Queueing Analysis of a Butterfly Network

Parimal Parag and Jean-François Chamberland
Department of Electrical and Computer Engineering
Texas A&M University, College Station, TX 77843-3128, USA
Email: {parimal, chmbrlnd}@tamu.edu

***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 have infinite backlogs. However, if a network supports delay-sensitive applications, traffic is often sparse 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 wireless resources in terms of bandwidth and transmit power. Together, these effects decrease the potential benefits of network coding. For real-time applications over wireless networks, it may be best not to combine information at the nodes.**

I. Introduction

Network coding is a new paradigm that has received much attention in the literature lately [1], [2], [3], [4]. It has the potential to improve the throughput and robustness of communication networks. This performance gain is achieved by relaxing the assumption 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, and 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 [5].

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, it 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 VoIP, video conferencing, gaming and electronic commerce are now commonly used by vanguardists on both wired and wireless devices. Future communication infrastructures are therefore expected to carry much larger volumes of data with varying quality of service (QoS) requirements. This paper seeks to provide preliminary answers to two important questions.

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, in part, by the concept of effective bandwidth, which identifies the data-rate needed by a source to fulfill its service requirement [6], [7]. The effective bandwidth of a source can be much larger than the average throughput it generates. The sparse traffic generated by delay-sensitive applications combined with the gains associated with statistical multiplexing act as to decrease the benefits of network coding. Therefore, it is not clear how much we gain with 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 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, the resources must be assigned as to create a suitable topology. While this enables efficient coding, there may be a non-negligible 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, but with physical resources allocated optimally in both cases. The second question we seek to address is the following. When is it relevant to create a topology suitable for network coding in a wireless environment?

These two questions are not only related through the rising popularity of real-time applications over wireless networks, their answers necessitate the development of analogous mathematical tools. This similarity motivates our joint treatment of these related topics. Specifically, we investigate the impact of network coding on the queueing behavior of wireless communication systems. We consider a simple scenario where

This work was supported, in part, by the National Science Foundation under the CAREER Award No. CCF-0747363.

978-1-4244-2571-6/08/$25.00

two varying rate sources communicate to multiple destinations through the notorious butterfly network. We analyze the performance of this system, and compute its achievable rate-region when the system operates under stringent service constraints. Due to the time-varying nature of arrivals and service, it is difficult to provide deterministic delay guarantees. Thus, we adopt a statistical QoS measure that captures the asymptotic decay-rate of buffer occupancy

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

where L is the equilibrium distribution of the buffer at the transmitter. The parameter θ reflects the perceived quality of the corresponding communication link: a larger θ implies a lower probability of violating the delay restriction or a tighter QoS constraint. This metric is closely tied to large-deviation theory and forms a basis for the concept of effective bandwidth, which has been studied extensively in the past [8], [9].

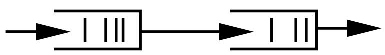

Fig. 1. Network with tandem queues

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 framework adopted in this paper, independent sources sharing the same link can be studied separately. This is one of the appealing properties of an analysis based on large deviations. The main challenge is to characterize the performance of the tandem network shown in Fig. 1. This network consists of two successive nodes where the input to the second node is the output of the preceding one. Due to page limitations, we can only state the results obtained for the tandem queue and apply them to study the butterfly network.

Fig. 2 shows the butterfly network we wish to study. We consider two distinct versions of the butterfly network. First, we analyze an AWGN limited network with constant and identical link capacities. This configuration is suitable for network coding and provides ground for our initial queueing comparison. Then, we examine a wireless network under a broadcast paradigm. In the latter case, we assume that network resources can be allocated among the various nodes to create non-identical links, thereby maximizing the performance of each configuration.

