Determining the Call Admission Region for Real-Time Heterogeneous Applications in Wireless TDMA Networks

Jeffrey M. Capone, Arizona State University and Ioannis Stavrakakis, Northeastern University

Abstract

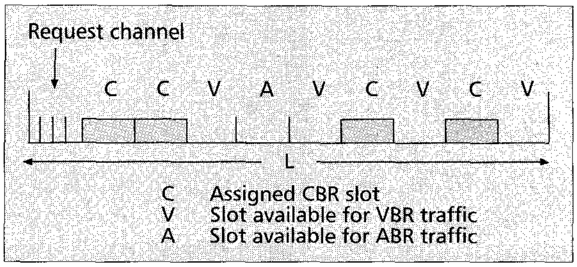
The call admission problem for a wireless packet-switched network supporting *homogeneous* applications — such as a cellular voice network — is an old one and has been extensively investigated in the past. The focus of the present article is to investigate the call admission region for a TDMA (wireless) system supporting *heterogeneous* real-time VBR applications with distinct QoS requirements and traffic characteristics. The QoS is defined in terms of a maximum tolerable packet delay and dropping probability; packets may be dropped due to delay violations or channel-induced errors. The call acceptance region is investigated in this article under the assumption that each user's per-frame resource (slot) requests are communicated to the scheduler (resource allocation authority). The call acceptance region is shaped by the QoS that can be delivered by the uplink scheduling policy. In the beginning of this article, some mechanisms employed to inform the scheduler of the users' requests along with some scheduling policies are discussed. While these scheduling policies identify QoS points in the call admission region, they cannot reveal the entire region. In this article, an approach is outlined on how to precisely determine the call admission region (largest set of QoS points delivered under any scheduling policy) for a TDMA system. This approach is shown to lead to the precise description of the reduced call admission region in the presence of channel errors. In order to demonstrate the performance improvement provided by an error control scheme, a simple QoS-sensitive forward error control protocol operating over an erroneous channel is employed, leading to an enlargement of the calculated call admission region. Finally, it is shown that feasible scheduling policies exist which do deliver at least the minimum QoS requirement of the applications whose associated QoS vector falls within the determined call admission region.

In a wireless network many users communicate over a shared unreliable channel. Several network architectures have been designed to facilitate this communication, but typically only cater to one traffic type. These architectures include a medium access control (MAC) protocol to coordinate the sharing of the channel and possibly an error control scheme to combat the unreliability of the wireless link. Two popular time sharing approaches are carrier sense multiple access (CSMA) and time-division multiple access (TDMA). In CSMA, the user senses the channel and, if it is idle, transmits a packet (information unit) without any coordination with the other users. Collisions are possible, in which case retransmissions are scheduled after a random amount of time. The service delivered to the supported applications may be classified as "best effort." In TDMA time is divided into periodic frames, and each frame contains a number of time slots. Each time slot is the time required for the transmission of a packet (plus some guard time). A base station (or central access point) coordinates usage of the time slots, which allows for service diversification. The downlink (base to terminal) traffic is broadcast from the base station, and therefore can be centrally controlled at the base station. The transmissions on the uplink (terminal to base) are distributed among the geographically dispersed users, and coordination typically takes place in two phases, call establishment and channel access.

At call establishment, users typically request access through a control channel with request packets. Since there is no coordination among these users, there is a possibility of collision between two or more request packets. To

Research for this article was supported in part by the National Science Foundation under Grant NCR 9628116.

0890-8044/98/$10.00 © 1998 IEEE


Request channel
C C V A V C V C V
L
C Assigned CBR slot
V Slot available for VBR traffic
A Slot available for ABR traffic

■ Figure 1. *A typical uplink TDMA frame structure supporting CBR, VBR, and ABR traffic classes.*

resolve such collisions, a contention resolution protocol is employed in the control channel. During call establishment, an amount of bandwidth (measured in time slots per frame) is requested for servicing the call. If the call is requesting constant bit rate (CBR) service (or circuit-switched), the user is allocated (scheduled) a fixed number of slots per frame for the duration of the call. If the request cannot be accommodated, the call is blocked from service. If the call is requesting variable bit rate (VBR) service, where the bandwidth needed to service the call may vary over time, the user must dynamically update its bandwidth requirements.

Due to the lack of resources in a wireless network, the continuous provision of information on the status of the requirements for the distributed sources may not be possible. The information needed to assign transmissions may only be updated periodically or at discrete moments, as in [1–5]. These systems use a combination of control channels, piggybacking on information-bearing packets, and polling procedures for the communication of allocation requests and transmission assignments between the base station and source. Once the request is received, a transmission scheduling policy at the base station must be employed to decide on the access rights to the channel. Therefore, channel access for VBR applications is achieved by engaging two mechanisms: the signaling mechanism, which conveys the bandwidth requests to the base station, and a transmission scheduling mechanism which decides the amount of resources allocated to each application.

