A Novel Scheme Using the Information of Departure Processes for Delay Guarantees of Distributed VBR Traffic

Zhi-Ren Chang I-Chung Lee Cheng-Shang Chang
Chien-Hsin Li and Ben-Li Sui

Abstract— **In this paper, we consider the problem of providing delay guarantees in a *distributed* environment, e.g., a wireless network or a cable network. In a distributed environment, the information of the arrival process is in general not available to the network. Due to the lack of such information, traffic regulators and scheduling policies discussed in the literature cannot be directly applied. To cope with the problem, we propose a *distributed* traffic regulator that uses the information of the *departure* process. Based on such distributed traffic regulators, we propose the Distributed Earliest Deadline First (DEDF) scheduling policy. For the DEDF scheme, we derive an admission control criterion and show that the maximum delay can be guaranteed if the criterion is satisfied.**

Keywords— **Distributed systems, traffic regulation, scheduling, admission control**

I. Introduction

Future wireless networks are expected to provide a wide variety of services through high data rate wireless links. Several classes of services, ranging from voice to data and multimedia services, are currently being standardized (see e.g., [19], [20]). The problem of providing quality-of-service (QoS) guarantees in a *distributed* environment, such as wireless networks and cable networks, has become an important research problem. Traditionally, such a problem is treated with peak rate allocation, such as R-ALOHA in [3] and PRMA in [15]. Peak rate allocation is most suitable for Constant-Bit-Rate (CBR) traffic and it may not be efficient for Variable-Bit-Rate (VBR) traffic. On the other hand, scheduling policies that are efficient for VBR traffic in a centralized environment, such as First-Come First-Served (FCFS), Earliest Deadline First (EDF), Static Priority(SP), Virtual Clock (VC) [24], Packetized Generalized Processor Sharing (PGPS) [18] and Service Curve Earliest Deadline first (SCED) [11], [21], are difficult to implement in a distributed environment as they all need the information of packet arrival times. Such information is in general not available in a distributed environment without explicitly and constantly exchanging information. As information exchange consumes bandwidth, it is essential to execute information exchange efficiently in order to provide QoS guarantees for VBR traffic in a distributed environment.

In [7], a polling scheme with nonpreemptive priority (PNP) is proposed for providing QoS guarantees for CBR and VBR traffic in a distributed environment. There they considered the up link of a multiple access channel coordinated by a system controller, called *base station* (or *head-end controller*). The PNP scheme is a polling token based priority scheme, where the priority of polling tokens is assigned in the order of CBR sources, VBR sources and non-realtime traffic. The base station sends out polling tokens to the sources. Packets are sent to the base station only when polling tokens are received. There are three major drawbacks for such a scheme:

(i) As it is a priority scheme, it suffers from the drawback of fairness for homogeneous sources. Homogeneous sources do not receive equal shares of bandwidth.
(ii) As shown from the simulation in [7], its admission control criterion may be too conservative for VBR sources. The number of VBR sources that can be admitted is quite limited.
(iii) Traffic regulators are implemented at VBR users. The base station has to trust all the VBR sources to be cooperative in order to guarantee QoS. As a result, there is no firewall protection from a malicious user (a user pumps more traffic than allowed by the contract).

To cope with these problems in the PNP scheme, we propose in this paper the Distributed Earliest Deadline First (DEDF) scheduling algorithm and derive its associated admission control criterion. As our scheme is based on EDF, it provides fair access to homogeneous users. Also, traffic regulators are implemented at the base station to provide firewall protection from malicious users. This is done by observing departure processes and exchanging information efficiently. Our simulation also shows improvement of bandwidth utilization. The offered load in our simulation can be as high as 86%. At such high load, packets suffer from random queueing delay, but they are still within their provable delay guarantees. We note the DEDF scheme is in fact a *centralized* scheduling scheme at the base station. The word "distributed" is for distributed VBR sources as the Distributed-Queueing Request Update Multiple Access (DQRUMA) scheme in [16].

The paper is organized as follows. In Section II, we for-

Winbond Electronics Corp., No. 4, Creation Rd. III, Science-Based Industrial Park, Hsinchu, Taiwan, R.O.C. cjchang2@winbond.com.tw
Dept. of Electrical Engineering, National Tsing Hua University, Hsinchu 30013 Taiwan, R.O.C. Email: iclee@gibbs.ee.nthu.edu.tw
Corresponding author. Dept. of Electrical Engineering, National Tsing Hua University, Hsinchu 30013 Taiwan, R.O.C. Email:cschang@ee.nthu.edu.tw
PRIMAX Electronics Ltd., 669 Ruey Kuang Road, Neihu 114, Taipei, Taiwan, R.O.C. Email: james.lee@primax.com.tw
Institute for Information Industry, 5FL., 116, Sec.2, Nan-King E.RD., Taipei, Taiwan, R.O.C. Email: sbli@netrd.iii.org.tw

malize the problem. In Section III, we briefly review the filtering theory for deterministic traffic regulation and service guarantees that will be used in this paper. In Section IV, we discuss how to regulate the traffic from a VBR source in a distributed network. We propose two distributed traffic regulators and analyze their delay and output burstiness. The first one, called DTR A in the paper, is capable of knowing whether the VBR source has packets to send by sensing carrier in a separate channel. DTR A uses such information and the information of the departure process to implement the maximal f-regulator. The second one, called DTR B in the paper, detects the time that the VBR source has packets to send with limited delay. Once such an instant is detected, DTR B then follows the algorithm in DTR A for traffic regulation. Intuitively, DTR B can be viewed as a concatenation of a device with bounded delay and DTR A. In Section V, we propose the DEDF scheme. The scheme is a combination of DTR B and the EDF scheduler. It performs distributed traffic regulation and scheduling simultaneously. To guarantee the delay for each VBR source in such a scheme, we derive a sufficient condition for admission control. Under such a condition, we show that all the VBR sources satisfy their delay constraints. We program the DEDF scheme and carry out various simulations for the scheme in Section VI. Simulation results show substantial improvement for fairness over the priority scheme in [7]. We conclude the paper in Section VII by discussing future research problems of the scheme.

II. Problem formulation and assumptions

As the network model in [7], [25], we consider a multiple access channel with distributed VBR sources (CBR sources can be viewed as special cases of VBR sources). Assume that these VBR sources communicate with each other via a system controller, called *base station* (or *head-end controller*). All packets are sent to the base station first and then relayed to their destinations. All the sources can send packets to the base station through a multiple access channel. This channel is called *up link*. On the other hand, the base station broadcasts the packets to all VBR sources through a broadcasting channel, called *down link*. The down link has almost the same characteristic as a general centralized link, where information, including packet arrival times, is available to the base station. Hence, we will focus on the up link, where information is distributed and kept to each user.

We assume that both the up link and the down link have zero propagation delay and that all the packets of these VBR sources are of the same size. The assumption of zero propagation delay may look unrealistic at the first glance as propagation delay may vary in a mobile environment. However, when the variation of propagation delay is negligible comparing to the transmission time of a packet, the propagation delay can be added into the transmission time of a packet in the analysis.

Packets generated by a VBR source is stored at a local buffer before transmission. The time a packet is generated is called its arrival time, and the time it reaches the base station is called its departure time. We assume the local buffer size is infinite (large enough to accommodate for the traffic generated by the VBR source). For each packet of a VBR source, there is a maximum delay constraint. The delay of a packet is defined to be the difference between its departure time and its arrival time.