To compare the queueing performance of network coding versus classical routing, we compute the achievable rate regions for both cases. Not too surprisingly, network coding outperforms classical routing for a network with identical link capacities. Although statistical multiplexing had the potential to offset some of the coding gain, classical routing remains a distant second to network coding for all QoS requirements. The wireless butterfly offers more interesting results. In this case, combining packets at intermediate node 3 doesn't always offer throughput benefits, and sometimes may even be harmful. This phenomenon depends on the physical location of the nodes. It results from the fact that network coding needs symmetric links between link 1–3 and link 1–5 for the coding gains to be realized. If the link capacities are not equal on both these links, then destination C needs to wait for packets from the slower link to be able to decode the cross-traffic from source B.

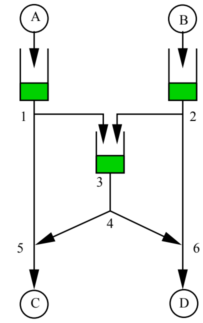

Fig. 2. Butterfly Network with Buffers of Interest

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. 2. A multicast scenario is considered, where independent sources A & B store their respective information in buffers at nodes 1 & 2, and must transmit their data to both destinations, C & D. To facilitate this, node 1 sends its packets to node 5 and a replica to the buffer at node 3. Similarly, node 2 forwards its packets to node 6 and to a buffer at node 3. Node 3 can take two courses of action; either it stores the packets from both sources and forward them successively to node 4 or it combines the packets algebraically before forwarding the data.

The first case will be called the *no coding* case. In this scenario, node 4 duplicates the received packets from node 3 and forwards copies to nodes 5 & 6. These destination nodes disregard the redundant information (or they could use it to improve the reliability of the previously received data) and retain the new information. For the second case, we consider a simple network coding scheme where node 3 adds the two streams of packets over $GF(2)$ and relays the coded packets to node 4 [5]. The latter duplicates the received packets and transmits them to nodes 5 & 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 it 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 not exceed θ_0 for any of the queues in the system.

For the sake of analysis, we assume that packets are infinitely divisible and hence arrival and service processes are fluid in nature. Accordingly, it becomes possible to define instantaneous arrival and service rates. Under this assumption, we have a single server fluid queue at all the nodes. We also take the buffers in the system to be arbitrary large. For finite buffers, a similar approach applies albeit with additional boundary conditions on the buffer occupancy.

A. Source Model

Many real-time traffic sources can be accurately represented by on-off sources [10]. 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 [11], [7], [8]. We postulate that source A is independent from source B and they both satisfy the following assumption.

Assumption 1: In an *on* state, the source emits packets at a constant peak rate into its corresponding data buffer, and remains idle otherwise. Moreover, the *off* and *on* times are independent and exponentially distributed.
The peak-rate for source at node $i \in \{1, 2\}$ is taken as a_i. Whereas, the mean *off* and *on* times are denoted by λ_i^{-1} and μ_i^{-1} respectively.

B. Queueing Model

We denote the capacity of link i–j by c_{ij}. That is, if there is a link between nodes i & j and the buffer associated to node i is non-empty, then this node 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 doesn't need to store any data. It only facilitates the forwarding of the packets it receives, to nodes 5 & 6. Hence, the buffer associated to node 4 is always empty.

Node 1 sends the same information to both nodes 3 & 5, and therefore retains data in its buffer until both receiving nodes have acquired the packet. 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 $c_2 = \min\{c_{23}, c_{26}\}$. More specifically, nodes 1, 2 and 3 transmit packets at rates c_1, c_2 and c_3, respectively, whenever their own buffers are non-empty. Note that, by construction, congestion can only occur at these 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. 2, for the butterfly network under consideration.

III. Key Results

In this section, we list the results needed to compute the achievable rate-regions for communicating through a butterfly network under specific QoS requirements. Let $L_1(t)$ be the amount of fluid at time t in an arbitrary large reservoir 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 ν. The amount of fluid in the latter buffer at time t is denoted by $L_2(t)$.

