15] R. Guérin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high speed networks," *IEEE J. Sel. Areas Commun.*, vol. 9, pp. 968–981, Sep. 1991.

[16] G. de Veciana, G. Kesidis, and J. C. Walrand, "Resource management in wide-area ATM networks using effective bandwiths," *IEEE J. Sel. Areas Commun.*, vol. 13, pp. 1081–1090, 1995.

[17] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, pp. 630–643, Jul. 2003.

[18] L. Liu, P. Parag, J. Tang, W.-Y. Chen, and J.-F. Chamberland, "Resource allocation and quality of service evaluation for wireless communication systems using fluid models," *IEEE Trans. Inf. Theory*, vol. 53, no. 5, pp. 1767–1777, May 2007.

[19] S. Bhadra and S. Shakkottai, "Looking at large networks: Coding vs. Queueing," in *Proc. 25th IEEE Conf. Comput. Commun.*, Apr. 2006, pp. 1–12.

[20] A. Eryilmaz, A. Ozdaglar, and M. Médard, "On delay performance gains from network coding," in *Proc. 40th Annu. Conf. Inf. Sci. Syst.*, Mar. 2006, pp. 864–870.

[21] B. Shrader and A. Ephremides, "A queueing model for random linear coding," in *Proc. Military Commun. Conf.*, Oct. 2007, pp. 1–7.

[22] B. Shrader and A. Ephremides, "On the queueing delay of a multicast erasure channel," in *Proc. IEEE Inf. Theory Workshop*, Oct. 2006, pp. 423–427.

[23] J. K. Sundararajan, D. Shah, and M. Médard, "Arq for network coding," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2008, pp. 1651–1655.

[24] P. Chaporkar and A. Proutiere, "Adaptive network coding and scheduling for maximizing throughput in wireless networks," in *Proc. 13th Annu. ACM Int. Conf. Mobile Comput. Netw.*, Sep. 2007, pp. 135–146.

[25] F. Xue, C.-H. Liu, and S. Sandhu, "MAC-layer and PHY-layer network coding for two-way relaying: Achievable rate regions and opportunistic scheduling," in *Proc. 44th Annu. Allerton Conf.*, Sep. 2007, pp. 396–402.

[26] X. He and A. Yener, "On the energy-delay trade-off of a two-way relay network," in *Proc. 42nd Annu. Conf. Inf. Sci. Syst.*, Mar. 2008, pp. 865–870.

[27] C.-H. Liu and F. Xue, "Network coding for two-way relaying: Rate region, sum rate and opportunistic scheduling," in *Proc. IEEE Int. Conf. Commun.*, May 2008, pp. 1044–1049.

[28] Y. E. Sagduyu and A. Ephremides, "Cross-layer optimization of MAC and network coding in wireless queueing tandem networks," *IEEE Trans. Inf. Theory*, vol. 54, no. 2, pp. 554–571, Feb. 2008.

[29] A. Eryilmaz and D. S. Lun, "Control for inter-session network coding," in *Proc. Workshop Netw. Coding Theory Appl.*, Jan. 2007.

[30] C. Rago, P. Willett, and Y. Bar-Shalom, "Censoring sensors: A low-communication rate scheme for distributed detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 32, pp. 554–568, Apr. 1996.

[31] R. J. Gibbens and P. J. Hunt, "Effective bandwidths for the multi-type UAS channel," *Queueing Syst.*, vol. 9, pp. 17–28, 1991.

[32] L. Liu, P. Parag, and J.-F. Chamberland, "Quality of service analysis for wireless user-cooperation networks," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3833–3842, Oct. 2007.

[33] D. P. Kroese and W. R. W. Scheinhardt, "Joint distributions for interacting fluid queues," *Queueing Syst.*, vol. 37, no. 1–3, pp. 99–139, Mar. 2001.

[34] S. Aalto and W. R. W. Scheinhardt, "Tandem fluid queues fed by homogeneous on–off sources," *Oper. Res. Lett.*, vol. 27, no. 2, pp. 73–82, Sep. 2000.

[35] N. Barbot and B. Sericola, "Exact stationary solution to tandem fluid queues," *Int. J. Simul.: Syst. Sci. Technol.*, vol. 4, no. 5–6, pp. 12–20, 2003.