In the following section, channel access techniques that are designed to support integrated services such as CBR and VBR are discussed, with emphasis on the support of VBR service. Access and scheduling for wireless voice networks employing speech activity detection is then discussed, providing insight into the development of the region of achievable quality of service (QoS). We next examine the impact of the wireless channel and a QoS-sensitive error control scheme on the region of achievable QoS. It is then shown that feasible scheduling policies exist which do deliver at least the minimum QoS requirement of the applications whose associated QoS vector falls within the determined call admission region.

Channel Access for Integrated Services

Channel access techniques for integrated services in a TDMA environment have been proposed in numerous publications (e.g., [1–10]). The common objective of these schemes is to enable sharing of resources among diverse applications while delivering the required QoS. Typically, the services to be supported are designed to be compatible with the prevalent network architecture for integrated services over fiber/copper-based channels, ATM. Some of the services offered by ATM to support the various traffic classes are CBR, VBR, and available bit rate (ABR). The focus of this article is on support of VBR service in a wireless network.

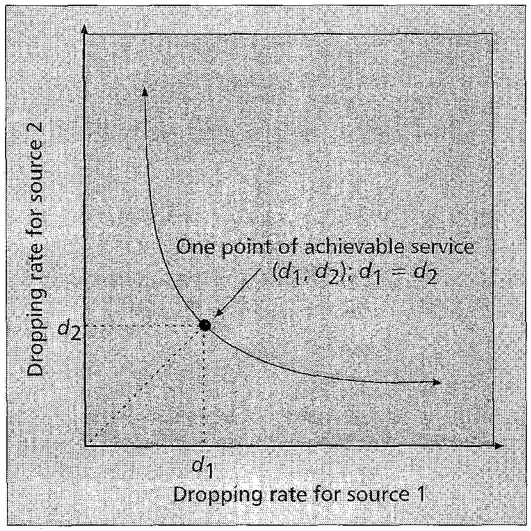
In [1], access and scheduling are based on a multiservice dynamic reservation (MDR) TDMA frame format. At the beginning of each uplink MDR TDMA frame, a slotted ALOHA control channel is employed for accommodating allocation requests, followed by a number of slots available to service the VBR traffic. A typical uplink TDMA frame where slots are allocated to various classes of service (e.g., CBR, VBR, and ABR), and service requests are processed at frame boundaries, is illustrated in Fig. 1.

In [1], the VBR traffic sources are assigned slots based on a usage parameter control (or scheduler), that achieves statistical multiplexing. The distributed-queuing request update multiple access (DQRUMA) protocol is proposed in [3]. In that work, the channel access is performed on a slot-by-slot basis by providing a request channel at the beginning of each uplink slot. The source may request access at the beginning of each uplink slot and listen for permission on the downlink for the next uplink slot. In addition, a contention-free request channel is formed by "piggy-backing" the allocation requests on the uplink information-bearing packets. The emphasis of this work is on the signaling of requests, not the scheduling of transmissions. The DQRUMA scheme is similar to the Dynamic Slot Assignment (DSA) protocol proposed in [2, 6]. In the DSA protocol, each uplink burst contains capacity requirements of the source that are used by the scheduler at the base station which performs a slot-by-slot allocation. In addition, a cyclical random access channel is provided to convey the requests from sources that have been previously inactive or have an urgent packet. The authors propose to use an earliest due date (EDD)-type policy for scheduling these transmissions. Another protocol, Service Integration for Radio Access (SIR^{++}), proposed in [7], is designed to service a generic mixture of traffic by allocating a fixed amount of slots per frame to each source, then allowing competition for the unused slots based on a random access approach. The authors attempt to tailor the performance delivered to the sources by managing the amount of fixed bandwidth and best-effort bandwidth allocated to each source. In all the above channel access schemes, signaling occurs at discrete and typically periodic instances — explicit and continuous exchange of control information needed for scheduling is not possible due to the limited resources (bandwidth).