We wish to find maximum rate a, such that the QoS constraint θ_0 is satisfied by both queues. Let θ_1 and θ_2 be the asymptotic buffer decay-rate governing the first and second queues in the tandem network, where

$$\theta_j = -\lim_{x\to\infty} \frac{\ln \Pr\{L_j > x\}}{x} \tag{2}$$

and L_j is the steady-state distribution of the jth buffer in tandem. Specifically, we wish to find set A such that

$$A(\theta_0, c, \nu) = \{a \in \mathbb{R}^+ : \theta_0 \le \min\{\theta_1, \theta_2\}\}.$$

Here, $a \in \mathbb{R}^+$ is admissible with $a \in A(\theta_0, c, \nu)$ if and only if

$$a \le \begin{cases} a_1(\theta_0, \nu) & 0 < \nu \le \nu^* \\ \min\{a_1(\theta_0, c), a_{u2}(\theta_0, c, \nu)\} & \nu^* < \nu < c; \end{cases} \tag{3}$$

where $a_1(\theta, c) \triangleq c + c\mu/(\lambda + c\theta)$ and the expression for $a_{u2}(\theta, c, \nu)$ is

$$c\left[1 + \frac{\mu}{\lambda}\left(\frac{-1 + \sqrt{1 - [(c-\nu)\frac{\theta}{\mu} - 1][(c-\nu)\frac{\theta}{\lambda} - 1]}}{(c-\nu)\frac{\theta}{\mu} - 1}\right)^2\right].$$

Parameter ν^* is given implicitly by

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

IV. Achievable Rate Regions

The results listed above can be used to characterize the achievable rate regions for the butterfly network under consideration, with independent *on-off* sources A & B. Let θ_i be the asymptotic decay-rate of buffer-occupancy for buffer at nodes $i \in \{1, 2, 3\}$, and let a_1 and a_2 be the peak-arrival rates from sources A and B, respectively. 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.,

$$\mathcal{R} = \left\{(a_1, a_2) \in \mathbb{R}^+ \times \mathbb{R}^+ : \theta_0 \le \min\{\theta_1, \theta_2, \theta_3\}\right\}. \tag{4}$$

A. Network Coding

For network coding, packets on 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 the buffer at node 3 oblivious of the other flow. Buffer 3 can be serviced at a maximum rate c_3. However, to avoid any decoding delay at the destinations, the service rates offered at nodes 1, 2 and 3 are $c_1' = \min\{c_1, c_3\}, c_2' = \min\{c_2, c_3\}$ and $c_3' = \max\{c_1', c_2'\}$. Using the notation of the previous section, we can write the achievable rate-region $\mathcal{R}$ for this system as $\mathcal{R}_{\mathrm{nc}} = A_1(\theta_0, c_1', c_1') \times A_2(\theta_0, c_2', c_2')$, where A_i is the achievable rate region corresponding to the source at node i.

B. Classical Routing

For the case of classical routing, consider two parallel buffers at node 3 with positive 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. The aggregate fluid in both the buffers will be greater than or equal to that of a single buffer with incoming flows from nodes 1 & 2 and serviced at rate c_3. Thus, the LDP of a single buffer at node 3 would be larger than the LDP of the aggregate of the parallel buffers. Yet, it can be shown that for independent flows and an optimal splitting, they are equal. Therefore the two independent flows can be 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 as well. For a given $\nu \geq 0$, we can find the allowable rate-pairs (a_1, a_2) such that the QoS constraint θ_0 is satisfied by the system. Accordingly, the achievable rate-region will be equal to the union of the regions corresponding to all possible values of ν. That is,

$$\mathcal{R}_{\mathrm{cr}} = \bigcup_{0 \leq \nu \leq c_3} A_1(\theta_0, c_1, \nu) \times A_2(\theta_0, c_2, c_3 - \nu); \qquad (5)$$

where A_i is the achievable rate region corresponding to the source at node i.

V. Wireless Butterfly Network

We study a wireless butterfly network under a broadcast paradigm. We assume that the system operates in frequency division multiplexing (FDM) mode. Consider the multicast scenario where two sources wish to communicate with two destination nodes. An additional node that acts as a relay is present to facilitate communication over the network. All the nodes have 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 non-interfering frequency bands.