[36] D. Mitra, "Stochastic theory of a fluid model of producers and consumers coupled by a buffer," *Adv. Appl. Prob.*, vol. 20, pp. 646–676, 1993.

[37] D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic theory of a data-handling system with multiple sources," *Bell Syst. Tech. J.*, vol. 61, no. 8, pp. 1871–1894, Oct. 1982.

[38] J. Virtamo and I. Norros, "Fluid queue driven by an M/M/1 queue," *Queueing Syst.*, vol. 16, no. 3–4, pp. 373–386, Sep. 1994.

[39] S. Aalto, "Characterization of the output rate process for a Markovian storage model," *J. Appl. Probab.*, vol. 35, no. 1, pp. 184–199, Mar. 1998.

[40] S. Aalto, "Output from an A-M-S type fluid queue," *Fundam. Role Teletraffic Evol. Telecommun. Netw.*, pp. 421–430, 1997.

[41] T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 2001.

[42] T. E. Stern and A. I. Elwalid, "Analysis of separable Markov-modulated rate models for information-handling systems," *Adv. Appl. Probab.*, vol. 23, pp. 105–139, 1991.

[43] S. Asmussen, *Applied Probability and Queues*, ser. Stochastic Modelling and Applied Probability, 2nd ed. New York: Springer-Verlag, 2003, vol. 51.
[44] J. R. Norris, *Markov Chains*, ser. Statistical and Probabilistic Mathematics. Cambridge, U.K.: Cambridge Univ. Press, 1998.
[45] M. Reed and B. Simon, *Functional Analysis: Volume IV*, ser. Methods of Modern Mathematical Physics, revised ed. New York: Academic, 1980.
[46] R. Carmona and J. Lacroix, *Spectral Theory of Random Schrödinger Operators*, ser. Probability and Its Applications. Boston, MA: Birkhäuser, 1990.
[47] M. Rubinovitch, "The ouput of a buffered data communication system," *Stochastic Processes Appl.*, vol. 1, pp. 375–382, 1973.
[48] O. Kella and W. Whitt, "A storage model with a two-state random environment," *Oper. Res.*, vol. 40, no. S2, pp. 257–262, May-Jun. 1992.
[49] O. J. Boxma and V. Dumas, "The busy period in the fluid queue," in *Proc. ACM SIGMETRICS Joint Int. Conf. Meas. Model. Comput. Syst.*, 1998, pp. 100–110.
[50] V. G. Kulkarni, "Effective bandwidths for Markov regenerative sources," *Queueing Syst.*, vol. 24, no. 1–4, pp. 137–153, Mar. 1996.
[51] M. Fabian, P. Habala, P. Hájek, V. M. Santalucía, J. Pelant, and V. Zizler, *Functional Analysis and Infinite-Dimensional Geometry*, ser. CMS Books in Mathematics. New York: Springer-Verlag, 1990.
[52] R. V. Churchill and J. W. Brown, *Complex Variables and Applications*, ser. Higher Mathematics, 5th ed. New York: McGraw-Hill, 1989.

Parimal Parag (S'05) received a dual degree (B. Tech. and M. Tech.) in electrical engineering, with specialization in communication systems, from Indian Institute of Technology, Madras, India, in 2004. He is currently working towards the Ph.D. degree at Texas A&M University, College Station.

He was at Los Alamos National Laboratory in summer 2007. His research interests include communication systems and networks, queueing theory, applied probability, estimation and detection theory and optimization methods.

Jean-Francois Chamberland (S'98–M'04–SM'09) received the B.S. degree from McGill University, Montreal, QC, Canada, in 1998, the M.S. degree from Cornell University, Ithaca, NY, in 2000, and the Ph.D. degree from the University of Illinois at Urbana-Champaign, Urbana, in 2004.

He joined Texas A&M University, College Station, in 2004, where he is currently an Assistant Professor in the Department of Electrical and Computer Engineering.

Dr. Chamberland has received a Young Author Best Paper Award from the IEEE Signal Processing Society, in 2006; and an Early Career Development (CAREER) Award from the National Science Foundation, in 2008.