As in a centralized link, one needs to carry out the following in order to guarantee maximum delay:

(i) Traffic regulation:

Each VBR source needs to be regulated before entering the network. If one builds the regulators at the users, then the base station has to trust all the VBR sources to be cooperative. This is the assumption used in [7], [25]. One serious drawback for such an assumption is that there is no firewall protection from a malicious user (a user pumps more traffic than allowed by the contract). In this paper, we consider having traffic regulators in the base station so that QoS can be guaranteed for cooperative VBR sources even though some of them are not. To characterize a VBR source, we adopt the concept of upper envelope in [9], [10], [4]. A VBR source A is f-upper enveloped if f is increasing and subadditive, i.e., $f(t_1) \leq f(t_2)$, $f(t_1) + f(t_2 - t_1) \leq f(t_2)$ for all $t_1 \leq t_2$, and A satisfies

$$A(t_2) - A(t_1) \leq f(t_2 - t_1), \quad \forall t_1 \leq t_2, \tag{1}$$

where $A(t)$ is the cumulative arrivals by time t. As discussed in [5], an f-upper enveloped source can be policed by feeding a source through the maximal f-regulator, i.e.

$$\tilde{B}(t) = \min_{0 \leq s \leq t} \left[A(s) + f(t-s) \right], \tag{2}$$

where $A(t)$ (resp. $\tilde{B}(t)$) is the cumulative arrivals (resp. departures) from the maximal f-regulator by time t. Note that the maximal f-regulator in (2) requires the information of the arrival process and such information is not available to the base station. The question is then how one implements such a regulator in the base station, where the information of the arrival process is not available.

(ii) Traffic scheduling:

In a centralized link, many scheduling policies such as First-Come First-Served (FCFS), Earliest Deadline First (EDF), Static Priority(SP), Virtual Clock (VC) [24], Packetized Generalized Processor Sharing (PGPS) [18] and Service Curve Earliest Deadline first (SCED) [11], [21], can be used for delay guarantees. All these scheduling policies need the information of the arrival process. As in traffic regulation, the question is how one implements such a scheduling policy in a distributed environment where the information of the arrival process is not available.

(iii) Admission control:

When a VBR source makes a request for connection, it needs to provide the base station its traffic characterization, including the traffic envelope f and the delay constraint. Based on the traffic characterization, the base station checks whether it can guarantee the delay of the existing connections and the new connection. If this is the case, the new connection is admitted. The problem is then to find the admission control criterion that guarantees the delay of all sources.

III. Review of the filtering theory

Our proposed solution to the problems in Section II is the Distributed Earliest Deadline First (DEDF) scheme. Our analysis for the DEDF scheme is based on the recently developed filtering theory for deterministic traffic regulation and service guarantees in [5], [12], [1], [2]. The main advantage of using such a theory is that it enables us to analyze the DEDF scheme systematically. Before we introduce the DEDF scheme, we first briefly review the filtering theory in this section (for more details of the theory, we refer to the book [6]). The filtering theory is developed under the $(\min, +)$-algebra, where one replaces the usual addition by the min operator and the usual multiplication by the addition operator. As in the classical linear system theory, the new filtering theory treats an arrival process A (or a departure process B) as a signal, and a network as a system. A (continuous-time) signal $f \equiv \{f(t), t \geq 0\}$ is a nonnegative and increasing function. A signal is said to be larger than another if all its values at any time t are larger. Two basic operations for signals under the $(\min, +)$-algebra are considered, the "addition" operation $\oplus$ and the "multiplication" operation $\star$.

(i) (min) the pointwise minimum of two signals:

$$(f \oplus g)(t) = \min[f(t), g(t)].$$

(ii) (convolution) the convolution of two signals:

$$(f \star g)(t) = \min_{0 \leq s \leq t}[f(s) + g(t-s)].$$

These two operations are associative, commutative and distributive, and one may use them as the usual addition and multiplication.

Based on the two operations for signals, two types of basic network elements are defined: the maximal f-regulator and the f-server. The maximal f-regulator with the input A yields the output $B = A \star f$ (when f is subadditive), and the f-server for the input A guarantees the output $B \geq A \star f$. The maximal f-regulator has the following three properties:

(TR 1) Traffic regulation: the output from the maximal f-regulator is f-upper enveloped for any input.

(TR 2) Optimality: the maximal f-regulator is the best causal traffic regulator that one can implement in terms of maximizing the number of cumulative departures from the regulator at any moment in time.

(TR 3) Conformity: if the input to the maximal f-regulator is already f-upper enveloped, then it is not affected by the regulator.

An ideal (σ, ρ)-leaky bucket is a special case of the maximal f-regulator with $f(t) = \sigma + \rho t$.

Schedulers, such as EDF [14], VC [24], [13], [22], PGPS [18] and SCED [11], [21], can be characterized by f-servers for certain f's. The representation of a server is not unique, and it may depend on the input. Certainly, the maximal f-regulator is an f-server. Among all f-servers, the O_d-server (with $O_d(t) = 0$ for $0 \leq t \leq d$ and $O_d(t) = \infty$ otherwise) is of particular interest. A server is an O_d-server for an input if and only if the server guarantees maximum delay d for that input.

Network elements can be joined by concatenation, "filter bank summation", and feedback to form a composite network element.

(i) Concatenation: a concatenation of an f_1-server for an input A and an f_2-server for the output from the f_1-server is an f-server for A, where $f = f_1 \star f_2$.

(ii) "filter bank summation": the "filter bank summation" of an f_1-server for A and an f_2-server for A is an f-server for A, where $f = f_1 \oplus f_2$.

(iii) Feedback: the feedback of an f-server is an f^*-server for A if $f(0) > 0$ and $A(t) < \infty$ for all t, where f^* is the subadditive closure of f in [5].

The maximum queue length (resp. maximum delay) of an f_2-server with an f_1-upper enveloped input is bounded by the maximum vertical (resp. horizontal) distance between f_1 and f_2. Moreover, the output is f_3^*-upper enveloped, where

$$f_3(t) = \sup_{s \geq 0}[f_1(t+s) - f_2(s)]. \tag{3}$$

IV. Distributed traffic regulation

A. Distributed traffic regulation with a separate carrier sensing channel

Our objective in this section is to propose a scheme that can implement the maximal f-regulator in a distributed environment. As discussed in the introduction, the maximal f-regulator in (2) requires the information of the arrival process, which is in general not available to the base station. The base station may acquire such information by polling the source. However, this consumes bandwidth and exchange of such information should be carried out only if needed.

Our first observation is that the base station has the information of the departure process. Though the base station cannot observe the arrival process directly, it can use the information of the departure process to deduce the information of the arrival process. For instance, let $A(t)$ and $B(t)$ be the cumulative arrivals and departures by time t. If the base station polls the source and the source replies with an end-of-file (EOF) message (no more packet to transmit at the time of polling), then the base station knows $A(t) = B(t)$ at that instant. In our setting, we assume that both $A(t)$ and $B(t)$ are left continuous.

Our second observation is that the maximal f-regulator may not need all the information of the arrival process. In the following, we show that one can implement the maximal f-regulator if the base station is capable of knowing whether the VBR source has a packet to send (it does not know how many of them). Such information can be obtained by sensing carrier in a separate channel. When the VBR source has at least a packet to send, it transmits a carrier at a certain frequency. We define a busy period of the VBR source the period of time that the base station detects such a frequency. The time between two successive busy periods is called an idle period. Define τ_k^b (resp.

τ_k^e) be the starting (resp. finishing) time of the k^{th} busy period. Consider the following distributed traffic regulator.

Distributed Traffic Regulator A (DTR A):

The VBR source is allowed to send a packet only when it receives a token from the base station. Let $B(t)$ be the cumulative departures by time t from the traffic regulator. In the k^{th} busy period, tokens are generated and sent at $\tau_{k,j}$, $j = 1, 2, \ldots$, till the end of the busy period, where

$$\tau_{k,j} = \inf\left\{t > \tau_k^b : g_k(t) \geq B(\tau_k^b) + j\right\}, \qquad (4)$$

and

$$g_k(t) = \min_{1 \leq \ell \leq k} \left[B(\tau_\ell^b) + f(t - \tau_\ell^b)\right]. \qquad (5)$$

Note that $g_k(t)$ is the number of tokens generated by time t if t is in k^{th} busy period ($\tau_k^b < t \leq \tau_k^e$).