Node 1 broadcasts its message to nodes 3 & 5, and node 2 does the same to nodes 3 & 6. Node 3 can broadcast its message to nodes 5 & 6. The packets transmitted by node 3 can be either multiplexed messages from nodes 1 & 2, or their algebraic sum. We call these modes of operation routing and network coding, respectively.

For simplicity, we assume that all the transmission links are time-invariant. We consider two cases: (i) the additive white Gaussian noise (AWGN) case, and (ii) channels that are subject to path loss components. Again, we suppose that every node is equipped with an arbitrary large buffer to store information before being sent to the destination. We also assume that a simple link layer acknowledgment scheme is present, so that data is 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

$$W \log_2 \left(1 + \frac{P}{N_0 W}\right) \qquad (6)$$

where P is the power of the received 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 (6) 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 links, and hence they become of equal capacity. We can denote the constant service rate offered by each link as $c = (W/3) \log_2 (1 + 3\gamma)$, where $\gamma = P/(N_0 W)$ is the observed SNR.

For the second case, we ignore fading and assume that received power decays exponentially in distance with an exponent α. With a spectral bandwidth allocation of ξW and a distance of d meters, the capacity of the corresponding link in bits/second becomes

$$c(d, \xi) = \xi W \log_2 \left(1 + \frac{\gamma}{\xi d^{\alpha}}\right). \qquad (7)$$

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

B. AWGN Links

First we consider the case when 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 & 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 & 2. Under a QoS constraint θ_0, the maximum possible rate received by destinations C and D are identical and equal to a_1 from source A and a_2 from source B, where

$$a_i = c \left(1 + \frac{\mu_i}{\lambda_i + \theta_0 c}\right), \quad i \in \{1, 2\}.$$

That is, the achievable rate region for source rate-pair is $\mathcal{R}_{\mathrm{nc}} = A_1(\theta_0, c, c) \times A_2(\theta_0, c, c)$.

Next, we consider the situation where node 3 simply forwards packets from nodes 1 & 2 to nodes 5 & 6. Since for all $\nu \in [0, c]$ we have $A_i(\theta_0, c, \nu) \subset A_i(\theta_0, c, c)$, the source rate-pairs (a_1, a_2) are limited by the congestion at node 3. The total achievable rate region for classical routing then becomes

$$\mathcal{R}_{\mathrm{cr}} = \bigcup_{0 \leq \nu \leq c} A_1(\theta_0, \nu) \times A_2(\theta_0, c - \nu).$$

We have plotted the results for different values of θ_0 in Fig. 3. System parameters chosen for this numerical study are $N_0 = 10^{-6}$W/Hz, $W = 22$MHz, $\lambda_1^{-1} = \lambda_2^{-1} = 650$ms, and $\mu_1^{-1} = \mu_2^{-1} = 352$ms. Additionally, we chose received power to be $P = 100$mW.

C. Links with Path Loss

We consider an example where the sources and destinations are located on the vertices of a perfect square of side length d, and the relay node lies on the perpendicular bisector of the edges connecting the two sources at a distance x. The distance from the two sources to the relay node being identical, we

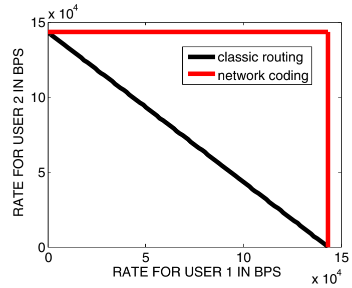

Fig. 3. Performance of a QoS butterfly network.

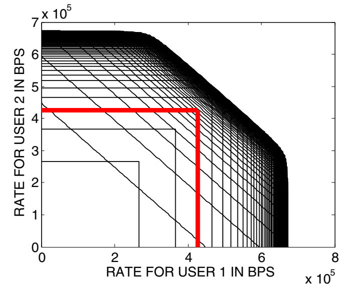

Fig. 4. Achievable Rate Region for $x/d = 0.6$.

assume that a fraction $\xi/2$ of the total bandwidth is allocated to each source; and the remaining $(1-\xi)W$, to the relay node.