The performance of the scheduling policies in [6, 7] are evaluated for different combinations of heterogeneous VBR sources. In [5, 8, 9], the scheduling policies are designed to meet (if achievable) the diverse QoS required by the set of VBR applications. In addition, a call admission policy is developed. In [5], the authors propose a polling strategy to accommodate the VBR traffic. A polling generation period is determined for each VBR source so that each packet meets its delay constraint. The authors develop a very conservative scheme and provide only for sufficient conditions to ensure that the scheduling policy can deliver service satisfying the delay constraints of the VBR sources. The authors claim that this may be fine in an environment where resources are abundant, such as ATM; however, in a wireless environment where resources are scarce, the efficiency of the scheme should be improved by considering probabilistic QoS guarantees. Probabilistic QoS guarantees are considered in [8, 9] and in this article. For example, packets from a particular source (such as voice or video) may only need to meet their delay constraint 99 percent of the time, and can tolerate being dropped otherwise. In this article, necessary and sufficient conditions in order for the probabilistic QoS to be achieved are presented, leading to bandwidth-efficient call admission control and scheduling algorithms.

TDMA Voice Networks

Before a scheduling policy can be designed, it should be known whether or not the required performance can be delivered to the set of VBR sources through some scheduling policy. In the following, this question will be addressed for a set of statistically identical VBR sources — a set of voice sources employing speech activity detection.



■ Figure 2. *Achievable service for a homogeneous system under a nondiscriminating policy.*

Many channel access protocols have been developed that exploit the pauses [11] in the traffic from voices sources, for example, PRMA++ and E-TDMA [12–14]. In these protocols, one time slot per TDMA frame is allocated to each active source. When an inactive source becomes active, the source notifies the base station of its activity by transmitting a request packet through the control channel. There is the possibility of packet collision in the control channel, however, typically the amount of resources appropriated to the control channel is enough so that the probability of an allocation request being successful is close to unity [13]. Once the base station receives this notification, the scheduler assigns to the source a slot (if available) in the TDMA frame. If a slot is not available in the current frame, the packet is not serviced and is dropped at the source as a result of excess delay [13, 14]. Therefore, the packet dropping probability is influenced by the amount of resources and the number of sources competing.

Consider a system of N voice sources each with activity factor γ accessing T slots in the uplink of a TDMA frame. T is the number of slots in the TDMA frame available to service the VBR applications (Fig. 1). In this system, the total number of requests that arrive in frame n, $\lambda_S(n)$, can be modeled as a binomial random variable [11] with mean $N\gamma$. As previously mentioned, if the voice packet is not serviced during the frame following its arrival, it is dropped. The system packet dropping probability,

$$\frac{b_S}{E[\lambda_S(n)]}=\frac{b_S}{N\gamma},$$

is then the ratio of the expected number of packets dropped per frame to the expected number of packets that arrive per frame [13, 14]. The expected number of packets dropped per frame (or system dropping rate) is

$$b_S = \{E[\lambda_S(n) \mid \lambda_S(n) > T] - T\}P(\lambda_S(n) > T). \qquad (1)$$

The formulation of Eq. 1 allows for the development of a call admission rule. Simply limit the maximum number of voice sources, N, so that the packet dropping probability

$$\frac{b_S}{N\gamma} \le 0.02,$$

where 2 percent is considered to be the maximum tolerable packet dropping probability. In this system, an explicit scheduling policy is not needed. The implicit nondiscriminating scheduling policy would induce the same (system) packet dropping probability to all sources as a result of their homogeneous traffic characteristics.

The above policy will not be able to deliver the QoS requirements to the supported applications if they have diverse traffic characteristics and/or QoS requirements. When heterogeneous sources with diverse QoS requirements share the same resources, an explicit scheduling policy is needed to prevent heavily active applications (sources) from capturing more resources and receiving better than the required service at the expense of compromising service to the other applications. In addition, in such a diverse environment it is not sufficient to only check that the *system* dropping rate (or dropping probability) is satisfied.

The QoS requirements of each application, i, can be described in terms of a maximum tolerable dropping probability p_i. The corresponding maximum tolerable packet dropping rate, d_i (measured in expected number of dropped packets per frame), is easily determined by $d_i = p_i\lambda_i$, $1 \le i \le N$. The QoS vector associated with the supported applications can be defined in terms of the (performance) packet dropping rate vector $\mathbf{d} = (d_1, d_2, \ldots, d_N)$.

Consider the following example where two heterogeneous sources, which may require multiple slots per frame, compete for $T = 5$ slots. Sources 1 and 2 have an average packet arrival rate of 3.0 and 1.6 packets/frame and a variance of 1.8 and 3.84 packets/frame, respectively. In this example, the QoS vector $\mathbf{d}$ is assumed to be $\mathbf{d} = (d_1, d_2) = (0.781, 0.099)$, and therefore the associated maximum tolerable system packet dropping rate is equal to $d_S = d_1 + d_2 = 0.790$. It turns out that the maximum tolerable system packet dropping rate is satisfied by the system (i.e., $b_S \le d_S = 0.790$). However, even if source 1 is removed from the system, it can be shown that the dropping rate delivered to source 2 is 0.10; thus, the QoS vector is not achievable under any policy.