Remark 1: At the first glance, one might think the inversion in (4) is difficult and the complexity for implementing such a regulator is high. This is not the case when the envelope f has a certain form. For example, if we consider the (σ, ρ)-leaky bucket, i.e., $f(t) = \rho t + \sigma$ for all $t > 0$, then

$$g_k(t) = g_k + \rho t + \sigma, \qquad (6)$$

where $g_k = \min_{1 \leq \ell \leq k}[B(\tau_\ell^b) - \rho \tau_\ell^b]$. Moreover, g_k can be computed recursively by

$$g_k = \min[g_{k-1}, B(\tau_k^b) - \rho \tau_k^b]. \qquad (7)$$

From (6), one then has

$$\tau_{k,j} = \max\left[\frac{B(\tau_k^b) - g_k - \sigma + j}{\rho}, \tau_k^b\right]. \qquad (8)$$

Thus, the base station only needs to store three parameters: (i) the starting time of the current busy period, τ_k^b, (ii) the cumulative departures by the starting time of the current busy period, $B(\tau_k^b)$, and (iii) the state of the recursive equation in (7), g_k.

More generally, we can consider the concatenation of m (σ_ℓ, ρ_ℓ)-leaky buckets (see e.g., [5]), i.e., $f(t) = \min_{1 \leq \ell \leq m}[\rho_\ell t + \sigma_\ell]$. One can easily show that

$$\tau_{k,j} = \max\Big[\max_{1 \leq \ell \leq m}\big[\frac{B(\tau_k^b) - g_{\ell,k} - \sigma_\ell + j}{\rho_\ell}\big], \tau_k^b\Big], \qquad (9)$$

where

$$g_{\ell,k} = \min[g_{\ell,k-1}, B(\tau_k^b) - \rho_\ell \tau_k^b], \quad \ell = 1, \ldots, m. \qquad (10)$$

In the following theorem, we state that DTR A is indeed the maximal f-regulator. The proof of the theorem can be found in the book [6], Problem 2.3.

Theorem 2: Let

$$\tilde{B}(t) = (A \star f)(t) = \min_{0 \leq s \leq t}[A(s) + f(t - s)]$$

be the output from maximal f-regulator (f is increasing and subadditive). Then $\tilde{B}(t) = B(t)$ for all t and DTR A is the maximal f-regulator.

Since DTR A is shown to be the maximal f-regulator, the output from DTR A is f-upper enveloped and it has no delay for an f-upper enveloped input.

B. Distributed traffic regulation with inband signaling

In the previous section, we assume that there is a separate carrier sensing channel that is capable of knowing whether the VBR source has a packet to send. However, assuming the existence of a separate carrier sensing channel may not be realistic. Instead, we propose in this section a scheme with inband signaling. As in Section IV-A, we are only dealing with a single VBR source in this section. The general case with multiple VBR sources will be considered in the next section.

The basic idea of our scheme is for the base station to detect the instant that the VBR source has a packet to send with limited delay. To do this, the VBR source is classified into two states: the idle state and the busy state. The base station puts the VBR source in the idle (resp. busy) state if it thinks the VBR source does not have (resp. has) a packet to send. It may happen that the VBR source does have a packet to send while the base station puts it in the idle state. To correct such an error, the base station limits the length of the period that the VBR source can be put in the idle state, i.e., the base station switches the VBR user into the busy state if the user has been put in the idle state over p time units, where p is the parameter designed by the base station. (In Section V-B, the parameter p is chosen to maximize the number of connections that can be admitted to the network.) It is clear that the delay to detect the instant that the VBR source has a packet to send is then bounded above by the parameter p. Once the VBR source is switched into the busy state, polling tokens are generated according to the algorithm in DTR A so that the output is still well regulated. The VBR source may notify the base station that it does not have any more packets to send by appending an end-of-file (EOF) message in the last packet or simply sending an EOF message if it does not have any packets to send (inband signaling). When this happens, the base station then switches the VBR source back to the idle state.

To be precise, we list the polling scheme as follows:

Distributed Traffic Regulator B (DTR B):

1. The base station maintains two states for the VBR source : the idle state and the busy state.

2. When the VBR source receives a polling token form the base station, consider the following conditions of the buffer in the VBR source.

2.1. Suppose that there are more than one packets in the buffer of the VBR source. After being polled, the VBR source transmits a packet.

2.2. Suppose that there is only one packet in the buffer of the VBR source. After being polled, the VBR source transmits a packet appended with an EOF message. When a packet appended with an EOF message is received, the base station changes the state from the busy state to the idle state at the current time. If this is the k^{th} time to receive a packet appended with an EOF message, it sets τ_k^e to be the current time. Furthermore, the base station sets τ_{k+1}^b to be the sum of the current time and p, and changes the state of the VBR source from the idle state to the busy state at the time τ_{k+1}^b.

2.3. Suppose that there are no packets in the buffer of the VBR source. After being polled, the VBR source replies with an EOF message. When an EOF message is received, the base station changes the state from the busy state to the idle state at the current time. If it has received k packets appended with EOF messages by the current time, it *resets* τ^b_{k+1} to be the sum of the current time and p, and changes the state of the VBR source from the idle state to the busy state at the time τ^b_{k+1}.
3. When the VBR connection is set up, the base station sets its initial state to be idle and sets τ^e_0 to be the current time. Furthermore, it sets τ^b_1 to be the sum of the current time and p, and changes the state from the idle state to the busy state at the time τ^b_1.
4. Let $B(t)$ be the cumulative number of packets received by time t from the VBR source. In the k^{th} time of the busy state, the polling tokens are generated and sent at $\tau_{k,j}$, $j = 1, 2, \ldots$, where

$$\tau_{k,j} = \inf \left\{ t > \tau^b_k : g_k(t) \geq B(\tau^b_k) + j \right\}, \tag{11}$$

and

$$g_k(t) = \min_{1 \leq \ell \leq k} \left[B(\tau^b_\ell) + f(t - \tau^b_\ell) \right]. \tag{12}$$

In our analysis, tokens can be divided into two classes. A token is called **effective** if the base station receives a packet or a packet appended with an EOF message when the token is served. If the base station receives only an EOF message when a token is served, then the token is called **ineffective**. As tokens are served immediately after being generated, $B(t)$ is the same as the number of effective tokens served by time t.

Theorem 3: The output $B(t)$ from DTR B is the output from the concatenation of an O_p-server and the maximal f-regulator.

The proof of Theorem 3 is given in Appendix A. As direct consequences of Theorem 3, the maximum delay for DTR B is bounded by p time units for an f-upper enveloped input. Also, the output $B(t)$ from DTR B is f-upper enveloped.

V. The DEDF scheme

In this section, we propose the Distributed Earliest Deadline First (DEDF) scheme to guarantee maximum delay for multiple VBR sources.

A. The algorithm

Suppose there are N VBR sources. The i^{th} VBR source is claimed to be f_i-upper enveloped and its maximum tolerable delay is d_i, $i = 1, 2, \ldots, N$. The basic idea of our DEDF algorithm is to combine DTR B and the EDF scheduler. Suppose that one chooses the parameter $p_{i,1}$ in DTR B for the i^{th} VBR source. Every token generated by DTR B for the i^{th} VBR source is assigned a deadline that is equal to the sum of the its generation time and another design parameter $p_{i,2}$. If these tokens are served before their deadlines, then the maximum delay incurred by the EDF scheduler is bounded above by $p_{i,2}$. In addition to the $p_{i,1}$ time units incurred by DTR B, the overall delay is then bounded above by $p_{i,1} + p_{i,2}$, provided that the input is f_i-upper enveloped. To be precise, we present the scheduling algorithm as follows:

The Distributed Earliest Deadline First (DEDF) Scheme:

1. For each VBR source, there are two states: the busy state and the idle state. The base station maintains a token queue for each VBR source. A token consists of two data fields: source identification and deadline.
2. The base station searches the head of the N token queues to find the one with the earliest deadline. (If there is no such a token, the base station stays idle.) It then grabs the token and polls the corresponding VBR source (from the source identification field). Suppose that the token with the earliest deadline is a token for the i^{th} VBR source. Consider the following conditions of the buffer in the i^{th} VBR source.
2.1. Suppose that there are more than one packets in the buffer of the i^{th} VBR source. After being polled, the i^{th} VBR source transmits a packet.
2.2. Suppose that there is only one packet in the buffer of the i^{th} VBR source. After being polled, the i^{th} VBR source transmits a packet appended with an EOF message. When a packet appended with an EOF message from the i^{th} source is received, the base station *clears the token queue of the i^{th} VBR source* and changes the state of the i^{th} source from the busy state to the idle state at the current time. If this is the k^{th} time to receive a packet appended with an EOF message from the i^{th} source, it sets $\tau^e_{i,k}$ to be the current time. Furthermore, the base station sets $\tau^b_{i,k+1}$ to be the sum of the current time and $p_{i,1}$, and changes the state of the i^{th} source from the idle state to the busy state at the time $\tau^b_{i,k+1}$.
2.3. Suppose that there are no packets in the buffer of the i^{th} VBR source. After being polled, the i^{th} source replies with an EOF message. When an EOF message from the i^{th} source is received, the base station *clears the token queue of the i^{th} VBR source* and changes the state of the i^{th} source from the busy state to the idle state at the current time. If it has received k packets appended with EOF messages from the i^{th} source by the current time, it resets $\tau^b_{i,k+1}$ to be the sum of the current time and $p_{i,1}$, and changes the state of the i^{th} source from the idle state to the busy state at the time $\tau^b_{i,k+1}$.
3. When the i^{th} VBR connection is set up (admitted by the criterion in the next section), the base station sets its initial state to be idle and sets $\tau^e_{i,0}$ to be the current time. Furthermore, it sets $\tau^b_{i,1}$ to be the sum of the current time and $p_{i,1}$, and changes the state from the idle state to the busy state at the time $\tau^b_{i,1}$.
4. In the k^{th} time of the busy state of the i^{th} VBR source, the tokens are generated at $\tau_{i,k,j}$, $j = 1, 2, \ldots$, where

$$\tau_{i,k,j} = \inf \left\{ t > \tau^b_{i,k} : g_{i,k}(t) \geq B_i(\tau^b_{i,k}) + j \right\}, \tag{13}$$

and

$$g_{i,k}(t) = \min_{1 \leq \ell \leq k} \left[B_i(\tau^b_{i,\ell}) + f_i(t - \tau^b_{i,\ell}) \right], \tag{14}$$

where $B_i(t)$ (left continuous) is the number of packets received by time t from the i^{th} VBR source. When a token of the i^{th} VBR source is generated, *its deadline is set to be the sum of the current time and* $p_{i,2}$.

Since tokens may be not be served immediately after they are generated, it is necessary for us to keep a token queue for each VBR source. The token queue of a VBR source is cleared every time when the busy state is changed to the idle state.

As discussed in Remark 1, computation complexity for (13) seems high as it depends on the whole history of the departure processes. However, it can be greatly simplified by considering envelope functions that are concatenation of leaky buckets. In that case, the base station only needs to store the following parameters: (i) the starting time of the current busy period for the i^{th} source, $\tau^b_{i,k}$, $i = 1, \dots N$, (ii) the cumulative departures by the starting time of the current busy period for the i^{th} source, $B_i(\tau^b_k)$, $i = 1, \dots, N$,and (iii) the states of the recursive equations in (10).

Remark 4: In Step 2, we say the base station stays idle when there is no token in the token queues. In fact, the base station could use the time to serve other traffic without delay constraint, including the Available Bit Rate (ABR) traffic and the signaling traffic. Such traffic can be served by the usual contention based medium access protocols. For example, the Group Randomly Addressed Polling (GRAP) in [8], [7] can be easily incorporated into our scheme to serve such traffic. As our focus is on serving VBR traffic, we will not pursue this issue further. Also, note that the Constant Bit Rate (CBR) traffic discussed in [7] is a special case of the VBR traffic with f being a linear function.

In our scheme, tokens can be divided into three classes. A token is called **effective** if the base station receives a packet or a packet appended with an EOF message when this token is served. If the base station only receives an EOF message when a token is served, then the token is called **ineffective**. At last, a token is classified as an **unserved** token if it is generated but not served by the base station.

Let $T_{i,B}(t)$ (resp. $T_{i,I}(t)$) denote the number of (resp. ineffective) tokens for the i^{th} source that are generated by time t. Assume that $T_{i,B}(t)$ and $T_{i,I}(t)$ are left continuous.

The following theorem is the main result for the DEDF scheme. Its proof is given in Appendix B.

Theorem 5: If all the *effective* tokens are served before their deadlines, then the DEDF scheme is an $\tilde{f}_i$-server for the i^{th} VBR source, where $\tilde{f}_i = O_{p_{i,1}+p_{i,2}} \star f_i$, $i = 1, \dots, N$. Moreover, the departure process B_i of the i^{th} VBR source from the DEDF scheme is $\hat{f}_i$-upper enveloped, where $\hat{f}_i(t) = f_i(t + p_{i,2})$. If, furthermore, A_i is f_i-upper enveloped, then the maximum delay for the i^{th} VBR source is bounded above by $p_{i,1} + p_{i,2}$.

Thus, if A_i is f_i-upper enveloped and $p_{i,1} + p_{i,2} \le d_i$, then all the packets from the i^{th} VBR source satisfy their delay constraints. This is true even when other sources do not conform to their traffic envelopes. Moreover, its output is still well regulated by $\hat{f}_i$. This can be further used for service guarantees in the following stages of its route. To visualize how the DEDF scheme works, we depict the block diagram of the DEDF scheme in Figure 1.

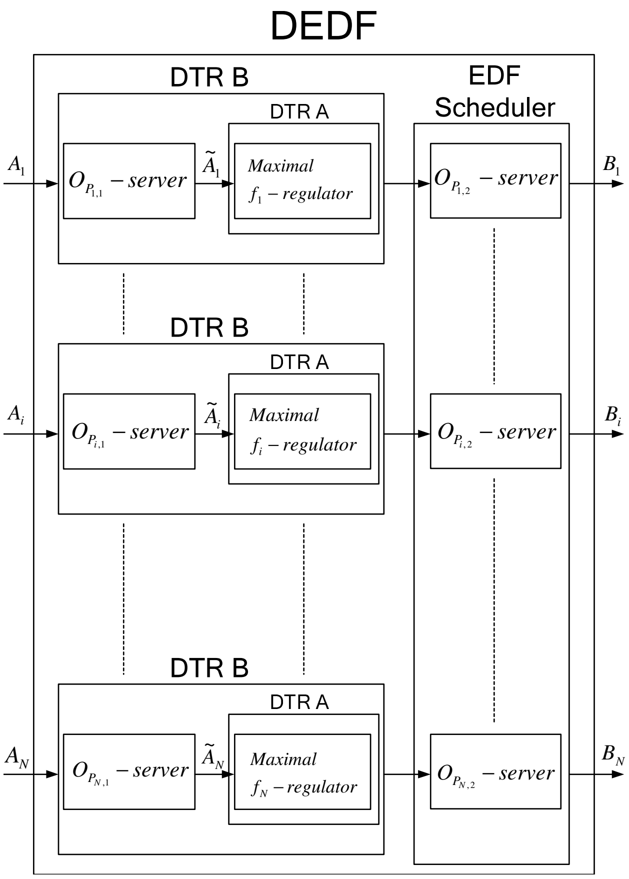

Fig. 1. The block diagram of the DEDF scheme

B. Admission control

In this section, we derive a sufficient condition for all the effective and ineffective tokens to be served before their deadlines. As the cumulative effective tokens generated by time t, i.e., $T_{i,B}(t)$, is the output from DTR B (with $p = p_{i,1}$) for the input A_i, $T_{i,B}$ is f_i-upper enveloped. Moreover, the cumulative ineffective tokens is bounded above by the worst case when they are periodically generated. Let t_I be the service time for serving an ineffective token and t_B be the service time for serving an effective token. Base on these two facts, we then have the sufficient condition

$$\sum_{i=1}^{N} f_i(t - p_{i,2} + t_B)t_B + \sum_{i=1}^{N} \lceil \frac{1}{p_{i,1}}(t - p_{i,2} + t_B) \rceil t_I \le t, \quad (15)$$

for all the *effective* and *ineffective* tokens to be served before their deadlines. We state this formally in Theorem 6. The proof of Theorem 6 is given in Appendix C.