To maximize the gains of network coding, we need to make the link capacities identical. The bandwidth allocation is done accordingly. Hence, we can write the capacity for the network coding case, as follows:

$$C(\xi^*) = \frac{\xi^*}{2} W \log_2 \left(1 + \frac{2\gamma}{\xi^* d_e^\alpha}\right)$$
$$= (1-\xi^*) W \log_2 \left(1 + \frac{\gamma}{(1-\xi^*) d_{35}^\alpha}\right);$$

where $d_e = \max(d, \sqrt{(d/2)^2 + x^2})$, $\gamma = P/(N_0 W)$, and the distance from the relay to a destination is $d_{35} = d_{36} = \sqrt{(d/2)^2 + (d-x)^2}$.

For the routing case, each source broadcasts its packets to the relay node, where they are buffered. The relay then simply forwards the received messages to both destinations using *generalized processor sharing*. The links from the sources to the relay node have identical capacity $C_{13} = C_{23}$, given by

$$C_{13}(\xi) = \frac{\xi}{2} W \log_2 \left(1 + \frac{2\gamma}{\xi d_{13}^\alpha}\right);$$

where $d_{13} = d_{23} = \sqrt{(d/2)^2 + x^2}$. Similarly, the link from the relay node to the destinations have capacity $C_{35} = C_{36}$, given by

$$C_{35}(\xi) = (1-\xi) W \log_2 \left(1 + \frac{\gamma}{(1-\xi) d_{35}^\alpha}\right).$$

Using the same queueing performance analysis as before and for a fixed x, we can find the achievable rate-region under QoS constraint θ_0 for network coding and traditional routing (for different values of ξ). The two regions are shown in Fig. 4 for transmit power of $P = 40$W and the same system parameters as before. Here, we have taken $d = 15$m and $\alpha = 1.8$. The region enclosed by the thick solid lines represent the achievable rate region achieved by network coding. Thin solid lines represent the same for classical routing for different values of ξ. Clearly, classical routing outperforms network coding in this case.

VI. Conclusion

We compared network coding to the classical routing for this QoS constrained communication system, and computed the achievable rate regions for both the cases. For AWGN case with identical link capacities, network coding outperforms classical routing for our communication system model. This essentially implies that network coding gains are more than the multiplexing gains achieved by routing for a symmetric network. However, we obtained more interesting results for the wireless butterfly network. In this case, combining packets at the intermediate node doesn't always offer gains and sometimes it is even harmful. Results depend on the topology of the butterfly network. For network coding to be useful, we need a symmetric direct link. It turns out that often it is better to route packets than trying to set up a direct link (if one can use the physical resources to set up a link). In our future work, we wish to study networks with varying service rates.

References

[1] R. Ahlswede and I. Csiszar, "To get a bit of information may be as hard as to get full information," *IEEE Transactions on Information Theory*, vol. 27, no. 4, pp. 398–408, July 1981.

[2] R. Koetter and M. Médard, "An algebraic approach to network coding," *IEEE/ACM Transactions on Networking*, vol. 11, no. 5, pp. 782–795, October 2003.

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

[4] D. S. Lun and A. Eryilmaz, "Control for inter-session network coding," in *Information Theory and Applications Workshop*, January-February 2007.

[5] R. Ahlswede, N. Cai, S.-Y. Li, and R. Yeung, "Network information flow," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1204–1216, July 2000.

[6] F. P. Kelly, "Effective bandwidths at multi-type queues," *Queueing Systems*, vol. 9, pp. 5–15, 1991.

[7] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective bandwidths for multiclass markov fluids and other atm sources," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 424–428, August 1993.

[8] A. I. Elwalid and D. Mitra, "Effective bandwidth of general markovian traffic sources and admission control of high speed networks," *IEEE/ACM Transactions on Networking*, vol. 1, no. 3, pp. 329–343, June 1993.

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

[10] C. Rago, P. Willett, and Y. Bar-Shalom, "Censoring sensors: a low-communication-rate scheme for distributed detection," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 32, no. 2, pp. 554–568, April 1996.

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