The main question investigated in this article is whether or not a scheduling policy exists that can deliver a given QoS vector $\mathbf{d}$. To answer this question, the region of achievable QoS vectors is established. It is based on a set of inequalities and an equality constraint derived by employing work-conserving (nonidling) arguments. Details regarding these derivations can be found in [8–10].

Achievable QoS and Call Admission Region in an Error-Free Environment

The region of achievable QoS is the set of points (performance vectors) that can be delivered under some policy. The determination of the region of achievable QoS leads to the development of a call admission rule. For example, if, with the addition of the new source, the new multidimensional target QoS vector is in the region of achievable QoS vectors, the call can be admitted since there exists some policy that can deliver the target service to each application. If the call cannot be admitted and more resources can be made available, a precisely defined region of achievable QoS can also be used to determine the minimum additional resources required in order for the new call to be admitted.

Considering the homogeneous case discussed in the previous section, the performance delivered to the applications under a nondiscriminating policy is known, and it is seen in Fig. 2 for a system of two sources where equal dropping rates are delivered to each source. This policy identifies one point in the region of achievable QoS. If service diversification is desired, it is important to establish the level of diversification possible; that is, the set of points (d_1, d_2) that can be delivered by any (discriminating and nondiscriminating) policy. Referring to Fig. 2, if it is desired to improve the performance delivered to source 1, what effect does this policy have on the QoS delivered to source 2, what are the bounds on the service delivered to each source under some policy, and how precise are these bounds?

In the following, the region of achievable QoS for a set of arbitrary VBR sources competing for an error-free channel is precisely described. In an error-free channel all transmitted packets are successfully received. In this case, the delivered QoS is shaped practically entirely by the packet discarding process at the transmitter (source) due to delay violations; the latter occur when the demand exceeds the amount of available resources for a sufficiently long period. Thus, the performance is limited by the amount of available resources (resource-limited). However, the achievable QoS in a wireless network is shaped not only by the amount of available resources and the employed resource management scheme, but also by the channel quality (interference). In addition, lower layers may substantially influence the design of higher-layer protocols. Therefore, the achievable QoS is shaped collectively by the capabilities of the physical channel, error control, and the MAC protocols. These issues are investigated in the next section.

Throughout this article, channel access is based on the general uplink TDMA frame format shown in Fig. 1, where T slots are available to service the N heterogeneous real-time VBR sources (applications). Since allocation requests are processed at the beginning of each frame and packets are not synchronously generated at frame boundaries, the sources are required to buffer the packets at least until the beginning of the following frame. As a result, the source packet arrival process can be described in terms of a general arrival process embedded at frame boundaries. No additional assumptions for the packet arrival process are necessary at this point.

Packets which cannot be transmitted over the frame (service cycle) following their arrival are considered to have excess delay and are dropped at the source (transmitter); that is, all applications have a common packet delay tolerance, but may have diverse packet dropping rates. Diverse maximum packet delays in addition to dropping rates are also considered. In both sections, the region of achievable QoS, and therefore the call admission region, is precisely described.

Common Delay Tolerance

At the beginning of each frame n each source i will request a random number of slots denoted by $\lambda_i(n)$. If the aggregate demand in frame n, $\Sigma_{i=1}^{N}\lambda_i(n)$, exceeds the number of slots available to the VBR traffic T — referred to as an *overloaded frame* — decisions must be made regarding the amount of service that will be provided to each source. The portion of slots under policy f allocated to source i, $a_i^f(n)$, may be less than required by that source, $\lambda_i(n)$, due to resource limitations. Packets from a source which do not receive service over the frame following their arrival are considered to have excess delay and are dropped at the source. The number of packets from source i dropped in frame n will depend on the scheduling policy[1] f and is given by

$$d_i^f(n) = \lambda_i(n) - a_i^f(n)\begin{cases} = 0 & \text{if } \sum_{j=1}^{N}\lambda_i(n) \le T \\ \ge 0 & \text{if } \sum_{j=1}^{N}\lambda_i(n) > T \end{cases} ,1 \le i \le N. \qquad (2)$$

Let $d_i^f = E[d_i^f(n)]$, $a_i^f = E[a_i^f(n)]$, and $\lambda_i = E[\lambda_i(n)]$ be the (assumed time-invariant) expected values of the associated quantities.