Theorem 6: Under the condition in (15), all the *effective* and *ineffective* tokens are served before their deadlines.

Note that in order for the base station to carry out the admission test, it has to have the information of the traffic envelope and the delay constraint of a VBR source. Such information can be transmitted via a contention based medium access protocol as described in Remark 4.

Though we only need to guarantee that the *effective* tokens are served before their deadlines in Theorem 5, we still need to guarantee both the *effective* and *ineffective* tokens as we are not able to distinguish a token until it is served. This makes our bounds a little bit conservative as illustrated in our simulations in Section VI.

One may define the (normalized) effective peak rate of the i^{th} VBR source as follows:

$$r_i = \sup_{t\geq 0} \frac{1}{t}\Big(f_i(t-p_{i,2}+t_B)t_B + \lceil \frac{1}{p_{i,1}}(t-p_{i,2}+t_B)\rceil t_I\Big). \tag{16}$$

It is easy to see that the admission criterion in (15) is satisfied if

$$\sum_{i=1}^{N} r_i \leq 1. \tag{17}$$

Moreover, one can choose the design parameter $p_{i,1}$ and $p_{i,2}$ such that the effective peak rate r_i is minimized under the constraint $p_{i,1}+p_{i,2} \leq d_i$. For example, if $f_i(t) = \rho_i t + \sigma_i$, then using the inequality $\lceil x\rceil \leq x+1$ yields

$$r_i \leq \max[\frac{\sigma_i t_B + t_I}{p_{i,2}-t_B}, \rho_i t_B + \frac{1}{p_{i,1}} t_I]. \tag{18}$$

One can then minimize the right side of (18) under the constraint $p_{i,1}+p_{i,2} \leq d_i$. Note that if all the VBR sources are homogeneous, then the effective peak rate criterion in (17) is the same as that in (15) and optimizing the parameter $p_{i,1}$ and $p_{i,2}$ in (18) yields the maximum number of VBR sources that can be admitted to the network.

VI. Simulation results

In this section, we perform three experiments to verify the DEDF scheme. In our experiments, we assume that the speed of the multiplexing channel is 10 $Mbits/sec$. All the packets are of the same length 1k bits, and all EOF messages are 50 bits. The length of each polling packet from the base station is also 50 bits. Thus, $t_B = 105$ μsec (a packet appended with an EOF message) and $t_I = 10$ μsec (a polling message + an EOF message).

The contract between the i^{th} VBR source and the base station is characterized by (ρ_i, σ_i, d_i) with the traffic envelope $f_i(t) = \rho_i t + \sigma_i$ and the maximal tolerable delay d_i. For each VBR source, we model its traffic as the output from the concatenation of an ON-OFF coder and a leaky bucket (See Figure 2). We assume that the coding speed of the i^{th} coder is c_i $kbits/sec$ and both the ON and OFF periods are $1/c_i$ sec, which is the period to generate a packet. Thus, the coder generates one packet if the period is ON. Otherwise, no packet is generated. In order to model the ON-OFF coder, we flip a coin to decide whether the next period is ON or OFF at the end of the current period. Let $p_{i,on}$ be the probability that the next period is an ON period.

The i^{th} ON-OFF coder is followed by a leaky bucket with the token generate rate ρ_i and the token buffer size $\sigma_i - 1$. It is known that the number of outputs from such a leak bucket during an interval of length t is bounded by $\lceil \rho_i t + \sigma_i - 1\rceil$. Thus, the output from such a leaky bucket is f_i-upper enveloped with $f_i(t) = \rho_i t + \sigma_i$ and it conforms to the traffic contract. The delay of a packet is measured between the time it departs from the leaky bucket and the time it is received by the base station.

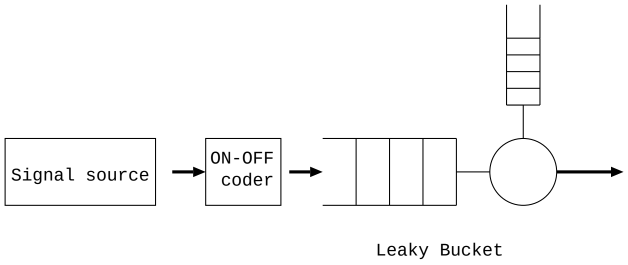

Fig. 2. The traffic model for a VBR source in the experiments

In the first experiment, we consider the case where all the VBR sources are homogeneous. We choose $p_{i,on} = p_{on} = 0.48$, $c_i = c = 900$ $kbits/sec$, $\sigma_i = \sigma = 24$, $\rho_i = \rho = 450$ $packets/sec$ and $d_i = d = 0.12$ sec for all i. The parameters $p_{i,1} = p_1$ and $p_{i,2} = p_2$ (for all i) are chosen so that the number of users (denoted by N) that can be admitted to the network is maximized. To do this, we use the effective peak rate criterion from (17) and (18) to derive the following inequality

$$N\big(\max[\frac{\sigma t_B + t_I}{p_2 - t_B}, \rho t_B + \frac{1}{p_1} t_I]\big) \leq 1 \tag{19}$$

Thus, from (19) and $p_1 + p_2 \leq d$, we find that the maximum number of VBR users admitted to the network N is 20. The corresponding p_1 and p_2 are 0.06667 sec and 0.05333 sec. Note that $p_1 + p_2 = d = 0.12$ sec. In this case, the total average offered load from the N VBR sources (without counting the traffic from polling and EOF messages) is $c \times p_{on} \times N = 8.6$ $Mbits/sec$. Thus, the channel utilization is at least 86%. From our experiment, the maximum delay (observed during the period of the simulation) ranges between 0.0705 sec and 0.0821 sec for these 20 sources. This shows that our deterministic bound is still conservative as discussed in Section V-B. In Figure 3 and Figure 4, we further plot the tail distributions of the delay of the packets for these 20 sources. From the tail distributions, we not only justify that all the packets satisfy their delay constraints, but also demonstrate the fairness among these 20 homogeneous sources as they all have similar distributions. This is a considerable improvement over the priority scheme in [7], where low priority sources suffer from large delay. Also, the offered load in the priority scheme in [7] has to be kept low due to the tight admission criterion. In the simulation in [7], the tail distributions are all uniformly distributed as there is almost no queueing delay and all the packets are simply waiting for the arrivals of their polling tokens. In the DEDF scheme, the tail distributions are not uniform. This shows that packets suffer from random queueing delay in the DEDF scheme.

To further understand the the trade-off between p_1 and p_2, we perform our second experiment. In the second ex-

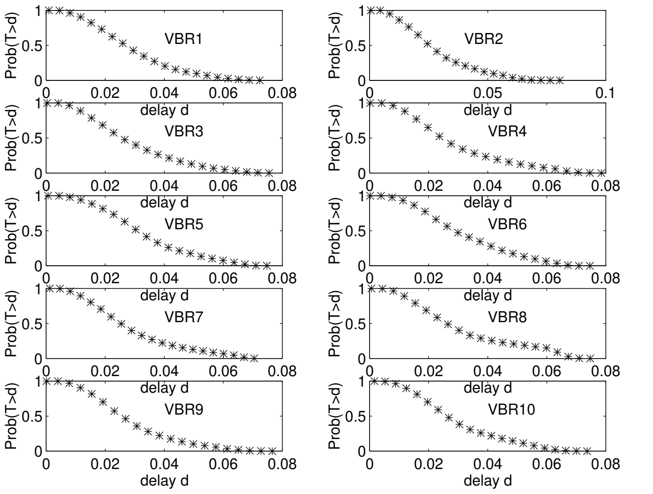

Fig. 3. The tail distributions of the delay for VBR sources 1-10 in the first experiment

