Guaranteed Quality-of-Service Wireless Access to ATM

Cheng-Shang Chang*, Jin-Fu Chang†, Kwang-Cheng Chen* and Ming-Young You*

January 11, 1996

Abstract

In this paper, we study the problem of wireless access to ATM. We consider three classes of ATM sources: constant bit rate (CBR), variable bit rate (VBR) and available bit rate (ABR). We propose a polling scheme with nonpreemptive priority. Under such a scheme, we derive sufficient conditions such that all the CBR sources satisfy their jitter constraints and all the VBR sources satisfy their delay constraints. The remaining bandwidth is used by the ABR sources, for which we adapt a random access scheme proposed by Chen and Lee (1994). For this random access scheme, we derive the throughput-offer load characteristic, and thus the capacity. Based on this, we propose adaptive random access schemes that track the offer load to its optimal value. Our simulations show that our adaptive schemes maintain a high throughput with respect to the whole range of system load.

Key words: ATM; Wireless networks; Multiple access; Quality of service; Polling

*Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30043, Taiwan, R.O.C. Email: cschang@ee.nthu.edu.tw, chenkc@ee.nthu.edu.tw, myyou@littlewood.ee.nthu.edu.tw

†Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, R.O.C. Email: jfchang@cc.ee.ntu.edu.tw.

1 Introduction

2 Problem formulation

Consider a multiple access problem with many users sharing a common wireless channel. For these users to communicate with each other, we assume there is a system controller, called *base station*, that is capable of broadcasting packets to all the users. We call this *down link*. Also, all the users are capable of sending packets to the base station through a multiple access channel. We call this multiple access channel *up link*. For the up link, we assume that different packets sent by different users at the same time are lost.

One of our objectives in this paper is to provide a scheme that supports ATM services in such a environment. Since the down link is a broadcasting channel from a centralized base station, its characteristic is not much different from a centralized ATM link, where information, such as packet arrival times and active users, is available to the base station. Hence, we will concentrate on the up link, where information is distributive and kept to each user.

We consider three classes of ATM traffic: (i) constant bit rate (CBR), (ii) variable bit rate (VBR), and available bit rate (ABR). As in ATM, we assume all the packets of these classes of traffic are of the same size, but needs not to be of the same format or size as an ATM cell for efficient transmission in a wireless channel. To formalize our problem, we make the following assumptions and characterizations on these three classes of traffic.

(A0) All packets have the same size.

(A1) A CBR source is characterized by two parameters (γ, δ), where γ is the rate of the source and δ is the maximum tolerable jitter (packet delay variation) for this source. Packets are generated periodically every $1/\gamma$ second and are stored in the Ready-to-Transmit (RTT) buffer. Jitter is defined to be the difference between the time of two successive departures and the time of two successive arrivals. Since a CBR source periodically generates a packet every $1/\gamma$ second, the maximum jitter is the maximum difference between two successive departures and $1/\gamma$. A CBR source can only be admitted when its maximum jitter constraint can be honored for every packet it generates.

(A2) A VBR source is characterized by three parameters (ρ, σ, d), where ρ is the average rate of the source, σ is the maximum burstiness of the source, and d is the maximum tolerable delay of the source. The first two parameters correspond to the token generation rate and the size of the token buffer of a leaky bucket in [24]. Thus, a VBR source is regulated by a (σ, ρ)-leaky bucket. We also assume that σ is an integer, but $1/\rho$ needs not be an integer. Packets passing through the leaky buckets are stored in the RTT buffer. The delay of a packet is defined to be the difference between the time it arrives at the base station and the time it arrives at the RTT buffer. As a CBR source, a VBR source can only be admitted when its maximum delay constraint can be honored for every packet it generates.

(A3) An ABR source neither has a jitter constraint nor a delay constraint. All ABR sources shares *fairly* and *efficiently* the remaining bandwidth from the CBR and VBR sources.

One of the controversial issues is whether packet dropping should be performed by the network or by the users. Here we implicitly shift the burden of packet dropping to the users of VBR sources. Packet dropping occurs in the leaky bucket that regulates the VBR traffic. Those packets that pass through a leaky bucket regulator are transmitted to the base station with a guaranteed maximum delay.

3 Polling with nonpreemptive priority for CBR and VBR sources

As discussed in the previous section, information is distributive in a wireless environment. Due to the lack of information, scheduling policies, such as First Come First Served (FCFS), Earliest Due Date (EDD), Static Priority (SP) and Generalized Processor Sharing (GPS) [20], are difficult to implemented without explicitly and constantly exchanging information. On the other hand, policies based on (peak rate) reservation, such as R-ALOHA [6] and PRMA [14], are too simple to provide efficient services for VBR sources. Since information exchange consumes precious bandwidth in a wireless channel, it is important to execute information exchange efficiently. In the following, we propose a Polling scheme with Nonpreemptive priority (PNP), where information exchange occurs in time to meet quality of services of CBR sources and VBR sources. The basic idea is polling token. Polling is done (for CBR and VBR) only when there are polling tokens.

Polling with nonpreemptive priority (PNP):

1. For each (γ, δ)-CBR source, its polling token is generated every $1/\gamma$ second in the base station.

2. For each (ρ, σ, d)-VBR source, its first polling token is generated in the base station p second after the connection is set up, where p is a design parameter to meet its delay constraint.

3. When the channel is cleared (the end of a transmission), the base station performs the following tasks:

   3.1. The base station first scans the polling token buffer for CBR sources according to a preset priority. If a polling token is found, its removes the polling token and polls the CBR source.

   3.2. If no polling tokens are found in the polling token buffer for CBR sources, it continues to scan the polling token buffer for VBR sources according to a preset priority. If a polling token is found for a VBR source, the base station polls the VBR source without removing the polling token.

   3.3. When there are no polling tokens are found for CBR and VBR sources, the base station switches to the service for ABR sources and signalling (as there are no time bounded packets that need to be served at the moment). The description of the service for ABR sources and signalling will be deferred to Section 4.

4. When a CBR source is polled, it transmits at most one packet from its RTT buffer.

5. When a VBR source is polled, it transmits a packet if there is one in its RTT buffer. Otherwise, it transmits an End-of-File (EOF) signal.

6. When an EOF signal from a VBR source is received, the base station removes the polling token of that source from the polling token buffer. Set up to generate the next polling token after p second.

7. When an ABR source is polled, it transmits at most one packet.

We depict the scheme in Fig. 1. The operation of the PNP scheme is asynchronous. Moreover, at given time t, it takes at most one polling time and one packet transmission time to clear the channel, assuming the time to send an EOF signal is shorter than the time to send a packet.

In the following theorem, we provide sufficient conditions for all the CBR packets to satisfy their maximum jitter constraints and for all VBR packets to satisfy their maximum delay constraints under the PNP scheme. We assume there are n_c sources, indexed by $i = 1, \ldots, n_c$, and n_v VBR sources, indexed by $j = 1, \ldots, n_v$. The preset priority in both CBR and VBR sources is in the increasing order. Thus, the first CBR source has the highest priority among all the sources. Let τ_p, τ_d and τ_e be the time to poll a source, the time to transmit a packet, and the time to transmit an EOF signal. We assume that $\tau_e \leq \tau_d$. Denote by (γ_i, δ_i) the traffic parameters for the i^{th} CBR source and (ρ_j, σ_j, d_j) the traffic parameters for the j^{th} VBR source. Let p_j be the polling token generation period for the j^{th} VBR source. Let $\lceil x \rceil$ be the integer ceiling of x.

Theorem 3.1 *Assume there are no transmission errors in the channel and (A0-2) hold.*

(i) *Let*

$$\delta_i^* = (\tau_p + \tau_d) + \sum_{k=1}^{i-1} \lceil \frac{\gamma_k}{\gamma_i} \rceil (\tau_p + \tau_d), \quad i = 1, \ldots, n_c. \tag{1}$$

If $\delta_i^* + (\tau_p + \tau_d) < 1/\gamma_i$ for all $i = 1, \ldots, n_c$, then the jitter of the i^{th} CBR source is bounded above by δ_i^*, $i = 1, \ldots, n_c$. If, furthermore, $\delta_i^* \leq \delta_i$ for all i, then all the packets generated by these n_c CBR sources meet their jitter constraints.

(ii) *Let*

$$\hat{\sigma}_0 = (\tau_p + \tau_d)(n_c + 1), \tag{2}$$

and

$$\hat{\rho}_0 = (\tau_p + \tau_d) \sum_{i=1}^{n_c} \gamma_i. \tag{3}$$

For $j = 1, \ldots, n_v$, let

$$\hat{\sigma}_j = (\tau_p + \tau_d)(\sigma_j + 1), \tag{4}$$

and

$$\hat{\rho}_j = (\tau_p + \tau_d)\rho_j + (\tau_p + \tau_e)/p_j. \tag{5}$$

For $j = 1, \ldots, n_v$, define recursively

$$d_j^* = p_j + \frac{\sum_{k=0}^{j} \hat{\sigma}_k + (j-1)(\tau_p + \tau_e) + (\tau_p + \tau_d) \sum_{k=1}^{j-1} \rho_k d_k^*}{1 - \sum_{k=0}^{j-1} \hat{\rho}_k}. \tag{6}$$

If $\sum_{k=0}^{n_v} \hat{\rho}_k \leq 1$, then the delay of the packets of the j^{th} VBR sources is bounded above by d_j^*, $j = 1, \ldots, n_v$. If, furthermore, $d_j^* \leq d_j$ for all j, then the delay constraints are satisfied for all the VBR sources.

To complete our scheme, we still need to engineer the polling token generation periods, $p_j, j = 1, \ldots, n_v$. In order to minimize overheads due to polling and EOF signals, one should have p_j as large as possible. The largest possible p_j can be obtained by solving $d_j^* = d_j$ in (6). On the other hand, it is very troublesome to change p_j when a new connection is set up or a connection is closed. In view of this, one might not want to push p_j to its maximum value. Instead, one should choose a moderate large p_j and use it until the connection is closed. As a compromise, here we suggest using

$$p_j = d_j/2. \tag{7}$$

We also suggest that the priority be assigned in the increasing order of the maximum delay constraint, i.e., the one with the smallest maximum delay constraint has the highest priority among all the VBR sources. In this case, admission control of a new (ρ, σ, d)-VBR source is simple. One first finds the largest index k such that $d_k \leq d$. Insert the new source and recompute (6) from $j = k+1$ upwards. If all these inequalities in Theorem 3.1(ii) are still satisfied, then the new source is admitted, otherwise it is rejected.

The admission control of a CBR source requires more computation. All the conditions in (i) and (ii) needs to be examined due to the priority nature of the scheme. Thus, it might be desirable to limit the impact of CBR sources to other traffic. Note from $\hat{\rho}_0$ and $\hat{\sigma}_0$ in (6) that the impact of CBR sources to VBR sources is only through the total number n_c and the total rate $\sum_{i=1}^{n_c} \gamma_i$ of CBR sources. If one can have preset limits for these two quantities, then the admission control of CBR sources and VBR sources can be done separately.

The rest of this section is devoted to the proof of Theorem 3.1.

Proof. (Theorem 3.1 (i)) As described in the PNP scheme, the polling tokens of the i^{th} CBR source are generated periodically with period $1/\gamma_i$. Thus, if we can show that each polling token of the i^{th} CBR source has a maximum wait $\tilde{\delta}_i$, then the jitter cannot be larger than $\tilde{\delta}_i$. Moreover, if $\tilde{\delta}_i + (\tau_p + \tau_d) < 1/\gamma_i$, the polling token is used before its next arrival. Thus, there is at most 1 i^{th} polling token in the base station at any time.