The individual dropping rate, d_i, can be controlled by the scheduling policy f. However, regardless of the scheduling policy, the total number of requests that exceed the number of resources T in each frame is conserved. Therefore, under any scheduling policy the expected number of packets dropped in the system is a constant. That is, the policy f will allow for diversification in the QoS delivered to the individual applications, but the system dropping rate remains conserved. This result can also be seen by summing Eq. 2 over all sources $i \in S = \{1, 2, \ldots, N\}$, and by considering the expected value of the associated quantity, which yields

$$d_S^f = E\left[\sum_{i=1}^{N} d_i^f(n)\right] = \left\{E\left[\lambda_S(n) \middle| \lambda_S(n) > T\right] - T\right\} P(\lambda_S(n) > T). \qquad (3)$$

As can be seen from Eq. 3, d_S^f is independent of the policy f; it only depends on the aggregate arrival process, $\lambda_S(n) = \Sigma_{i\in S}\,\lambda_i(n)$, and the number of resources T. Therefore, the system dropping rate d_S^f is conserved under any work-conserving policy f and denoted b_S.

This conservation property constitutes a necessary condition for a QoS vector $\mathbf{d} = (d_1, d_2, \ldots, d_N)$ to be achieved by some policy f, associated with the sum of the components of the QoS vector, $d_S = \Sigma_{i\in S}\, d_i = b_S$. In the special case of a homogeneous system, such as the wireless voice network described earlier, satisfying the system performance is sufficient to guarantee that the target QoS vector is achievable. This result has been established in [8].

For a heterogeneous environment, satisfying the system dropping rate does not ensure that the QoS vector is achievable (as shown earlier). The lower bound on the achievable performance for a subset g of sources under some policy f must also be considered. This lower bound is found by considering the subset of sources in isolation and examining the induced performance.

Let d_g^f denote the average subsystem g packet dropping rate under policy f, defined by

$$d_g^f \overset{\Delta}{=} E\left[\sum\nolimits_{i\in g} d_i^f(n)\right] = \sum\nolimits_{i\in g} E\left[d_i^f(n)\right] = \sum\nolimits_{i\in g} d_i^f, \quad \forall g \subset S.$$

That is, d_g^f is equal to the aggregate packet dropping rate associated with sources in group g only, under policy f; all N sources in S are assumed to be present and served under policy f.

The lower bound on the performance delivered to set g is found by considering a system in which only sources in g are present; sources in set $\{S - g\}$ are considered removed and all of the resources are allotted to sources in g. This lower bound is given by

$$b_g = \{E[\lambda_g(n) \mid \lambda_g(n) > T] - T\}P(\lambda_g(n) > T), \qquad (4)$$

where $\lambda_g(n)$ denotes the aggregate arrival rate from sources in set g, $g \subseteq S$, that is, $\lambda_g(n) = \Sigma_{i\in g}\,\lambda_i(n)$.

It is apparent that no policy can deliver a lower dropping rate than b_g to sources in set g when all sources in S are present; therefore, the following conditions,

$$d_g \ge b_g \quad \forall g \subseteq S, \qquad (5)$$

$$d_S = b_S, \qquad (6)$$

are necessary in order for a QoS vector $\mathbf{d}$ to be achieved by some scheduling policy f. It can also be shown that the condi-

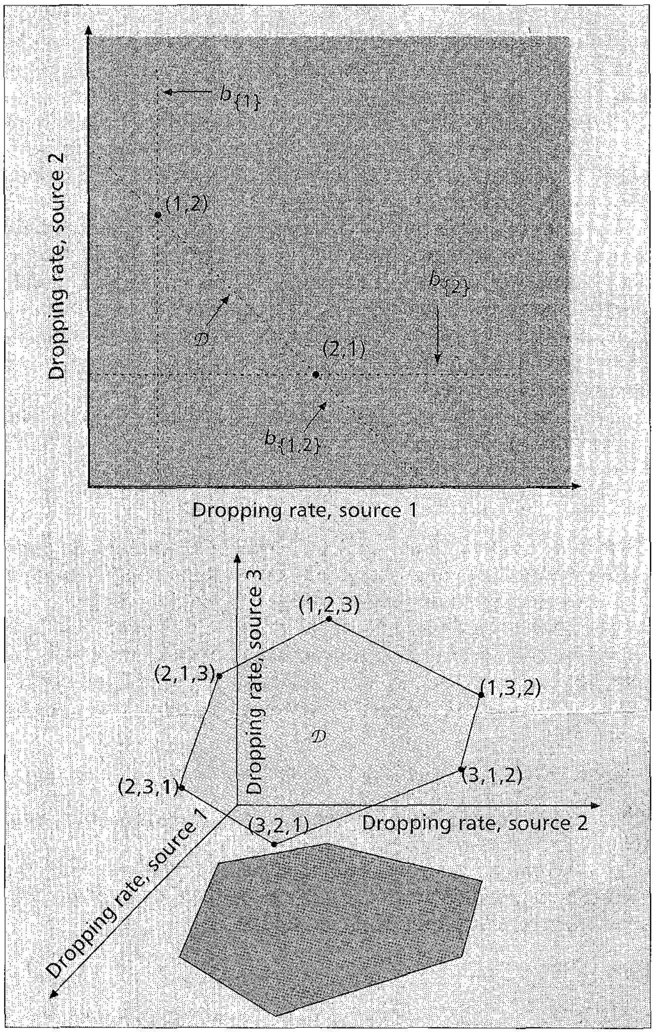
[1] *Throughout this article all scheduling polices are work-conserving, that is, no slots are left idle when there are packets present to transmit.*