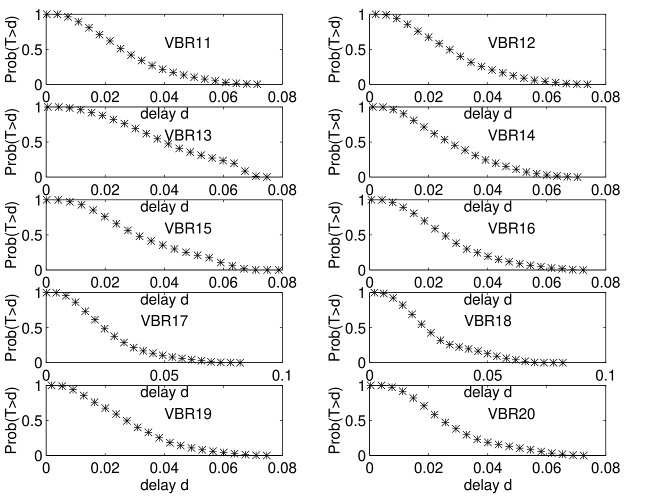

Fig. 4. The tail distributions of the delay for VBR sources 11-20 in the first experiment

periment, we use the same traffic model in the first experiment. We assume there are 10 VBR sources, i.e., $N = 10$. In this experiment, the average offered load (without counting the traffic from polling and EOF messages) is roughly 43%. We vary p_1 and p_2 on the line $p_1 + p_2 = d$, and collect two simulation results: the (observed) maximal delay and the waste ratio (the ratio of the time consumed by the ineffective tokens to the total simulation time in percentage). The total number of served packets for each VBR source is about 20,000. We report the results in Table I. Since the offered load in this experiment is not as heavy as that in the first experiment, tokens are served almost immediately after they are generated. As a result, the major portion of the delay comes from the delay to generate a token, i.e., p_1. This explains why the observed maximum delay is closed to p_1 in Table I. A more interesting observation is that the waste ratio is insignificant for a fairly large range of p_1. The waste ratio becomes significant when p_1 is close to the inverse of average traffic contract rate, $1/\rho = 0.0022$ sec. In this case, the base station is polling too often and there is a significant amount of ineffective tokens. This experiment also confirms that our choice of $p_1 = 0.06667$ sec in the first experiment is a good choice. Such a choice does not waste too much bandwidth. It also maximizes the number of VBR users that can admitted to the network.

system parameters (p_1, p_2) $(secs)$	maximal delay d_{max} $(secs)$	waste ratio (%)
(0.11, 0.01)	0.112713	0.00
(0.10, 0.02)	0.104857	0.00
(0.09, 0.03)	0.091514	0.00
(0.08, 0.04)	0.084087	0.00
(0.07, 0.05)	0.074288	0.00
(0.06, 0.06)	0.065211	0.00
(0.05, 0.07)	0.056979	0.00
(0.04, 0.08)	0.046723	0.00
(0.03, 0.09)	0.032149	0.00
(0.02, 0.10)	0.022319	0.00
(0.01, 0.11)	0.011514	0.0026
(0.005, 0.115)	0.005907	0.0943
(0.0025, 0.1175)	0.003398	0.8164
(0.00125, 0.11875)	0.002090	3.4519
(0.000625, 0.119375)	0.001389	9.59

TABLE I

THE MAXIMAL DELAY AND WASTE-RATIO FOR DIFFERENT (p_1, p_2) COMBINATIONS IN THE SECOND EXPERIMENT

In the third experiment, a non-homogeneous case is considered. There are two types of different VBR traffic models. The coder speed c_i of the first (resp. second) type is $2000 kbits/sec$ (resp. $64 kbits/sec$) with $p_{i,on} = 0.72$ (resp. $p_{i,on} = 0.75$). The contract for the first type is $(\rho_i, \sigma_i, d_i) = (1500\ packets/sec, 6, 0.03\ sec)$, and the contract for the second type is $(\rho_i, \sigma_i, d_i) = (50\ packets/sec, 4, 0.15\ sec)$. Note that the first type has a more stringent delay constraint. The parameters $(p_{i,1}, p_{i,2})$ are chosen in the way that the effective peak rate in (19) is minimized under the constraint that $p_{i,1} + p_{i,2} \le d_i$. For the first (resp. second) type, it is found that $p_{i,1} = 0.02585\ sec$ and $p_{i,2} = 0.00415\ sec$ (resp. $p_{i,1} = 0.0695\ sec$ and $p_{i,2} = 0.0805\ sec$). We assume 15 users are admitted: the first six are in first type (type I) and the others are in the second type (type II). In this case, the admission criterion in (15) is satisfied. The total average offered load for this case is

$$2000\ kbits/sec \times 0.72 \times 6 + 64\ kbits/sec \times 0.75 \times 9$$
$$= 9.072\ Mbits/sec.$$

Thus, the channel utilization is at least 90.72%. The maximum delay observed for each source during the simulation is reported in Table II. We also plot the tail distributions of the delay of the packets in the third experiment in Figure 5. Once again, this demonstrates the improvement of fairness within each type as they have similar distributions within the same type.

Sources $N = 15$	traffic types	maximal delay d_{max} $(secs)$
1	I	0.0267
2	I	0.0276
3	I	0.0262
4	I	0.0260
5	I	0.0261
6	I	0.0260
7	II	0.0909
8	II	0.0818
9	II	0.0906
10	II	0.0860
11	II	0.1114
12	II	0.0793
13	II	0.0801
14	II	0.0799
15	II	0.1004

TABLE II
THE PARAMETERS AND THE MAXIMAL DELAY FOR THE VBR SOURCES IN THE THIRD EXPERIMENT

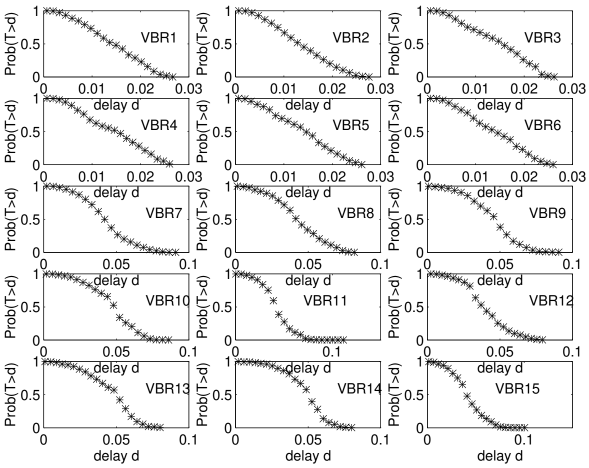

Fig. 5. The tail distributions of the delay for VBR sources 1-15 in the third experiment

VII. CONCLUSIONS AND FUTURE RESEARCH

In this paper, we proposed the DEDF scheme for providing delay guarantees in a distributed environment. To our best knowledge, this is the first scheme that uses the information of the departure processes to provide delay guarantees among all the time-limited polling schemes in the literature. The DEDF scheme consists of two parts: (i) DTR B and (ii) a token based EDF scheduler. DTR B, as a distributed traffic regulator, ensures that the traffic coming into the network is well regulated. This not only acts as a policing device for a VBR source, but also provides firewall protection from malicious users and catastrophic transmission errors. As DTR B also uses the information of the departure process of a VBR source to compute the token generation times, bandwidth utilization is more efficient than fixed (peak) rate regulation, where tokens are generated periodically. We also showed in Remark 1 that the token generation times can be computed easily if the envelope is a concatenation of leak buckets. For such an envelope, the computational complexity is of the same order as that of Virtual Clock. Under the admission criterion in (15), the token based EDF scheduler then ensures that all the tokens are served before their deadlines, which in turn guarantees that all the packets satisfy their delay constraints. As most EDF schedulers, the token based EDF scheduler can provide a continuous spectrum of delay bounds for VBR sources. Our experimental results show substantial improvement of fairness over the priority scheme in [7].

There are some interesting problems that arise from this research.