Now we prove this by induction. Our induction hypotheses are (I) $\tilde{\delta}_i \leq \delta_i^*$ and (II) $\tilde{\delta}_i + (\tau_p + \tau_d) < 1/\gamma_i$, where δ_i^* is defined in (1). Consider the first CBR source, its maximum wait is for the channel to be cleared since we assume a nonpreemptive priority. The channel will be cleared within a polling time τ_p and a packet transmission time τ_d. Thus, $\tilde{\delta}_1 = \tau_p + \tau_d = \delta_1^*$. After the channel is cleared, the first CBR source is polled and it transmits at most one packet. To ensure this is done before the arrival of next polling token, a sufficient condition is $\tilde{\delta}_1 + \tau_p + \tau_d < 1/\gamma_1$. But this is exactly our assumption $\delta_1^* + (\tau_p + \tau_d) < 1/\gamma_1$.

Now suppose our induction hypotheses hold up to the $i-1^{th}$ CBR source. We argue by contraction that $\tilde{\delta}_i \leq \delta_i^*$. Let the arrival instant of a "marked" polling token as time 0. Suppose

$\tilde{\delta}_i > \delta_i^*$. Then it follows from the priority scheme that up to time δ_i^* the channel must be serving all the CBR sources, $k = 1, \ldots, i-1$. From the induction hypothesis (II), every packet of these $i-1$ CBR sources is served before the generation of next polling token. Thus, the total amount of packets that can be served within $(0, \delta_i^*)$ for these $i-1$ CBR sources is at most $\sum_{k=1}^{i-1} \lceil \gamma_k \delta_i^* \rceil$. Adding $\tau_d + \tau_p$ for nonpreemptive priority, the total amount of time to serve these packets is bounded above by

$$\sum_{k=1}^{i-1} \lceil \gamma_k \delta_i^* \rceil (\tau_p + \tau_d) + (\tau_p + \tau_d). \tag{8}$$

Since we assume $\delta_i^* + (\tau_d + \tau_p) < 1/\gamma_i$ in (i), we have $\delta_i^* < 1/\gamma_i$ and the quantity in (8) is bounded above by

$$\sum_{k=1}^{i-1} \lceil \frac{\gamma_k}{\gamma_i} \rceil (\tau_p + \tau_d) + (\tau_p + \tau_d) = \delta_i^*, \tag{9}$$

where the equality follows from the definition of δ_i^* in (1). Thus, the channel cannot be busy all the time in $(0, \delta_i^*)$ and we reach a contradiction. This shows $\tilde{\delta}_i \le \delta_i^*$. Also, from the assumption $\delta_i^* + (\tau_d + \tau_p) < 1/\gamma_i$, the marked polling token of the i^{th} CBR source will be removed before the generation of next polling token. ∎

Our proof for Theorem 3.1(ii) is based the calculus developed by Cruz [10, 11].

Definition 3.2 *Consider a nonnegative, left limit, and right continuous stochastic process* $A \equiv \{a(t), t \ge 0\}$. *Let* $A(t_1, t_2) = \int_{(t_1 t_2]} a(t)dt$. *We say that* A *is* (σ, ρ)*-upper constrained if* $A(s, t+s) \le \rho t + \sigma$ *for all* $s, t \ge 0$. *Similarly,* A *is* (σ, ρ)*-lower constrained if* $A(s, t+s) \ge \rho t - \sigma$ *for all* $s, t \ge 0$.

Remark 3.3 Note that the number of departures in $(t_1, t_2]$ from a (σ, ρ)-leaky bucket is bounded above by $\sigma + \lceil \rho(t_2 - t_1) \rceil$. Using the upper bound $\lceil x \rceil \le x + 1$, the departure process from a (σ, ρ)-leaky bucket is $(\sigma + 1, \rho)$-upper constrained.

Now we prove Theorem 3.1(ii).

Proof. (Theorem 3.1(ii)) For the first VBR source, one can map it to a classical vacation model with machine breakdowns and repairs. The polling token indicates whether the server is available or not. Once the RTT buffer is empty, the server takes a p_1 second vacation. Since the first VBR source only use the remaining bandwidth from all the CBR sources, the amount of bandwidth available to the first VBR source is a time varying process. This corresponds to machine breakdowns and repairs in the vacation model.

To be precise, we let $C_1 \equiv \{c_1(t), t \geq 0\}$ be the stochastic process that denotes the available bandwidth to the first VBR source at time t. If the channel is available to the first VBR source at time t, then $c_1(t) = 1$. Otherwise, $c_1(t) = 0$. Now we find a lower constraint for the time varying process C_1. As discussed in the proof of Theorem 3.1(i), the maximum number of packets from the n_c CBR sources that can be served in an interval $(t_1, t_2]$ is at most $\sum_{i=1}^{n_c} \lceil \gamma_i(t_2 - t_1) \rceil$. Applying the inequality $\lceil x \rceil \leq x + 1$ yields the bound $\sum_{i=1}^{n_c}(\gamma_i(t_2 - t_1) + 1)$. Adding $\tau_p + \tau_d$ for nonpreemptive priority, the bandwidth that is available to the first VBR source in $(t_1, t_2]$ is at least $t_2 - t_1 - (\tau_p + \tau_d)(1 + \sum_{i=1}^{n_c}(\gamma_i(t_2 - t_1) + 1))$. Thus, we have the following lower bound

$$C_1(t_1, t_2) \geq (1 - (\tau_p + \tau_d) \sum_{i=1}^{n_c} \gamma_i)(t_2 - t_1) - (\tau_p + \tau_d)(n_c + 1). \tag{10}$$

In other word, C_1 is $(\hat{\sigma}_0, 1 - \hat{\rho}_0)$-lower constrained, where $\hat{\sigma}_0$ and $\hat{\rho}_0$ are defined in (2) and (3).

Note from Remark 3.3 that the number of departures in $(t_1, t_2]$ from a (σ_1, ρ_1)-leaky bucket is $(\sigma_1 + 1, \rho_1)$-upper constrained. Let $A_1 \equiv \{a_1(t), t \geq 0\}$ be the amount of workload (actual data and polling overheads) that arrives at the channel at time t. Then

$$A_1(t_1, t_2) \leq (\tau_p + \tau_d)(\rho_1(t_2 - t_1) + \sigma_1 + 1). \tag{11}$$

This shows A_1 is $(\hat{\sigma}_1, (\tau_p + \tau_d)\rho_1)$-upper constrained, where $\hat{\sigma}_1$ is defined in (4).

Consider an instant that an EOF signal from the first VBR source is received. Note that the moment the connection is set up is equivalent to the case that an EOF signal is received. Mark the instant as time 0. Let $q_1(t)$ be the amount of backlogged workload (in the RTT buffer) from the first VBR source in the channel at time t. Since an EOF signal is received at time 0, we know $q_1(0) = 0$. Moreover, the polling token will be generated at time p_1. Before p_1, the first VBR source cannot access the channel even when the channel is available. Thus, until the next EOF signal, one has

$$q_1(t) = A_1(0, t) - C_1(p_1, t). \tag{12}$$

Since the service policy is First Come First Served (FCFS) within each VBR source, the delay for an arrival at time t is bounded by the amount of time needed to deplete $q_1(t)$. The time to deplete $q_1(t)$ is bounded by

$$\inf\{d \geq 0 : A_1(0, t) - C_1(p_1, t + d) \leq 0\}. \tag{13}$$

Maximizing over t, we have the following upper bound for the maximum delay

$$d_1^* = \sup_t \inf\{d \geq 0 : A_1(0,t) - C_1(p_1, t+d) \leq 0\}. \tag{14}$$

Since we assume $\sum_{k=0}^{n_v} \hat{\rho}_k \leq 1$, we have $\hat{\rho}_0 + (\tau_p + \tau_d)\rho_1 \leq 1$. Applying the upper constraint for A_1 in (11) and the lower constraint for C_1 in (10) yields

$$d_1^* = p_1 + (\hat{\sigma}_0 + \hat{\sigma}_1)/(1 - \hat{\rho}_0). \tag{15}$$

This completes the argument for the first VBR source.

The argument for the j^{th} VBR source is essentially the same as that for the first VBR source. Note that the lower constraint for the channel needs to be modified since the j^{th} VBR source only uses the remaining channel from all the CBR sources and the first $j-1$ VBR sources. In addition to the workload in (10), one needs to add the workload from the first $j-1$ VBR sources. Since the maximum delay of the k^{th} VBR source is bounded above by d_k^*, $k = 1, \ldots, j-1$, the number of packets from the k^{th} source that can served in $(t_1, t_2]$ is bounded above by $\sigma_k + \lceil \rho_k(t_2 - t_1 + d_k^*) \rceil$. When the RTT buffer of the k^{th} VBR source is empty, its polling token is generated periodically with period p_k. This implies there are at most $\lceil (t_2 - t_1)/p_k \rceil$ EOF signals in $(t_1, t_2]$ since there is an EOF signal associated with each polling token. Adding the overheads for polling and EOF signals, the amount of workload from the k^{th} source that can be served in $(t_1, t_2]$ is bounded above by (cf. the argument for (11))

$$\begin{aligned}
&(\tau_p + \tau_d)(\rho_k(t_2 - t_1) + \sigma_k + 1 + \rho_k d_k^*) + (\tau_p + \tau_e)(\frac{1}{p_k}(t_2 - t_1) + 1) \\
&= \hat{\rho}_k(t_2 - t_1) + \hat{\sigma}_k + \rho_k d_k^*(\tau_p + \tau_d) + (\tau_p + \tau_e),
\end{aligned} \tag{16}$$

where $\hat{\sigma}_k$ and $\hat{\rho}_k$ are defined (4) and (5). Thus, the channel is $(\sum_{k=0}^{j-1} \hat{\sigma}_k + (\tau_p + \tau_d) \sum_{k=1}^{j-1} \rho_k d_k^* + (j-1)(\tau_p + \tau_e), 1 - \sum_{k=0}^{j-1} \hat{\rho}_k)$-lower constrained. Parallel to the argument for the first VBR source, the maximum delay of the j^{th} VBR source is bounded above by

$$p_j + \frac{\sum_{k=0}^{j} \hat{\sigma}_k + (\tau_p + \tau_d) \sum_{k=1}^{j-1} \rho_k d_k^* + (j-1)(\tau_p + \tau_e)}{1 - \sum_{k=0}^{j-1} \hat{\rho}_k}. \tag{17}$$

■

4 Group randomly address polling for ABR sources and signalling

In this section, we continue our discussion for ABR sources and signalling. As CBR sources and VBR sources have higher priority, ABR sources and the traffic generated by signalling can only use the remaining bandwidth. Our objective in this section is to provide a scheme that share the remaining bandwidth *fairly* and *efficiently*.

To ease our exploitation, we may view the remaining channel as a time varying channel, and forget the impacts of CBR sources and VBR sources. We also concentrate on the ABR sources only, and treat the traffic generated by signalling as part of ABR sources (the issue for separating signalling and ABR sources will be addressed elsewhere). When we say a source in this section, it should be understood to be an ABR source. Also, the channel is referring to the remaining channel.