tions in Eqs. 5 and 6 are sufficient [8]. That is, the performance given by any vector in the region described by Eqs. 5 and 6 is delivered by some work-conserving policy f.

Let $\mathcal{D}$ denote the collection of all vectors **d** satisfying Eqs. 5 and 6. Then by definition $\mathcal{D}$ is a convex polytope [15]. Using results from convex polytopes [15], any vector in set $\mathcal{D}$ can be expressed as a convex combination of extreme points (vertices) of $\mathcal{D}$; that is, $\mathcal{D}$ may be expressed as the convex hull of its extreme points, $\mathcal{D} = \text{conv}[\exp(\mathcal{D})]$.

In addition, it can be shown [8] that $\mathbf{d}^*$ is a vertex of set $\mathcal{D}$ iff $\mathbf{d}^*$ is a dropping rate vector resulting from an Ordered Head of Line (O-HoL) priority service policy $\pi = (\pi_1, \pi_2, ..., \pi_N)$; $\pi_i \in \{1, 2, ..., N\}$, $\pi_i \neq \pi_j$, $1 \leq i, j \leq N$. The index of π_i indicates the order of the priority given to the π_i source. None of the π_j sources, $j > i$, may be served as long as packets from sources π_k, $k \leq i$, are present.

Figure 3 provides a graphical illustration of the region $\mathcal{D}$ for the case of $N = 2$ and $N = 3$ sources. The extreme points correspond to QoS vectors **d** induced by the $N!$ O-HoL priority policies $\pi = (\pi_1, \pi_2, ..., \pi_N)$. Referring to the upper part of Fig. 3, it may be observed that the policy $(\pi_1, \pi_2) = (1, 2)$ corresponds to the intersection of the line for the lower bound on the packet dropping rate for source 1, $b_{\{1\}}$, with the system dropping rate, $b_{\{1,2\}}$. Similarly, the second extreme point induced by the policy, $(\pi_1, \pi_2) = (2, 1)$, is the intersection of the lower bound on the packet dropping rate for source 2, $b_{\{2\}}$, with the system dropping rate, $b_{\{1,2\}}$. Similar observations can be made for region $\mathcal{D}$ for a system of $N = 3$ sources, shown in the lower part of Fig. 3.



■ Figure 3. *The region (polytope) $\mathcal{D}$ for a system with two and three sources.*

As can be seen in these figures, the aggregate dropping for the system is conserved. In addition, a performance bound on each subset of sources can be identified. These bounds are established from the set of inequalities in Eq. 5. From the region of achievable QoS described above, the call admission region can be precisely described.

The call admission (CA) region is the set of QoS vectors for which a policy that delivers the target or "better" QoS vector can be found, and is used to control admission into the network. Admission is based on the current state of the network, the new source characteristics, and QoS requirements. For example, if, with the addition of the new source, the new multidimensional target QoS vector is in the CA region, the call can be admitted. If the new multidimensional target QoS vector is not in the CA region and more resources cannot be made available, the call is blocked.

The call admission region, $\mathcal{A}$, associated with the region of achievable QoS vectors $\mathcal{D}$ is established by relaxing the equality condition on the system performance. It is defined to be the region of vectors **d** satisfying

$$\begin{aligned} &d_g \geq b_g \; \forall g \subset S, \\ &d_S \geq b_{\{S\}}. \end{aligned} \tag{7}$$

It is shown in [8] that if $\mathbf{d} \in \mathcal{A}$ then there exists a vector $\mathbf{d}' \in \mathcal{D}$ which is such that $d_i' \leq d_i$ for all $\forall i \in S$. This result implies that if the target QoS vector **d** is in $\mathcal{A}$, there exists a policy which can deliver **d** or "better" (i.e., less than the dropping rate required by any source); the CA region for systems with two and three sources is depicted in Fig. 4.