(i) Retransmission or forward error correction:
Though the DEDF scheme guarantees that all the packets satisfy their delay constraints, it should be understood this only holds for packets without transmission errors. Suppose that the error probability of transmitting a packet is P_e. Then the DEDF scheme only guarantees that a packet meets its deadline with probability P_e, assuming polling and EOF messages are transmitted correctly. In view of this, if we are asking higher probabilistic quality of service, we have to resort to either retransmission or feedforward error correction (FEC). The latter adds a fixed overhead to each packet and thus reduces the bandwidth efficiency. The former might lead to tighter deadlines as one has to reserve time for retransmission. This also means a tighter admission criterion. Trade-offs between retransmission and FEC is certainly worth further study.

(ii) Time varying capacity:
Suppose that an adaptive rate FEC is used in (i). Then the capacity of the channel is time varying. In this case, the number of VBR sources that can be admitted should be reduced so that QoS can still be maintained when the capacity is time varying.

(iii) Random access:
Suppose the objective is to provide a probabilistic type of QoS as discussed in (i). Instead of using polling, one may consider a contention based medium access protocol to inform the base station that a VBR source has a packet to send. However, as a contention based protocol usually yields a random access delay, a long access delay might result in many overdue packets. The control of QoS is much more difficult than the polling scheme presented in this paper.

(iv) Propagation delay:
In this paper, we assume negligible propagation delay so that polling can be used for information exchange. In the case of long propagation delay, acquiring the information where a VBR source has a packet to send is time-consuming. A standard approach is to time-interleave the scheme into different phases. Within each phase, the scheme is then operated at a lower transmission speed so that the propagation delay is negligible comparing to the packet transmission time. However, this might lose some

multiplexing gain due to time-interleaving. It is not clear whether the effective bandwidth approach (see e.g., [6] and references therein) would be more efficient under such an environment.

Acknowledgments

This research is supported in part by the National Science Council, Taiwan, R.O.C., under Contract NSC-89-2213-E007-127, the Institute for Information Industry (III), Taiwan, R.O.C. under Contract "Real-time video transmissions in wireless networks," and the program for promoting academic excellence of universities 89-E-FA04-1-4.

Appendix

I. Appendix A

In this section, we prove Theorem 3. Consider a fictitious arrival process $\tilde{A}$ that is constructed by the arrival process A as follows:
(i) Every arrival of A in the busy state is also an arrival of $\tilde{A}$ with the same arrival time.
(ii) Every arrival of A in the idle state is also an arrival of $\tilde{A}$ with the arrival time delayed to the beginning of the next busy state.
As the length of the idle state is bounded above by p (Steps 2 and 3 of DTR B), the delay between an arrival in $\tilde{A}$ and its corresponding arrival in A is at most p. Thus, $\tilde{A}$ can be viewed as the output of an O_p-server with the input A.

For the fictitious arrival process $\tilde{A}$, the base station is always in the busy state when there are arrivals. Thus, for $\tilde{A}$, DTR B is equivalent to having the carrier sensing capability, and DTR B is equivalent to DTR A. As DTR A is the maximal f-regulator (Theorem 2), DTR B can be viewed as the concatenation of an O_p-server and the maximal f-regulator for A.

II. Appendix B

In the section, we prove Theorem 5. First, one observes that $T_{i,B}(\tau^b_{i,k}) = B_i(\tau^b_{i,k})$ for all k as there is no backlogged tokens for the i^{th} VBR source at $\tau^b_{i,k}$. Thus, in view of (13) and (14), $T_{i,B}$ is the output from DTR B (with $p = p_{i,1}$) for the input A_i. It then follows from Theorem 3 that

$$T_{i,B} \geq A_i \star O_{p_{i,1}} \star f_i. \tag{20}$$

Since we assume that all the *effective* tokens are served before their deadlines, one has $B_i(t) \geq T_{i,B}(t - p_{i,2})$ for all t, where $p_{i,2}$ is the parameter added in the deadline of an effective token in Step 4 of the DEDF scheme. Thus,

$$B_i \geq T_{i,B} \star O_{p_{i,2}}. \tag{21}$$

In conjunction with (20),

$$B_i \geq A_i \star (O_{p_{i,1}} \star f_i) \star O_{p_{i,2}} = A_i \star (O_{p_{i,1}+p_{i,2}} \star f_i) = A_i \star \tilde{f}_i.$$

This shows that the DEDF scheme is an $\tilde{f}_i$-server for the i^{th} VBR source. Also, as $T_{i,B}$ is the output from DTR B, $T_{i,B}$ is f_i-upper enveloped from Theorem 3. Thus, it follows from (21) and (3) that B_i is $\hat{f}_i$-upper enveloped with $\hat{f}_i(t) = f_i(t + p_{i,2})$.

As the horizontal distance between $\tilde{f}_i$ and f_i is $p_{i,1}+p_{i,2}$, it then follows from the filtering theory that the maximum delay for the i^{th} VBR source is bounded above by $p_{i,1}+p_{i,2}$ if A_i is f_i-upper enveloped.

III. Appendix C

In this section, we prove Theorem 6. As in [11] and [17], we prove this by contradiction. Suppose that the time D is the first missed deadline, i.e., a token with deadline D is not served before D. Let s denote the last time before D at which a token with deadline larger than D is initially served. If there is no such time, let s denote the last time before D when the base station becomes busy. Let C denote the set of tokens which are with deadlines less than or equal to D and which are served after s. Let W be the total time to serve the tokens in C.

First, we derive a lower bound for W. Since all the tokens served in the interval $[s + t_B, D]$ have no deadlines larger than D by the definition of s, they all belong to the set C. Since the server is busy in this interval and by the assumption that time D is the first missed deadline, we have

$$W > D - s - t_B. \tag{22}$$

Next, we derive an upper bound for W. Let X be the set of VBR users that have tokens in the set C. There are two facts for the VBR users in X. The first fact is that all the tokens of the VBR users in X which are served before s have deadlines less than or equal to D. Otherwise, there exists a token that is served after s and it has deadline larger than D. This contradicts the definition of C and X.

The second fact is that all the token queues of the VBR users in X have no backlog at time s. Otherwise, the event that a token with deadline larger than D is served at s would not occur as a token with an earlier deadline would be chosen first. Thus, one has for $i \in X$ that

$$T_{i,B}(s) = B_i(s), \tag{23}$$

and that

$$T_{i,I}(s) = I_i(s), \tag{24}$$

where $I_i(t)$ denote the number of ineffective tokens served by time t.

Let $N_{i,B}(t)$ denote the number of effective tokens that are with deadlines less than or equal to t. Since all the effective tokens are assigned deadlines as the sum of their generation time and $p_{i,2}$, we have

$$N_{i,B}(t) = T_{i,B}(t - p_{i,2}). \tag{25}$$

The total time W to serve tokens in the set C is the sum of the time to serve effective tokens which is denoted by W_B and the time to serve ineffective tokens which is denoted by W_I. From the first fact that all the tokens of the VBR users in X served before s have deadlines less than or equal to D, we have

$$\begin{aligned} W_B &= \sum_{i:i\in X} \big(N_{i,B}(D) - B_i(s)\big)t_B \\ &= \sum_{i:i\in X} \big(T_{i,B}(D-p_{i,2}) - B_i(s)\big)t_B \quad \text{(from (25))} \\ &= \sum_{i:i\in X} \big(T_{i,B}(D-p_{i,2}) - T_{i,B}(s)\big)t_B. \quad \text{(from (23))} \end{aligned}$$

As argued in the proof of Theorem 5, $T_{i,B}$ is the output from DTR B and thus f_i-upper enveloped (Theorem 3). This leads to the following upper bound

$$W_B \le \sum_{i:i\in X} f_i(D-p_{i,2}-s)t_B. \tag{26}$$

Let $N_{i,I}(t)$ denote the number of ineffective tokens with deadlines less than or equal to t. One has from the first fact described above that $W_I = \sum_{i:i\in X}\big(N_{i,I}(D) - I_i(s)\big)t_I$. Thus, we have $W_I = \sum_{i:i\in X}\big(N_{i,I}(D) - T_{i,I}(s)\big)t_I$ from (24). Since the ineffective tokens with deadlines less than or equal to t are generated by time $D - p_{i,2}$, considering the worst case when all the ineffective tokens are generated periodically with period $p_{i,1}$ yields

$$W_I \le \sum_{i:i\in X} \lceil \frac{1}{p_{i,1}}(D-p_{i,2}-s)\rceil t_I. \tag{27}$$

From (26) and (27), we have

$$\begin{aligned} W &= W_B + W_I \\ &\le \sum_{i:i\in X} f_i(D-p_{i,2}-s)t_B \\ &\qquad + \sum_{i:i\in X} \lceil \frac{1}{p_{i,1}}(D-p_{i,2}-s)\rceil t_I. \end{aligned} \tag{28}$$

Combining (28) with (22), we reach a contradiction with (15) (by substituting $D - s - t_B$ in (15)). Thus, all the tokens are served before their deadlines.