Suppose there are N ABR sources. An ABR source is called *active* if it has a packet to send. An ideal scheme for these N ABR sources to share the channel *fairly* is to create a deterministic mapping (permutation) so that every active ABR source takes turn to send its packet. Once the mapping is set up, the (remaining) channel is fully utilized. To achieve this, every (ABR) source requires two pieces of information: (i) the set of active sources, and (ii) the deterministic mapping. A simple way to achieve this is to reserve each source a packet-to-send slot (or any other bandwidth allocation scheme to transmit this information). Active sources mark their corresponding slots so that the base station can have the information of the set of active sources. Then the base station polls active sources one by one to create the deterministic mapping. The scheme is known as the bit-map protocol (see e.g., Tanenbaum [22]). Clearly, the time taken to send the information of the set of active sources and the time to poll active sources for the creation of the deterministic mapping are overheads of the channel. Under such a scheme, the channel can be efficiently utilized if the length of the packet-to-send slot is much smaller than the packet size and almost the N sources have packets to send.

In a wireless environment, the set of sources is time varying, which in turn implies the set of active sources is also time varying. Moreover, only a small portion of sources are active. In such an environment, it may not be economical to reserve each source a packet-to-send slot. To deal with the problem that only a small portion of sources are active, one can partition sources into groups and have sources in the same group sharing a packet-to-send slot. This creates a contention problem since there are probably more than one active sources in a group. We will address this problem in details later. Since the set of sources is time varying, grouping cannot be done deterministically. Thus, a random and dynamic grouping scheme is needed.

4.1 GRAP with a fixed number of groups

Group Randomly Addressed Polling (GRAP) [8] is one of the representative protocols for the random and dynamic grouping scheme discussed previously, which we outline below.

Group Randomly Addressed Polling (GRAP):

1. At the beginning of every GRAP cycle, the base station broadcasts a READY signal to all the sources.

2. After receiving the READY signal, every active source joins randomly one of the groups with equal probability, and marks the corresponding packet-to-send slot for that group.

3. The base station then polls the marked groups one by one. Active sources in the group being polled send their packets. If there is only one active source in a group, then this active source will transmit its packet successfully. However, if there are more than one active sources in a group, then more than one (different) packets will be transmitted after being polled. This results in a collision and all collided packets are lost.

4. After the polling cycle being completed, repeat the procedure from Step 1.

Note that the marking process can be done either explicitly by transmitting an identical message to the base station, e.g., an carrier of a certain frequency, or implicitly by probing (or polling) from the base station.

To find the capacity (the maximum achievable throughput) of GRAP, we study its throughput-offer load characteristic. Suppose there are b groups and n active sources. Define the offer load $G = n/b$, the average number of active sources per group. The throughput is defined to be the proportional amount of time that successful packets are transmitted. Let τ_m be the time of a packet-to-send slot, τ_p be the time to poll a group, τ_d be the time to transmit a packet, and τ_c be the time to detect a collision.

Theorem 4.1 *When $b \to \infty$, the throughput-offer load characteristic is given as follows:*

$$S(G) = \frac{Ge^{-G}\tau_d}{\tau_m + (\tau_p + \tau_d)Ge^{-G} + (\tau_p + \tau_c)(1 - e^{-G} - Ge^{-G})}, \tag{18}$$

where S is the throughput and G is the offer load. Moreover, the capacity of GRAP is $S(G^*)$, where $0 < G^* < 1$ is the unique solution of

$$G = 1 - \frac{\tau_p + \tau_c}{\tau_m + \tau_p + \tau_c} e^{-G}. \tag{19}$$

Note that the capacity in Theorem 4.1 is only for a large number of groups. It is possible to have a higher capacity for a small number of groups.

In the case that $\tau_m = \tau_p = 1$, (18) reduces to that for nonpersistent CSMA/CD in [23, 18]. That GRAP and nonpersistent CSMA/CD have the same throughput characteristic is not surprising. In CSMA, polling is done implicitly by carrier sensing. The first empty slot after a (successful or collided) transmission is equivalent to the READY signal in GRAP. For each active source in CSMA, it still selects randomly a group to join. However, this information is kept to itself and not transmitted to the base station. Once an active source senses an empty slot, it increases its own counter and it knows which group is being "polled." The difference between GRAP and CSMA is that active sources in CSMA reselects a group to join after each transmission, while active sources in GRAP stay in the same group until the end of the whole cycle. In corresponding to the distribution of balls into urns, CSMA is sampling *with* replacement and GRAP is sampling *without* replacement. In view of this, they have the same throughput characteristic when the number of active sources is large. The advantage of using sampling with replacement (CSMA) is in light load, where an active source need not wait for the completion of the marking process of the packet-to-send slots. Similar comments were made by Chlamtac [9] and Scholl [21] for the bit map protocol. On the other hand, using sampling without replacement (GRAP) allows us to collect statistics for adaptive algorithms when the system load changes. We will discuss this more in Section 4.2. Note that implicit polling using carrier sensing can only be done in a very reliable channel, e.g., a wired cable. In the case that some active sources are not able to detect the carrier from other sources, the problem of hidden terminals might reduce the performance tremendously. In this case, explicit polling is needed.

Analogous to the interpretation of the throughput characteristic for CSMA in [23, 18], τ_m is the fixed overhead to detect empty groups (idle periods in CSMA). With empty group detections, we now only have either a successful transmission or a collision. The probability of a successful transmission can be approximate by $\text{Prob}(X = 1|X > 0)$, where X is a Poisson random variable with parameter G. Similarly, the probability of a collision can be approximately by $\text{Prob}(X \geq 2|X > 0)$. The equation (18) is then the weighted average of these three quantities. In comparison with the ALOHA family in [1, 2], the gain is obviously in the empty group detection ability in GRAP. Note that in the slotted ALOHA, the probability of a successful transmission is simply $\text{Prob}(X = 1)$, where X is again a Poisson random variable with parameter G. Conditioning on the event $\{X > 0\}$ increases the performance of GRAP. Also, as commented in [13] for CSMA, the gain for implementing a tree algorithm [4, 5] in GRAP is marginal since the collision probability is very small.

In the case without collision detection, i.e., $\tau_c = \tau_d$, we replace (19) in (18) and have

$$S(G^*) = \frac{\tau_d}{\tau_p + \tau_d + \tau_m} e^{-G^*}. \tag{20}$$

Note that $\tau_d/(\tau_p + \tau_d + \tau_m)$ is the throughput for the scheme without collision (one packet-to-send slot, one polling time and one packet transmission time). The factor e^{-G^*} is the degradation due to collisions. Since $G^* < 1$, the number of groups should be larger than the number of active sources. Moreover, in view of (19), the optimal offer load G^* is close to 0 if $t_m << \tau_p + \tau_d$. In practice, we often has $\tau_p << \tau_d$. This implies high throughput can still be achieved without collision detection if packet-to-send slot is much shorter than the packet transmission time.

Note that the throughput-offer load characteristic in (18) can be extended to the case with variable packet lengths. Now τ_d is interpreted as the expected packet transmission time. If a collision detection mechanism is implemented, then there should be a limit for the minimum packet length for collision detection. In the case without collision detection, τ_c is the expected packet transmission time of the longest packet among the collided packets. Thus, $\tau_c \geq \tau_d$ and the optimal offer load for variable packet length is not larger than that for the case with a fixed packet length. This suggests using more efforts on collision avoidance for the case with variable packet lengths.

Now we prove Theorem 4.1.

Proof. (Theorem 4.1) Recall we assume n active sources and b groups. The offer load G is defined to be the ratio n/b. Denote by N_1 (resp. N_2) the number of groups that have only one active source (resp. at least one active sources). Thus, $N_2 - N_1$ is the number of collisions in a cycle. Define the cycle time

$$C = \tau_m b + N_1(\tau_p + \tau_d) + (N_2 - N_1)(\tau_p + \tau_c) \tag{21}$$

to be the total amount of time to mark the packet-to-send slots, poll the marked groups, transmit packets from the active sources, and detect collisions. Define the throughput

$$S = \frac{E[N_1]\tau_d}{E[C]}. \tag{22}$$

Note that $N_1\tau_d$ is the time for successful transmission in the cycle. According to the renewal theory, S is the fractional of time that packets are transmitted successfully if the same experiment is repeated forever. To compute $E[N_1]$, let X_i, $i = 1, \ldots, n$, be the indicator random variable (r.v.) that takes value 1 if the i^{th} active source transmits its packet successfully and 0 otherwise. Thus, $N_1 = \sum_{i=1}^{n} X_i$. Note that

$$E[X_1] = \text{Prob}(X_1 = 1) = (1 - \frac{1}{b})^{n-1} \tag{23}$$

since the i^{th} active source transmits its packet successfully only when the other active sources are distributed to other groups. Also, X_i's are exchangeable r.v.'s and one has

$$E[N_1] = nE[X_1] = n(1 - \frac{1}{b})^{n-1}. \tag{24}$$

To compute $E[N_2]$, we let Y_j, $j = 1, \ldots, b$, be the indicator r.v. that takes value 1 if the j^{th} group contains at least one active sources and 0 otherwise. Similarly, we have $N_2 = \sum_{j=1}^{b} Y_j$ and Y_j's are exchangeable r.v.'s. Note that $Y_1 = 0$ when every active source is distributed to other groups. Thus, $E[Y_1] = 1 - (1 - \frac{1}{b})^n$ and

$$E[N_2] = bE[Y_1] = b(1 - (1 - \frac{1}{b})^n). \tag{25}$$

In view of (21), (24) and (25), one has from (22) that

$$S = \frac{n(1-\frac{1}{b})^{n-1}\tau_d}{\tau_m b + (\tau_p + \tau_d)(n(1-\frac{1}{b})^{n-1}) + (\tau_p + \tau_c)(b(1-(1-\frac{1}{b})^n) - n(1-\frac{1}{b})^{n-1})}. \tag{26}$$

Now the throughput is a function of G and n. Replacing this in (26) yields

$$S = S_n(G) = \frac{G(1-\frac{G}{n})^{n-1}\tau_d}{\tau_m + (\tau_p + \tau_d)G(1-\frac{G}{n})^{n-1} + (\tau_p + \tau_c)(1-(1-\frac{G}{n})^n - G(1-\frac{G}{n})^{n-1})}. \tag{27}$$

To derive the throughput for a large number of groups and a large number of active sources, we let $n \to \infty$. Thus,

$$S(G) = \lim_{n\to\infty} S_n(G) = \frac{Ge^{-G}\tau_d}{\tau_m + (\tau_p + \tau_d)Ge^{-G} + (\tau_p + \tau_c)(1 - e^{-G} - Ge^{-G})}. \tag{28}$$

Simple calculus shows that the optimal offer load satisfies the following equation

$$G = 1 - \frac{\tau_p + \tau_c}{\tau_m + \tau_p + \tau_c} e^{-G}. \tag{29}$$

Let $h(G) = G - 1 + \frac{\tau_p+\tau_c}{\tau_m+\tau_p+\tau_c}e^{-G}$. Then $h(0) < 0$ and $h(1) > 0$. Moreover, $h'(G) = 1 - \frac{\tau_p+\tau_c}{\tau_m+\tau_p+\tau_c}e^{-G} \geq 0$ for all $G \geq 0$. Thus, there exists a unique optimal offer load $0 < G^* < 1$.

4.2 Adaptive GRAP

In view of the throughput-offer load characteristic in (18), it is desirable to operate GRAP close to the optimal offer load G^*. In other words, when the number of active sources is increased, one should also increase the number of groups. The idea of tracking the optimal offer load has been discussed extensively for other protocols (see e.g., [22]). The objective of this section is to develop an adaptive algorithm for GRAP. Our algorithm is based on the following well known central limit theorem (CLT) for random allocation of balls into urns (see e.g. [12], pp. 313).