Diverse Delay Tolerance

In this section a nontrivial extension to the work presented earlier is developed by allowing for further diversification in the QoS. By considering diverse maximum tolerable delays in addition to dropping probabilities, some new and interesting problems emerge. The region of achievable QoS vectors is established for policies that are work-conserving and satisfy the earliest due date (EDD) service criterion (WC-EDD); such policies are known to optimize the overall system performance [16]. Under the WC-EDD family of policies, denoted $\mathcal{F}$, the conservation property is maintained which makes determining the region of achievable QoS possible. The key to understanding the conservation property for this system is to examine the residual traffic process.

In this system, depending on the QoS requirements, packets that cannot be transmitted over the frame following their arrival may be dropped (due to delay violation) or delayed to compete for service in the next frame. As seen in Fig. 1, a TDMA system in which arrivals are considered at frame boundaries may be modeled in terms of a discrete time system in which packet delays are measured in frames (1 frame = L time units). In this system, the N VBR sources are partitioned into two classes, $S_1 = \{1, 2, ..., K\}$ and $S_2 = \{K + 1, K + 2, ..., N\}$. Packets generated from sources in S_1 have a common maximum delay tolerance of L time units (1 frame) and packets generated from sources in S_2 have a common maximum delay tolerance of $2L$ time units (2 frames).

New arrivals from sources in set S_2 that are not serviced in the present frame form the residual traffic in the following frame. Consider the residual traffic as illustrated in the realization depicted in Fig. 5.

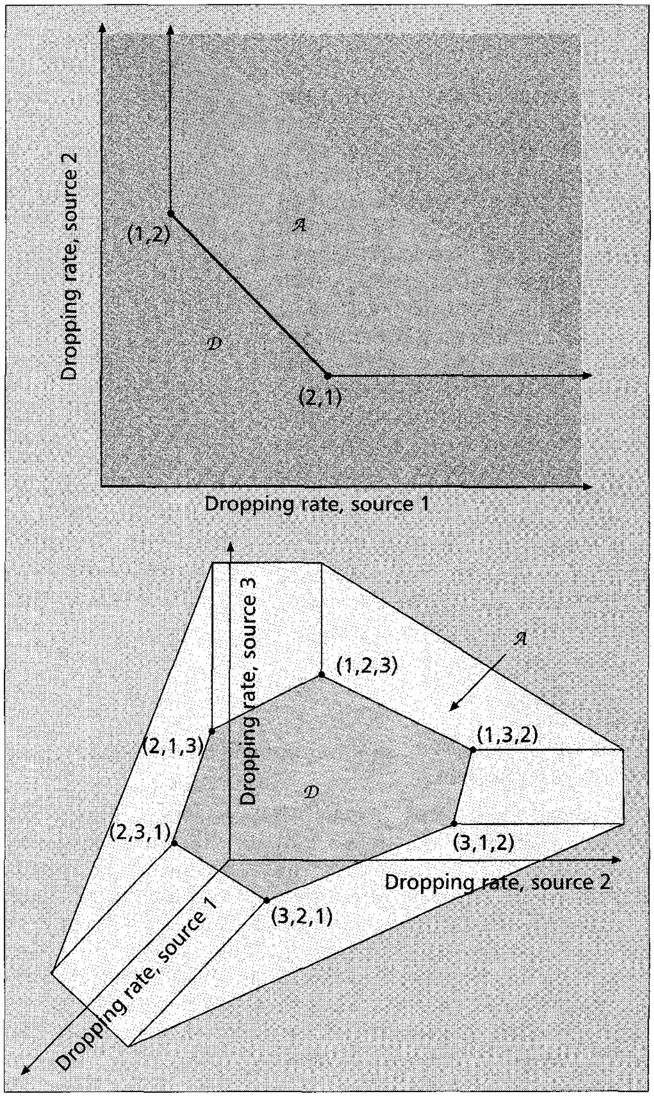
In the realization[2] in Fig. 5, if a WC-EDD policy is employed, then four out of the five packets with a service deadline (or due date) in the current frame will be served, the remaining one dropped, and the new arrival from class S_2 will form the residual traffic in the next frame $(n + 1)$. According to any WC-EDD policy, all three packets with service deadlines in frame $(n + 1)$ will be served during the frame, as well as one of the two packets with service deadlines in frame $(n + 2)$. Which new arrival is serviced in frame $(n + 2)$, and which forms the residual, is a function of the scheduling policy. As long as a WC-EDD policy is employed, the aggregate residual traffic is conserved. That is, the scheduling policy allows for diversification of the individual residual traffic, but the aggregate is conserved. Since the aggregate residual traffic is conserved, the aggregate dropping rate is also conserved.

Although the total aggregate residual traffic $\lambda^r_{S_2}$ is independent of the policy f, the aggregate residual traffic from any subset g, $g \subset S_2$, is dependent on the selected policy.