References

[1] R. Agrawal and R. Rajan, "Performance bounds for guaranteed and adaptive services," *IBM RC 20649*, 1996. Also in *Proc. 34th Allerton Conf. on Comm., Cont. & Comp.*, Monticello, IL, Oct. 1996.

[2] J.Y. Le Boudec, "Application of network calculus to guarantee service networks," *IEEE Transactions on Information Theory*, Vol. 44, pp. 1087-1096, 1998.

[3] 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.

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

[5] C.S. Chang, "On deterministic traffic regulation and service guarantees: a systematic approach by filtering," *IEEE Transactions on Information Theory*, Vol. 44, pp. 1097-1110, 1998.

[6] C.S. Chang. *Performance Guarantees in Communication Networks*. Springer-Verlag, 2000

[7] C.S. Chang, K.C. Chen, M.Y. You, and J.F. Chang, "Guaranteed quality-of-service wireless access to ATM networks," *IEEE Journal on Selected Areas of Communications*, Vol. 15, pp. 106-118, 1997.

[8] K.C. Chen, "Medium access control of wireless local area networks for mobile computing," *IEEE Network*, Sep./Oct., pp. 50-63, 1994.

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

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

[11] R.L. Cruz, "Quality of service guarantees in virtual circuit switched networks," *IEEE Journal on Selected Areas of Communications*, Vol. 13, pp. 1048-1056, 1995.

[12] R.L. Cruz and C.M. Okino, "Service guarantees for window flow control," *Proc. 34th Allerton Conf. on Comm., Cont. & Comp.*, Monticello, IL, Oct. 1996.

[13] N.R. Figueria and J. Pasquale, "An upper bound on delay for virtualclock service displine," *IEEE/ACM Transactions on Networking*, Vol. 3, pp. 399-408, 1994.

[14] L. Georgiadis, R. Guerin, and V. Peris, "Efficient network QoS provisioning based on per node traffic shaping," *IEEE/ACM Transactions on Networking*, Vol. 4, pp. 482-501, 1996.

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

[16] M.J. Karol, Z. Liu, and K.Y. Eng, "Distributed-queueing request update multiple access (DQRUMA) for wireless packet (ATM) networks," *IEEE Int. Conf. Commun.*, Seattle, U.S.A, 1995.

[17] C.L Liu and J.W. Layland, "Scheduling algorithms for multiprogramming in a hard-real-time environment," *Journal of the Association for computing Machinery*, Vol. 20, pp. 46-61, 1973

[18] 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 Transactions on Networking*, Vol. 1, pp. 344-357, 1993.

[19] R. Prasad, M. Alam and J.R. Farsertou, "Wireless Next Generation: 4G?" Multiaccess, Mobility and Teletraffic in Wireless Communications: Vol. 5, edited by G. Stüber and B. Jabbari, Klumer Academic Publisher, 2000, pp. 305-316

[20] X. Qiu, K. Chawla, L.F. Chang, J. Chuang, N. Sollenberger and J. Whitehead, "RLC/MAC Design Alternatives for Supporting Integrated Services over EGPRS," *IEEE Personal Communication*, Vol. 7, No. 2, April 2000, pp.20-33

[21] H. Sariowan, "A service-curve approach to performance guarantees in integrated-service networks," Ph.D. Dissertation, University of California, San Diego, 1996.

[22] G.G. Xie and S.S. Lam, "Delay guarantee of virtual clock server," *IEEE/ACM Transactions on Networking*, Vol. 3, pp. 683-689, 1995.

[23] H. Zhang, "Service displines for guaranteed performance service in packet-switching networks," *Proceedings of the IEEE*, Vol. 83, pp. 1374-1396, 1995.

[24] L. Zhang, "Virtualclock: a new traffic control algorithm for packet switching networks," *ACM Transactions on Computer Systems*, Vol. 9, pp. 101-124, 1991.

[25] D. Zhao, X. Shen and J.W. Mark, "Efficient Call Admission Control for Heterogeneoes Services in Wireless Mobile ATM Networks," IEEE Communication Magazine, Oct 2000, pp. 72-78

Zhi-Ren Chang received the B.S. and M.S. degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, R.O.C. in 1995 and 1997, respectively. From 1997 to 1999, he served as a lieutenant in the army. He joined the Winbond Electronics Corp. in 1999 and has been engaged in the design of VLSI circuits. He participated in the projects of ISDN and Bluetooth. His research interests are communication systems and networking. His email address is cjchang2@winbond.com.tw.

I-Chung Lee received the B.S. degree in Control Engineering from National Chiao-Tung University in 1987, and the M.S. degree in Electrical Engineering from National Taiwan University, Taiwan, ROC, in 1989. During 1989-1991, he became a commander of communication platoon while performing his obligatory service in the army. During 1991-1992, he worked for Vidar-SMS, Inc., developing switching power supply. Since 1992, Mr. Lee has been as Lecturer of Automation Engineering at Ta-Hwa Institute of Technology. Currently, he is a Ph.D. candidate in Dept. of Electrical Engineering, National Tsing-Hwa University. His research interests include performance evaluation and guarantees of communication networks. His email address is iclee@gibbs.ee.nthu.edu.tw.

Cheng-Shang Chang (S'85-M'86-M'89-SM'93) received the B.S. degree from the National Taiwan University, Taipei, Taiwan, in 1983, and the M.S. and Ph.D. degrees from Columbia University, New York, NY, in 1986 and 1989, respectively, all in Electrical Engineering. From 1989 to 1993 he was employed as a Research Staff Member at the IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y. Since 1993, he has been with the Department of Electrical Engineering at National Tsing Hua University, Taiwan, R.O.C., where he is a Professor. His current research interests are concerned with queueing theory, stochastic scheduling and performance evaluation of telecommunication networks and parallel processing systems. Dr. Chang received the IBM Outstanding Innovation Award in 1992, and the Outstanding Research Award from the National Science Council, Taiwan, in 1999. He is the author of the book "Performance Guarantees in Communication Networks," and he served as an editor for *Operations Research* from 1992 to 1999. His email address is cschang@ee.nthu.edu.tw.

Chien-Hsin Li received the B.S. degree in Electronic Engineering Technology from National Taiwan Institute of Technology, Taiwan in 1983, and the M.S. degree in Electrical Engineering from National Taiwan University, Taiwan, ROC, in 1985. He was a part-time student and Ph.D. candidate in Dept. of Information Science of National Tsing-Hua University from 1989 to 1994. During 1985-1989, he worked for III (Institute for Information Industry) as an engineer and project leader, developing tools and utilities for software development environment. During 1989-1995, he worked part-time employee for III, in charge the plan and consultant for project planning. During 1995-2000, he worked for III, developing firmware and software for wireless communication. Since May 2000, he moved work to Primax Electronics Ltd. working for the product development of Bluetooth wireless technologies. His research interests include software engineering, communication networks, and protocol of wireless communications. His email address is james.lee@primax.com.tw.

Ben-Li Sui, photograph and biography not available at the time of publication.