Theorem 4.2 *As* $b \to \infty$, $\frac{b-N_2-be^{-G}}{\sqrt{b}e^{-G}\sqrt{e^G-G-1}}$ *converges in distribution to the standard normal random variable, where* N_2 *is the number of marked groups (marked packet-to-send slots) in a cycle, and* $G = n/b$ *is the offer load.*

From Theorem 4.2, it follows that for $b >> 1$ we have a parameter estimation problem for the parameter G with respect the family of density functions of the sample N_2

$$f(N_2, G) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\big(\frac{b - N_2 - be^{-G}}{\sqrt{b}e^{-G}\sqrt{e^G - G - 1}}\big)^2\Big). \tag{30}$$

The maximum likelihood (ML) estimate yields

$$\hat{G} = \log \frac{b}{b - N_2}. \tag{31}$$

To track the optimal offer load, we need to model the new arrivals of active sources during the cycle. Consider the case with a finite number of sources N. We assume a nonactive source becomes an active source in a random amount of time that is exponentially distributed with mean $1/\mu$. This random amount of time is independent of all else. This model implies the number of new active arrivals is a Binomial distribution with parameters $N - n$ and $1 - e^{-\mu C}$, where n is the number of active sources at the beginning of the cycle and C is the cycle time defined in (21). Note that C is observable by the base station. The average number of new active sources arriving in the cycle is $(N - n)(1 - e^{-\mu C})$. This suggests using $(N - \hat{n})(1 - e^{-\mu C})$ as an estimate for the number of new active sources arriving in the cycle, where $\hat{n} = b\hat{G}$ is the estimate for the number of active sources,

Assuming all the collided active sources remain active for the next cycle, the number of active sources at the end of the cycle can be estimated by

$$\hat{n}_e = (\hat{n} - N_1) + (N - \hat{n})(1 - e^{-\mu C}), \tag{32}$$

where N_1 is the number of successful transmissions in the cycle, and $\hat{n} - N_1$ is the estimate for the number of collided active sources in the cycle. Note that N_1 is also available to the base station. Replacing (31) in (32) yields

$$\hat{n}_e = b \log \frac{b}{b - N_2} - N_1 + (N - b \log \frac{b}{b - N_2})(1 - e^{-\mu C}). \tag{33}$$

With the estimation in (33), we modify GRAP as follows:

Adaptive GRAP:

1. At the beginning of every cycle, the base station broadcasts not only the READY signal but also the number of groups, b, where $b = \lceil n/G^* \rceil$ and n is an estimate for the number of active users from Step 4.

4. The base station keeps track of the number of successful transmissions N_1, the number of marked groups N_2, and the cycle time C. Form the the estimate for the number of active sources

$$n = b \log \frac{b}{b - N_2} - N_1 + (N - b \log \frac{b}{b - N_2})(1 - e^{-\mu C}). \tag{34}$$

Repeat the procedure from Step 1.

The dynamic control scheme was first used in Lam and Kleinrock [17] for controlling ALOHA. The method of estimating G based on the number of empty slots in recent past was also used for controlling ALOHA in [22], pp. 270. The purpose of using an adaptive algorithm is to keep the offer load close to the optimal load offer G^* (see also [16, 25]). This not only increases the throughput, but also keeps the system stable (away from multiple eqilibrium points). For the case with a fixed number of groups b (without adaptation), applying the *load line* concept in [15] to the throughput-offer load characteristic in (18) yields

$$\mu \tau_d (N - Gb) = \frac{G e^{-G} \tau_d}{\tau_m + (\tau_p + \tau_d) G e^{-G} + (\tau_p + \tau_c)(1 - e^{-G} - G e^{-G})}. \tag{35}$$

Like most random access protocols, the solution for G may not be unique. This shows the existence of multiple equilibrium points of the fixed GRAP. For the physical interpretation of the behavior of multiple eqilibrium points, see e.g. the stochastic catastrophe theory in [19].

To implement the scheme, the base station needs to know the total number of sources. This is usually kept in a database in a wireless environment. The other parameter μ also needs to be obtained off line. If μ is not available, then one might just use the first term in (33) as the estimate, i.e., $n = b \log \frac{b}{b-N_2}$. Certainly, the estimate without the parameter μ might be very rough, but it might be corrected by resampling described below.

After the marking phase in Step 2, one might obtain from the ML estimator in (31) a new estimate for the offer load. Call this estimate $\hat{G}_2$. From the CLT in Theorem 4.2, one have a certain confidence on the estimate $\hat{G}_2$. If $\hat{G}_2 >> G^*$, then we know the prediction of the offer load in Step 4 of the previous cycle is not good, and it is expected to have a lot of collisions if Step 3 is carried out. Thus, it might be better to increase the number of groups according to the new estimate $\hat{G}_2$ and go back to Step 1 to resample once. Since the marking phase is done twice, the throughput-offer load characteristic is as follows:

$$S_2(G) = \frac{Ge^{-G}\tau_d}{2\tau_m + (\tau_p + \tau_d)Ge^{-G} + (\tau_p + \tau_c)(1 - e^{-G} - Ge^{-G})}. \tag{36}$$

From this, we know resampling is not needed if $S(\hat{G}_2) \geq S_2(G^*)$. On the other hand, resampling will be beneficial if resampling yields the perfect tracking of the optimal offer load G^* and $S(\hat{G}_2) < S_2(G^*)$. Thus, we modify the adaptive GRAP as follows:

Adaptive GRAP with resampling:

2. After receiving the READY signal, every active source joins randomly one of the b groups with equal probability, and marks the corresponding packet-to-send slot for that group. The base station keeps track of the number of marked groups N_2 and form the estimates

$$\hat{G}_2 = \log \frac{b}{b - N_2}, \tag{37}$$

and

$$\hat{n} = b\hat{G}_2. \tag{38}$$

If $G_2 > G^*$ and $S(\hat{G}_2) < S_2(G^*)$, go back to Step 1.

Resampling might ease the burden of the estimator in Step 4. For instance, the information of N and μ may not be needed. In our simulations in Section 5, we find that resampling without using the estimation in (34) is good enough for tracking the optimal offer load.

5 Experimental results

In this section, we perform several experiments for the PNP scheme (for CBR sources and VBR sources) in Section 3 and GRAP (for ABR sources) in Section 4.

In our experiment for the PNP schemes, we consider a wireless channel with link speed 10M bits/sec. There are 12 CBR sources and 16 VBR sources. All packets of these sources are 1k bits. The length for polling is 50 bits and the length for EOF is also 50 bits. Thus, we have $\tau_d = 0.1 msec$, $\tau_p = \tau_e = 5\mu sec$. To simplify our presentation, we shall set $\tau_d = 1$ and normalize every parameter with respect to τ_d. In the case, $\tau_p = \tau_e = 0.05$. In Table 1, we show the (normalized) traffic characterizations, the theoretical upper bounds in Theorem 3.1(i) for the (normalized) maximum jitter, the (normalized) maximum jitter in the experiment of these 12 CBR sources. For instance, the first CBR source has the rate 2k packets/sec and the tolerable jitter 0.2 msec. The bound in Theorem 3.1(i) for the maximum jitter is 0.105 msec and the maximum jitter in the whole simulation is 0.10496 msec. We also plot the tail distributions of these CBR sources in Fig. 2, 3 and 4. These show the bounds in Theorem 3.1(i) are very conservative as they are derived from the worst case analysis.

For each VBR source, we assume it is an output from an ON-OFF coder with coding speed 32kbits/sec. Since the packet length is 1k bits. There will be a packet coming out from a coder every 1/32 second if the coder is ON. To model the ON-OFF effect, we assume both the ON and OFF periods are 1/32 second, and at the end of every ON-OFF period we flip a coin to decide whether it is an ON period in the next ON-OFF period. The probability of being an ON period is listed in the column of Bernoulli parameter in Table 2. In Table 2, we also show the (normalized) traffic characterizations, the theoretical upper bounds in Theorem 3.1(i) for the (normalized) maximum delay, the (normalized) maximum delay in the experiment of these 16 CBR sources. For instance, the Bernoulli parameter of the first VBR source is 0.62. This implies its normalized rate (average packet arrival rate) is $0.62/32 \approx 0.0193$. It is then regulated by a leaky bucket with the normalized rate 0.0196 and buffer size 7. The normalized tolerable delay is 1200 packets. The bound in Theorem 3.1(ii) for the normalized maximum delay is 646.52 (packets) and the normalized maximum delay in the whole simulation is 600.6 packets. We also plot the tail distributions of these VBR sources in Fig. 5, 6 7 and 8. Once again, these show the bounds in Theorem 3.1(ii) are very conservative. Moreover, the straight lines in these figures show the delay distributions of these sources are almost uniformly distributed over $[0, 600]$. This implies most VBR packets are simply waiting for the arrivals of their polling tokens ($p_j = d_j/2 = 1200/2 = 600$).

We perform two experiments for GRAP. We assume the same channel speed, packet length, polling length, and EOF length in the experiment for the PNP scheme. Also, we assume the length of a packet-to-send slot is 50 bits and the base station is not capable of detecting

CBR sources	Characteristics (γ, δ)	Jitter bound in (1)	Max jitter in simulation
1	$(0.2, 2)$	1.05	1.0496
2	$(0.1, 4)$	3.15	2.0933
3	$(0.1, 6)$	4.20	3.1424
4	$(0.05, 12)$	9.45	4.1771
5	$(0.01, 60)$	48.30	5.3644
6	$(0.01, 60)$	49.35	6.2656
7	$(0.075, 80)$	70.35	6.2656
8	$(0.075, 80)$	71.40	6.1266
9	$(0.064, 100)$	85.05	7.0381
10	$(0.032, 200)$	168.00	6.2752
11	$(0.032, 200)$	169.05	7.2622
12	$(0.032, 200)$	170.10	7.1049

Table 1: CBR sources under the PNP scheme.

VBR sources	Bernoulli parameter	Characteristics (ρ, σ, d)	Delay bound in (1)	Max delay in simulation
1	0.62	$(1.96 \times 10^{-3}, 7, 1200)$	646.52	600.6
2	0.57	$(1.83 \times 10^{-3}, 6, 1200)$	665.36	601.5
3	0.55	$(1.77 \times 10^{-3}, 6, 1200)$	684.23	601.1
4	0.55	$(1.77 \times 10^{-3}, 6, 1200)$	703.25	599.9
5	0.53	$(1.68 \times 10^{-3}, 5, 1200)$	720.25	601.1
6	0.50	$(1.58 \times 10^{-3}, 6, 1200)$	739.56	603.8
7	0.48	$(1.54 \times 10^{-3}, 5, 1200)$	756.63	600.6
8	0.45	$(1.44 \times 10^{-3}, 6, 1200)$	776.09	601.9
9	0.44	$(1.41 \times 10^{-3}, 6, 1200)$	795.55	602.5
10	0.435	$(1.37 \times 10^{-3}, 51200)$	812.84	602.6
11	0.43	$(1.36 \times 10^{-3}, 5, 1200)$	830.23	603.6
12	0.4235	$(1.38 \times 10^{-3}, 5, 1200)$	847.76	604.7
13	0.41	$(1.33 \times 10^{-3}, 5, 1200$	865.53	601.4
14	0.35	$(1.14 \times 10^{-3}, 6, 1200)$	885.68	606.0
15	0.32	$(1.06 \times 10^{-3}, 4, 1200)$	900.82	608.1
16	0.3	$(1.00 \times 10^{-3}, 4, 1200$	915.87	610.2

Table 2: VBR sources under the PNP scheme.

collision. Thus, after normalization, $\tau_d = \tau_c = 1$, and $\tau_p = \tau_e = \tau_m = 0.05$. The corresponding throughput-offer load characteristic in Theorem 4.1 is plotted in Fig. 9. The capacity is found to be 0.6906. This occurs when the offer load $G = 0.27$. We assume there are 20 ABR sources ($N = 20$) in the first GRAP experiment, and 200 ABR sources in the second GRAP experiment. In both experiments, the amount of time for a nonactive source to become active is exponentially distributed with mean $1/\mu$, independent of all else. An active source remains active until its packet is transmitted successfully. Varying the parameter μ, we plot in Fig. 10 and Fig. 11 the throughputs of the fixed GRAP, the adaptive GRAP, and the adaptive GRAP with resampling for these two experiments. In the experiments for the adaptive GRAP, we use (34) to track the offer load. This requires the knowledge of μ and N. While in the experiments for the adaptive GRAP with resampling, we only use (37) and (38) to track the offer load. The estimation in (34) is not used, and they does not require the knowledge of μ and N. We use batch means to analyze the data [3]. Each point in the plot is the mean of 40 batches, with 500 successful transmissions in each batch. In both experiments, the 95% confidence intervals for these points are smaller than the corresponding markers, except for a couple of points around the sudden drop of the throughput in the fixed GRAP. There we place two circles to depict the 95% confidence intervals. From Fig. 10 and 11, it is clear that adaptive GRAP and adaptive GRAP with resampling outperform the fixed GRAP. Though there are still gaps between the peak throughputs of both adaptive GRAPs and the capacity, their throughputs do not drop as μ increases (careful readers might notice there is a small drop for the adaptive GRAP in the case $N = 20$). This is quite in contrast to the performance of the fixed GRAP, where the throughput drops tremendously as μ increases beyond a certain point. An interesting observation is that both curves for the adaptive GRAP with resampling are increasing in μ. Our explanation for this is that the central limit theorem in Theorem 4.2 comes into effect as the system load increases (μ increases) so that the tracking of the optimal offer load in (37) is more accurate for large μ.

6 Conclusions and future research

In this paper, we proposed the PNP scheme for real-time ATM sources in a wireless environment. Under the scheme, we derived sufficient conditions such that all the CBR sources satisfy their jitter constraints and all the VBR sources satisfy their delay constraints. The remaining bandwidth is then used by the ABR sources under the GRAP scheme. We also derived the throughput-offer load characteristic for GRAP, and thus the capacity for GRAP. Based on this, we proposed adaptive GRAPs which control the offer load so that the offer load is close to its optimal value. Our simulations show that the PNP scheme indeed guarantees the quality of service of real-time CBR and VBR sources. Moreover, adaptive GRAPs outperform the fixed GRAP and maintain a high throughput with respect to the whole range of system load.

From this research, some interesting problems arise.

(i) The PNP scheme provides *deterministic (hard)* quality of service for CBR sources and VBR sources under the *deterministic* characterizations in (A1) and (A2). From our simulations, these deterministic bounds are very conservative. This may be fine in a environment where resource is abundant, such as ATM. However, in a wireless environment where resource is scarce, the efficiency should be improved. To improve efficiency, researchers tend to consider *probabilistic (soft)* quality of service. Among all the probabilistic traffic characterizations, the approach by the theory of effective bandwidth (see e.g. [7] for a review) seems promising.

(ii) The PNP scheme is simple in theory and intuitive to understand. However, its admission criteria may require extensive computing if the number of sources is large. Moreover, a priority based scheme seems unfair for sources that have similar traffic characteristics. To be fair for homogeneous sources, it is important to have a round-robin like scheme. Certainly, difficulty arises for meeting quality of service for schemes without priority.

(iii) As in (ii), admission control for the PNP scheme might require extensive computing. This might cause a problem in handoff and roaming. Thus, it is crucial to have trunk reservation for handoff, but how to engineer it is a problem.

(iv) In the paper, we assume there are no transmission errors. If there is a transmission error for a real-time packet, we simply assume it is lost. Whether retransmission for a corrupted real-time packet is worthwhile is a problem that needs further investigation.

Acknowledgments: this research is supported in part by the National Science Council, Taiwan, R.O.C. under Contract NSC 85-2221-E-002-013.

References

[1] N. Abramson, "The ALOHA System- another alternative for computer communications," *Proc. FJCC*, pp. 281-285, 1970.

[2] N. Abramson, "Packet Switching and Satellites," *Proc. NCC,* pp. 695-702, 1973.

[3] P. Bratley, B.L. Fox and L.E. Scrage, *A Guide to Simulation.* New York: Springer-Verlag, 1987.

[4] J.I. Capetanakis, "Generalized TDMA: the multi-accessing tree protocol," *IEEE Trans. Commun.*, Vol. COM-27, pp 1476-1484, 1979.

[5] J.I. Capetanakis, "Tree algorithms for packet broadcast channels," *IEEE Trans. Inf. Theory*, Vol. IT-25, pp. 505-515, 1979.

[6] W. Crowther, R. Rettberg, D. Walden, S. Ornstein and F. Heart, "A system for broadcast communication: Reservation-ALOHA," *Proc. 6th Hawaii Int. Conf. Syst. Sci.*, pp. 596-603, 1973.

[7] C.S. Chang and J.A. Thomas, "Effective Bandwidth in High Speed Digital Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 13, pp. 1091-1100, 1995.

[8] K.C. Chen and C.H. Lee, "Group randomly addressed polling for multi-access wireless data networks," *IEEE International Conference on Communications*, New Orleans, pp. 913-917, 1994. "An updated version of GRAP," *IEEE Documentation P802.11/93-106.*

[9] I. Chlamtac, "Radio packet broadcast computer network- The broadcast recognition access method," M.S. thesis, Dept. of Mathematical Sciences, Tel Aviv University, 1976.

[10] R.L. Cruz, "A calculus for network delay, Part I: Network elements in isolation," *IEEE Tran. Inform. Theory*, Vol. 37, pp. 114-131, 1991.

[11] R.L. Cruz, "A calculus for network delay, Part II: Network analysis," *IEEE Tran. Inform. Theory*, Vol. 37, pp. 132-141, 1991.

[12] Y.S. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales.* New York: Springer-Verlag, 1988.

[13] R.G. Gallager, "A perspective of multiaccess channels," *IEEE Trans. Inform. Theory*, Vol. IT-31, pp. 124-142, 1985.

[14] D.J. Goodman, R.A. Valenzuela, K.T. Gayliard and B. Ramamurthi, "Packet researvation multiple access for local wireless communications," *IEEE Trans. Commun.*, Vol. 37, pp. 885-890, 1989.

[15] L. Kleinrock and S.S. Lam, "Packet switching in a multiple access broadcast channel: performance evaluation," *IEEE Trans. Commun.*, Vol. COM-23, pp. 410-423, 1975.

[16] L. Kleinrock and Y. Yemini, "An optimal adaptive scheme for multiple access broadcast communication," *Proc. ICC*, pp. 7.2.1 to 7.2.5., 1978.

[17] S.S. Lam and L. Kleinrock, "Packet switching in a multiaccess broadcast channel," *IEEE Trans. Commun.*, Vol. COM-23, pp. 891-904, 1975.

[18] J.S. Meditch and C.-T. A. Lea, "Stability and optimization of CSMA and CSMA/CD channels" in *Advances in Local Area Networks* (edited K. Kümmerla, J. O. Limb and F.A. Tobagi), New York: IEEE PRESS, 1987.

[19] R. Nelson, "Stochastic catastrophe theory in computer performance modeling," *Journal of the Association for Computing Machinery*, Vol. 34, pp. 661-685, 1987.

[20] A.K. Parekh and R.G. Gallager, "A generalized processor sharing approach to flow control in integrated service networks: the single-node case," *IEEE/ACM Trans. Networking*, Vol. 1, pp. 344-357, 1993.

[21] M. Scholl, "Multiplexing techniques for data transmission over packet switched radio systems," Ph.D. thesis, Computer Science Dept., UCLA, 1976.

[22] A.S. Tanenbaum, *Computer Networks.* New Jersey: Prentice-Hall, Inc. 1981.

[23] F.A. Tobagi and V.B. Hunt, "Performance analysis of carrier sense multiple access with collision detection" in *Advances in Local Area Networks* (edited K. Kümmerla, J. O. Limb and F.A. Tobagi), New York: IEEE PRESS, 1987.

[24] J.S. Turner, "New Directions in Communications (or Which Way to the Information Age?)," *IEEE Communication Magazine*, Vol. 24, No. 10, Oct. 1986.

[25] Y. Yemini, "On channel sharing in discrete-time packet switched, multiaccess broadcast communication," Ph.D. thesis, Computer Science Dept., UCLA, 1978.

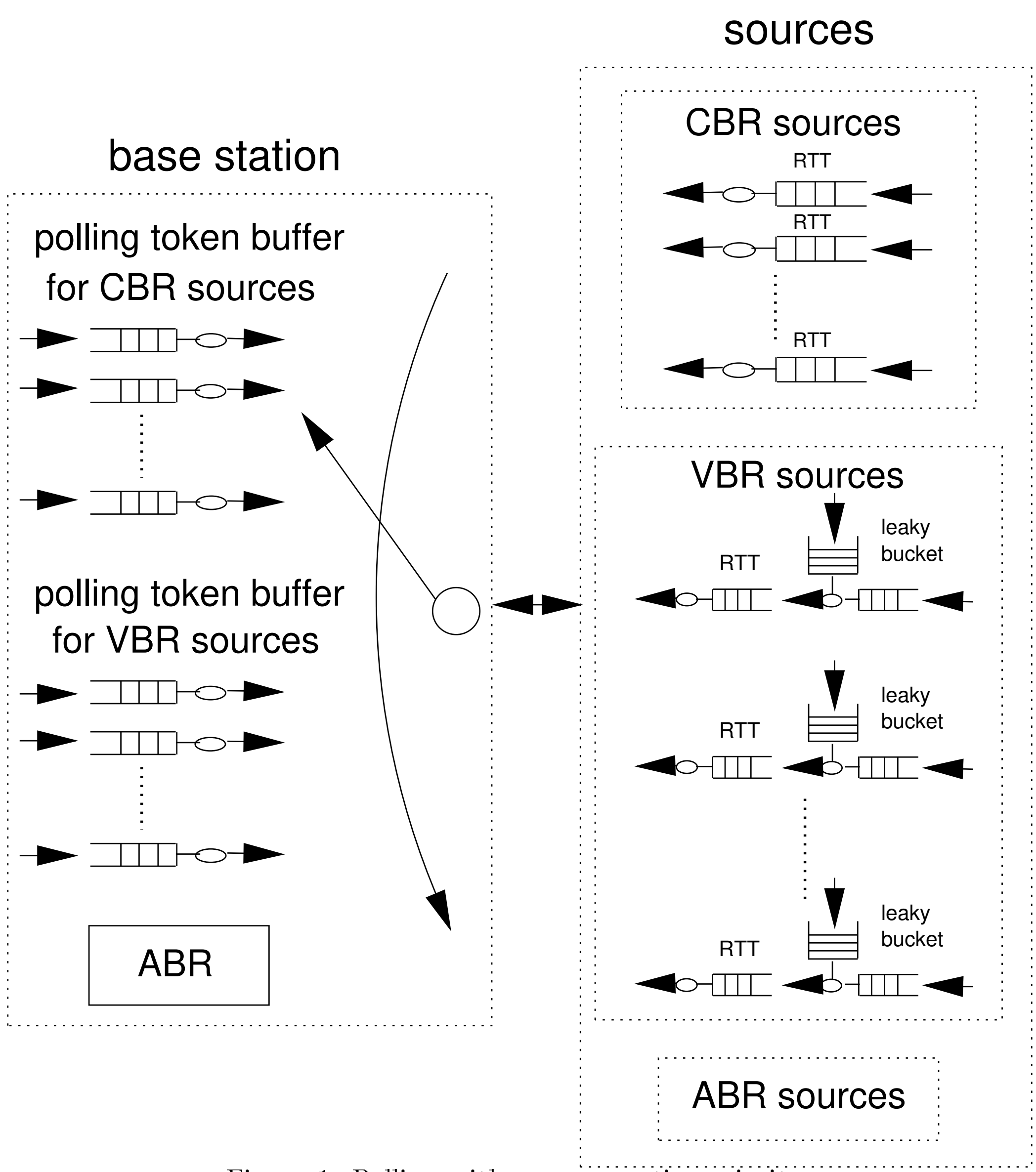

Figure 1: Polling with nonpreemptive priority

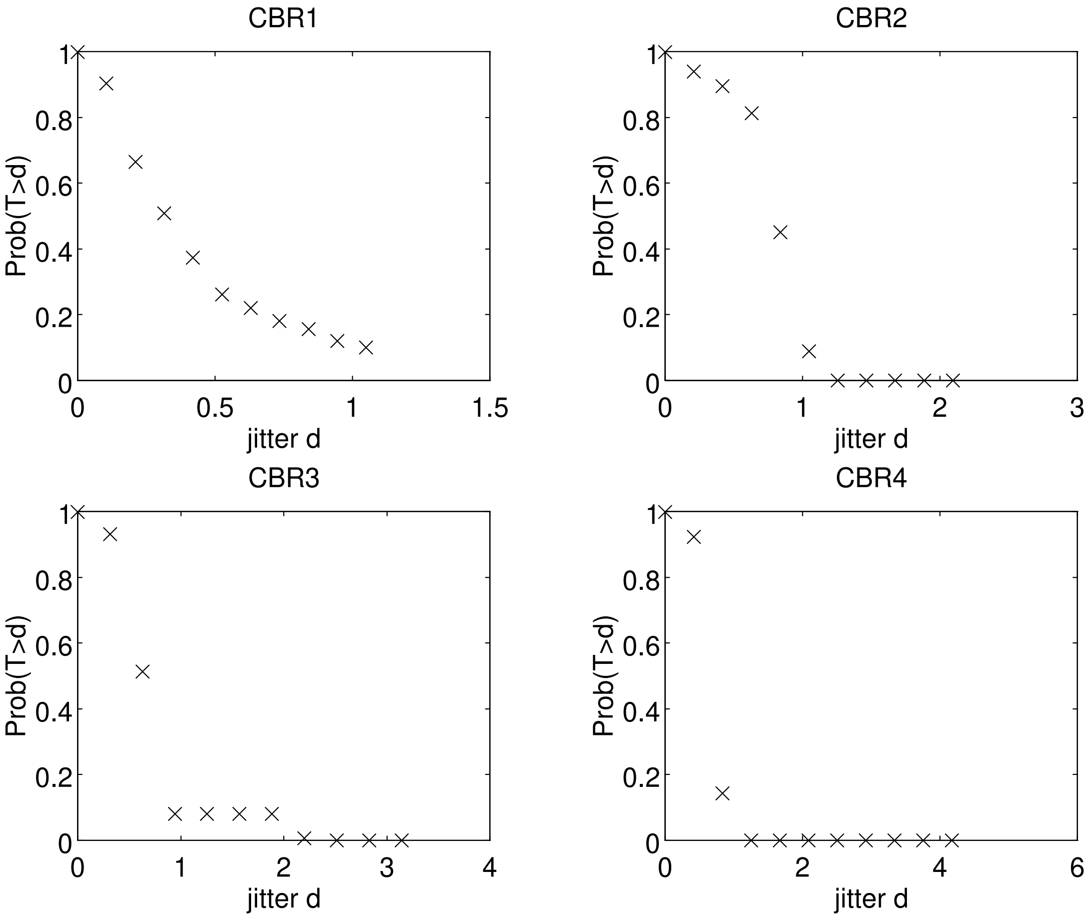

Figure 2: Jitter tail distributions for CBR 1-4

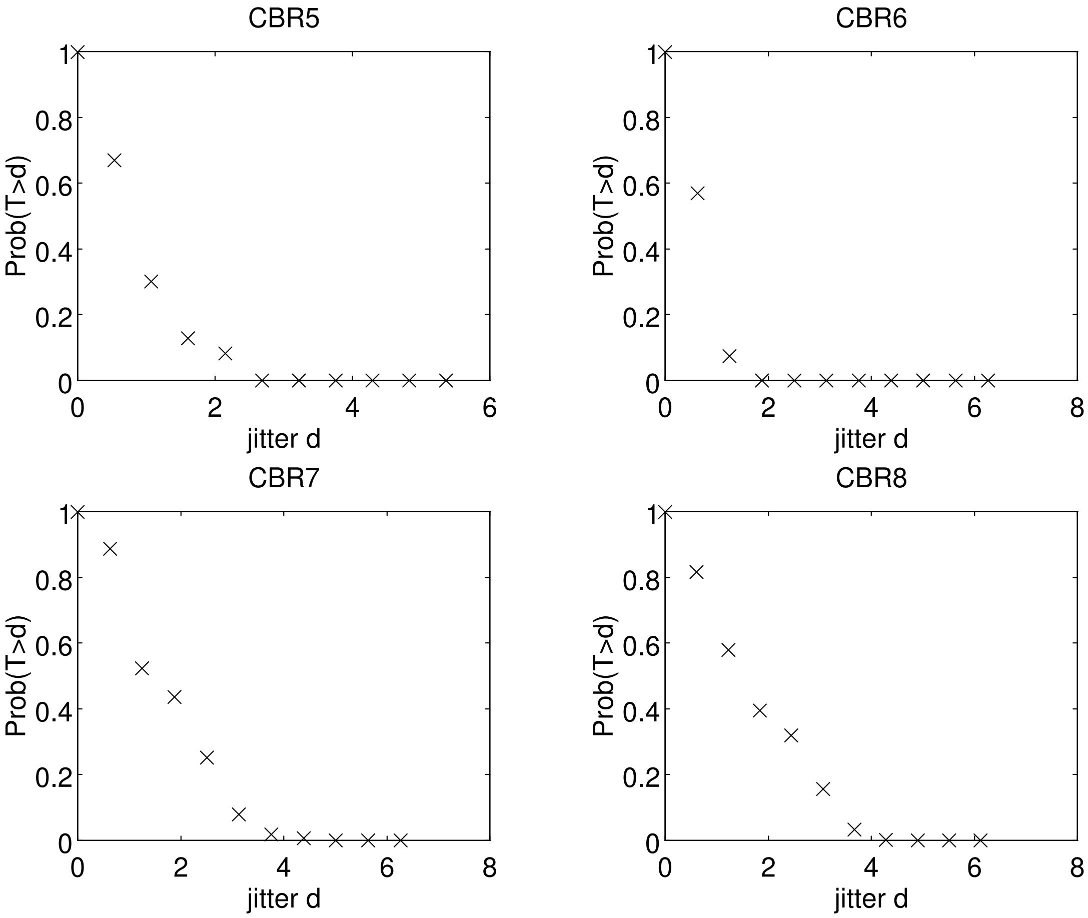

Figure 3: Jitter tail distributions for CBR 5-8

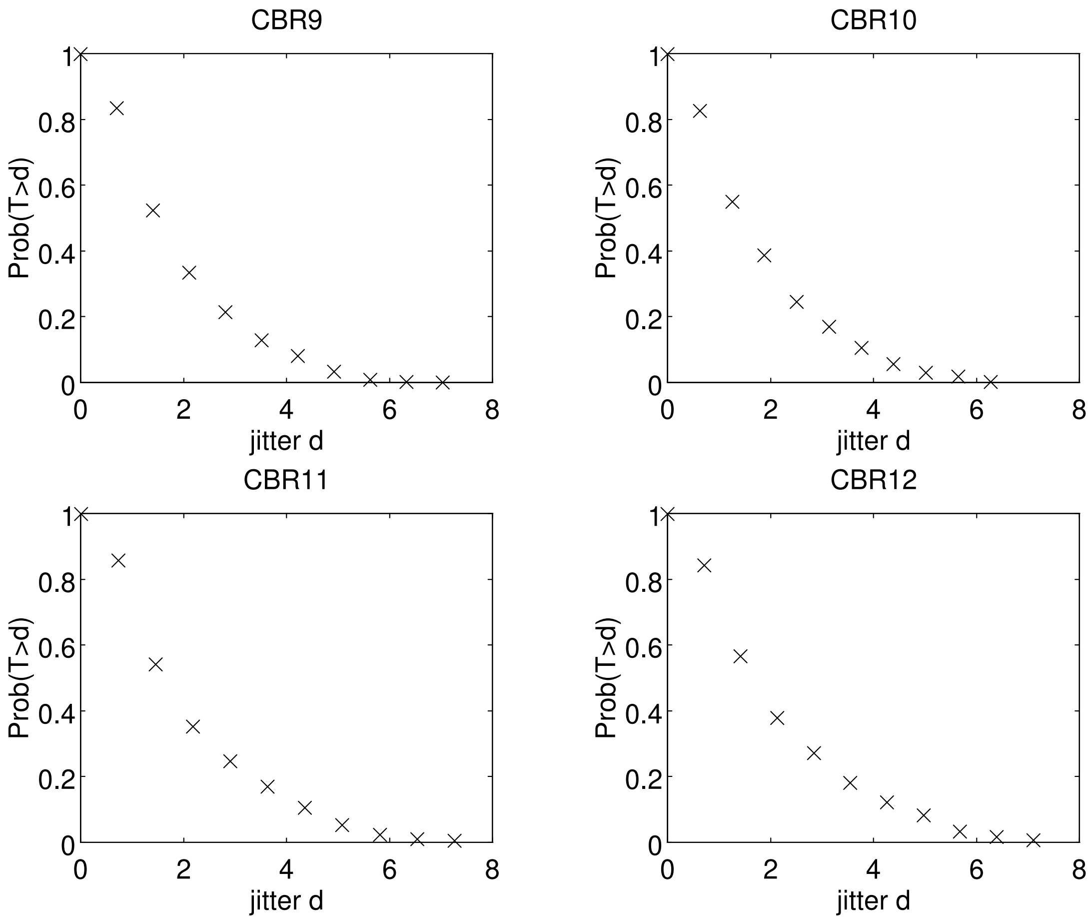

Figure 4: Jitter tail distributions for CBR 9-12

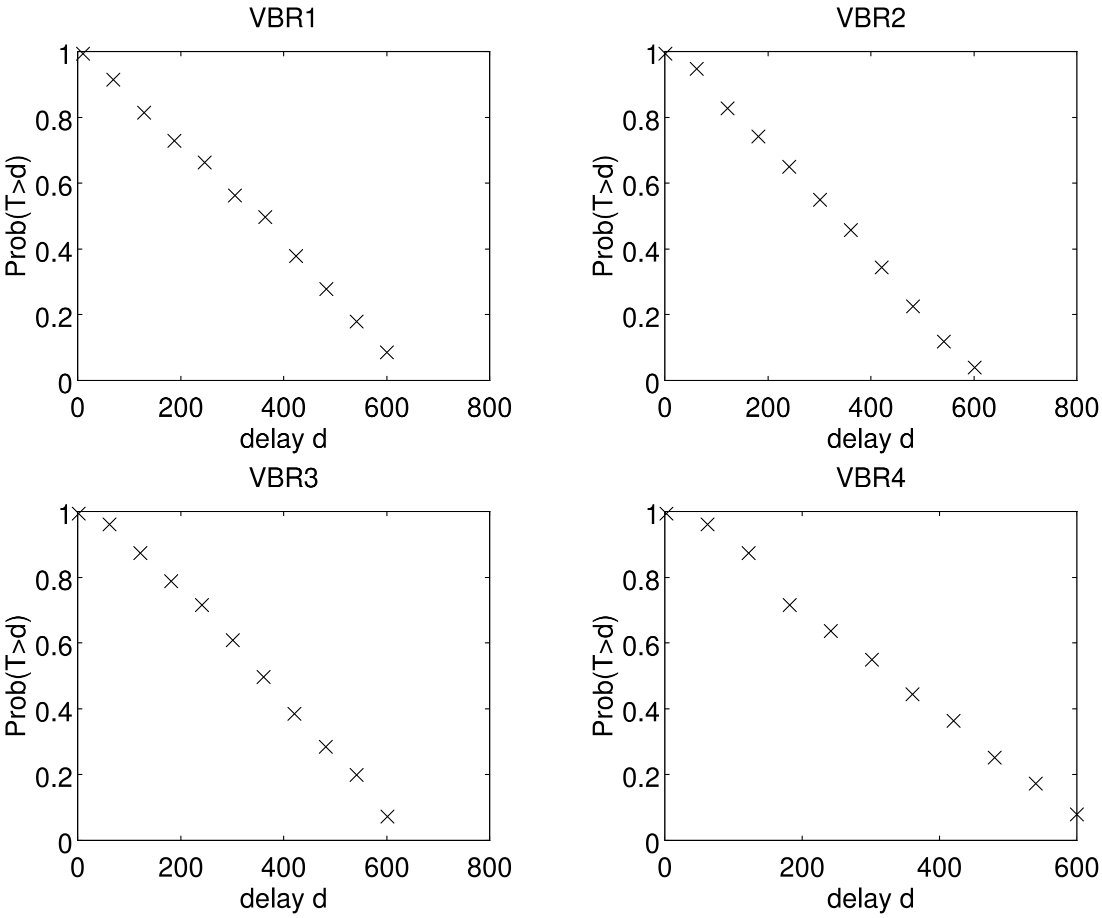

Figure 5: Delay tail distributions for VBR 1-4

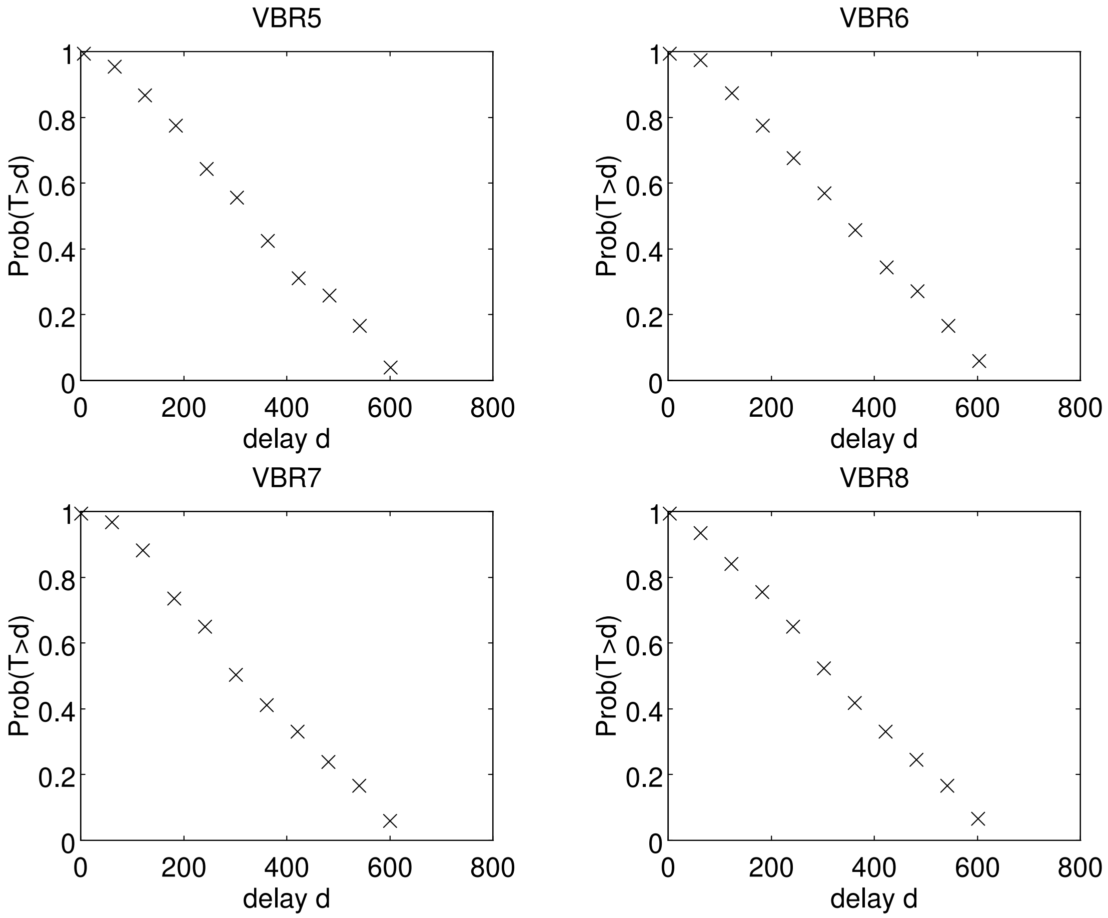

Figure 6: Delay tail distributions for VBR 5-8

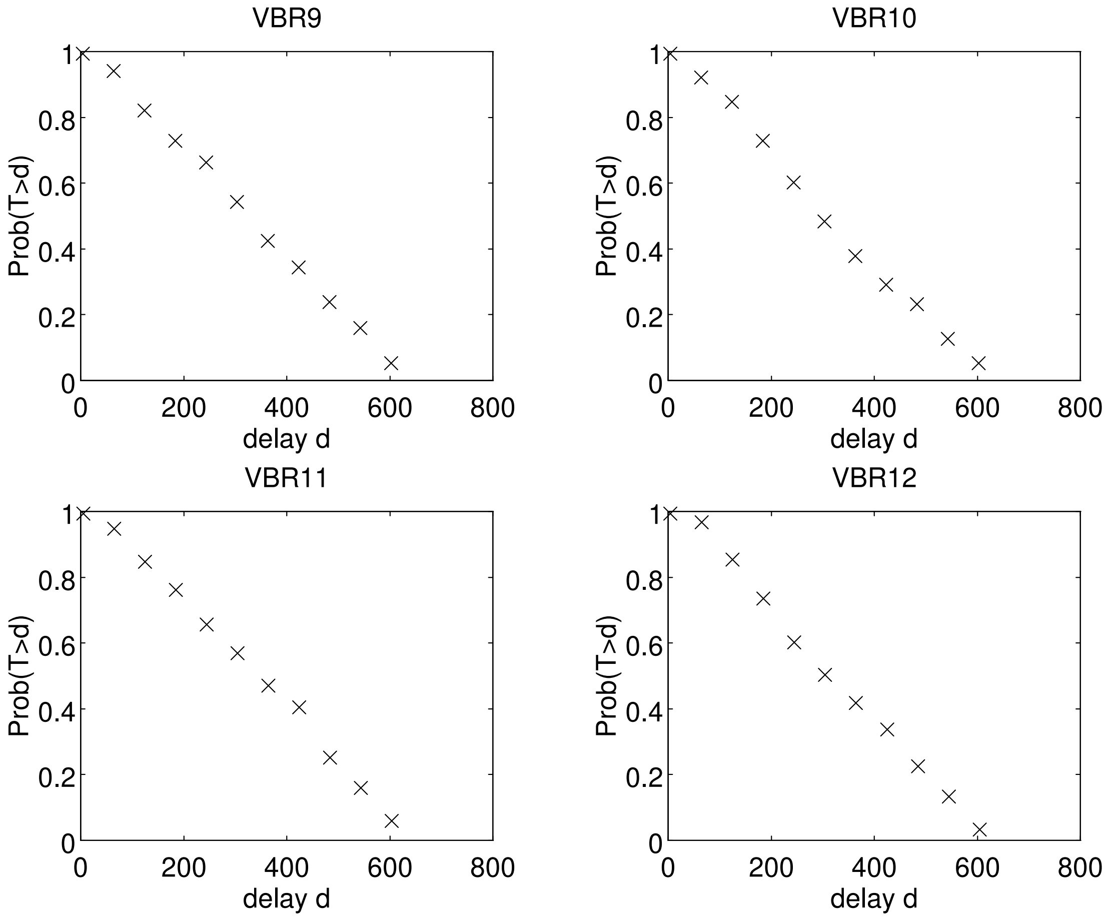

Figure 7: Delay tail distributions for VBR 9-12

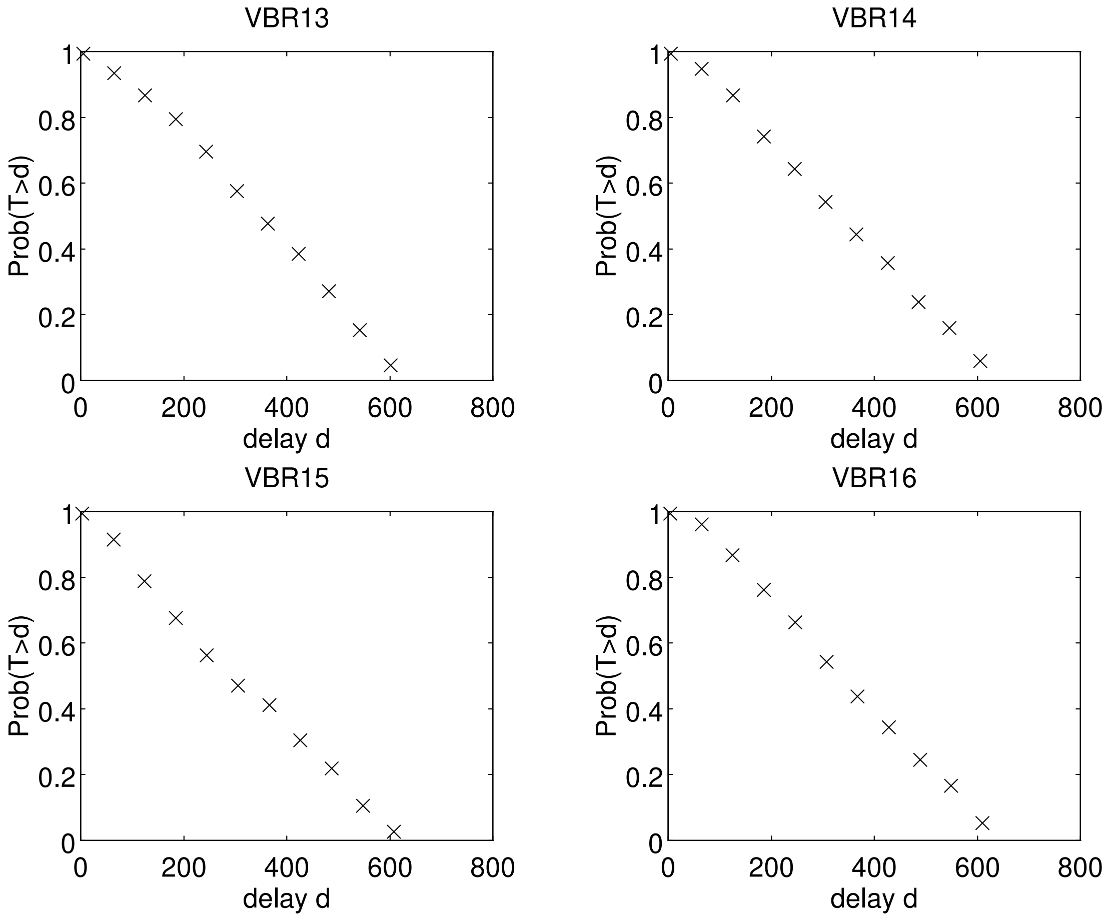

Figure 8: Delay tail distributions for VBR 13-16

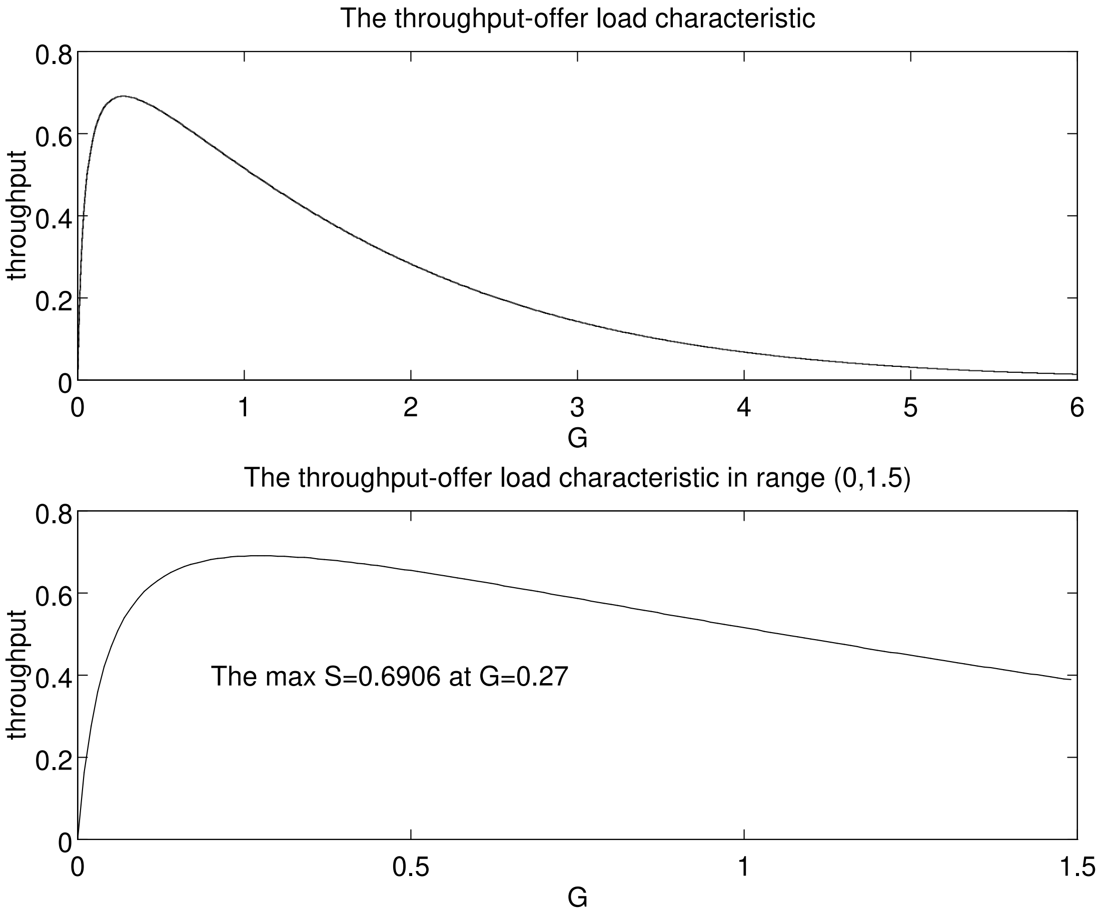

Figure 9: Throughput-offer load characteristic

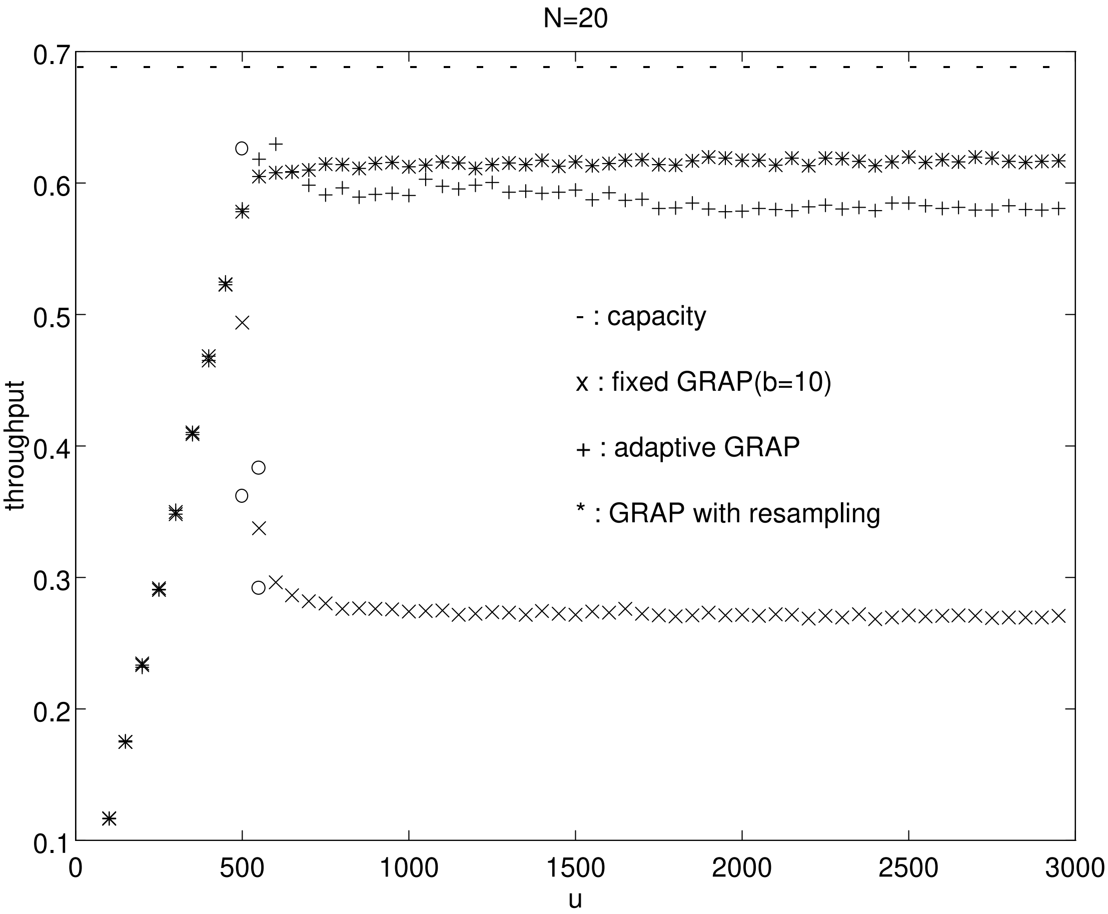

Figure 10: GRAPs with 20 ABR sources

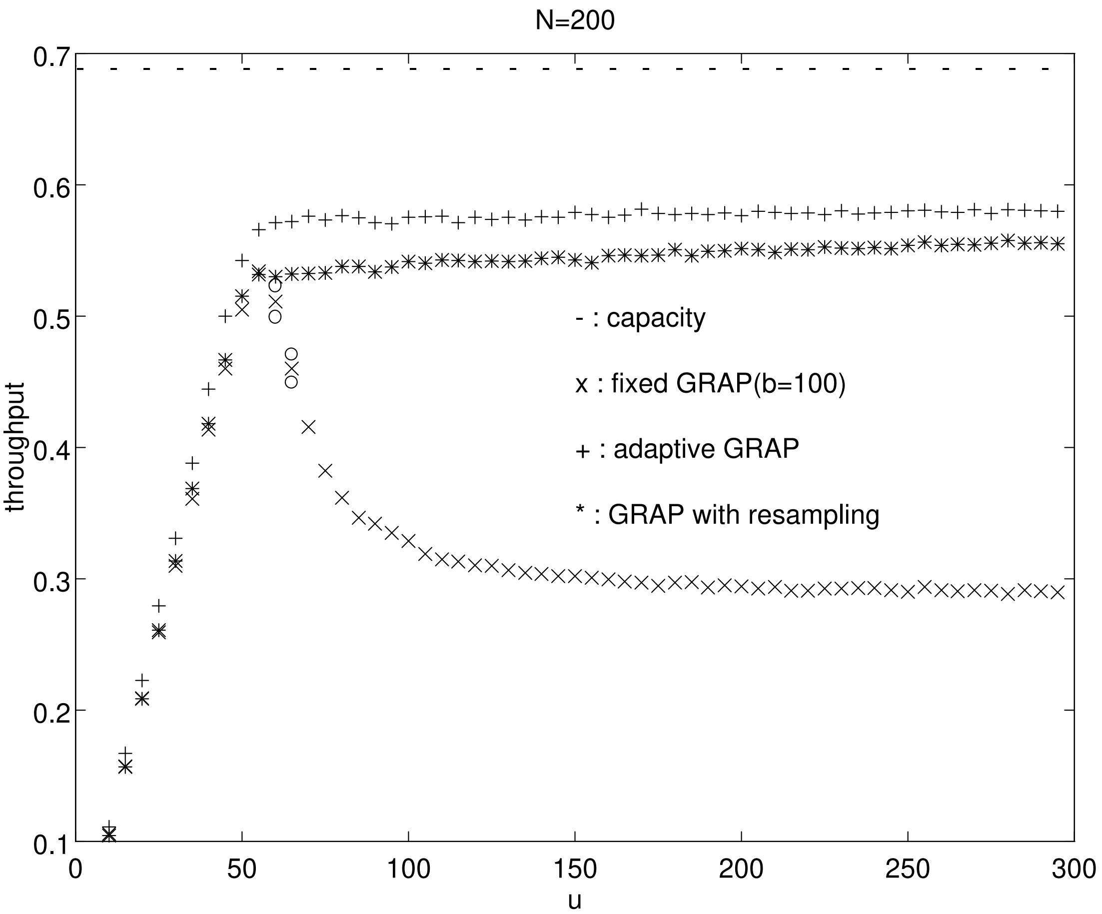

Figure 11: GRAPs with 200 ABR sources