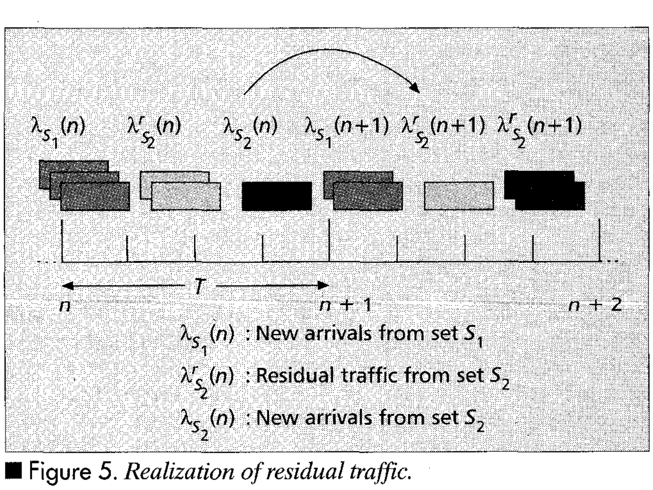
Future evolution of the total aggregate residual traffic process, $\lambda^r_{S_2}(n + 1)$, depends on the present values of $\lambda^r_{S_2}(n)$, $\lambda_{S_1}(n)$, and $\lambda_{S_2}(n)$. Therefore, if an independent identically distributed (iid) arrival process is assumed, then $\lambda^r_{S_2}(n)$ is Markovian and its stationary distribution can be easily derived [9]. Once the distribution of the residual traffic is found, a similar approach is used to establish necessary and sufficient conditions in order for a QoS vector **d** to be achieved by a WC-EDD policy f.

The extreme points of the region of achievable QoS (established for WC-EDD policies) are the dropping rate vectors resulting from Deadline-Sensitive Ordered-HoL (DSO-HoL) policies. A DSO-HoL priority service policy is defined as a policy which first separates packets into two sets: those with a service deadline in the current frame and those with a service deadline in the next frame. Packets from sources with a service deadline in the current frame are serviced according to a priority service policy, $\pi = (\pi_1, \pi_2,\dots,\pi_N)$; $\pi_i \in \{1, 2,\dots,N\}$, $\pi_i \neq \pi_j$, $1 \leq i,j \leq N$. The index of π_i indicates the order of the priority given to the π_i source to service packets having a service deadline in the current frame. None of the π_j sources, $j > i$, may be served as long as packets with current service deadline from sources π_k, $k \leq i$, are present. After servicing the packets having a current deadline, the same service policy, $\pi = (\pi_1, \pi_2, \dots, \pi_N)$, is followed for packets from sources that are present which do not have a current deadline.

From the example shown in Fig. 5, it is evident that the employed EDD policy imposes restrictions on the level of QoS diversification that could be achieved otherwise. For instance, new arrivals from class S_2 cannot be serviced in the presence of packets from class S_1, imposing a limit on the minimum dropping rate for sources in class S_2. This limit is higher than the dropping rate achieved if, for instance, all packets (new and residual) from sources in class S_2 had service priority over those in S_1.



■ Figure 4. *Call admission region for a two- (N = 2) and three- (N = 3) source system.*

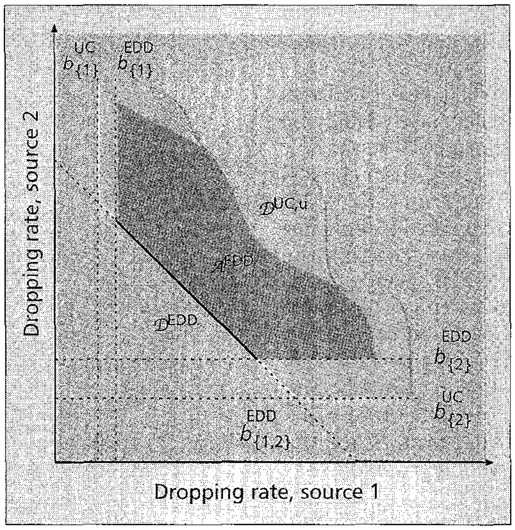

$\lambda_{S_1}(n)$
$\lambda^r_{S_2}(n)$
$\lambda_{S_2}(n)$
$\lambda_{S_1}(n+1)$
$\lambda^r_{S_2}(n+1)$
$\lambda^r_{S_2}(n+1)$
T
n
n + 1
n + 2
$\lambda_{S_1}(n)$: New arrivals from set S_1
$\lambda^r_{S_2}(n)$: Residual traffic from set S_2
$\lambda_{S_2}(n)$: New arrivals from set S_2

■ Figure 5. *Realization of residual traffic.*

In most of this section, the family of WC-EDD policies is considered for the following reasons. First, the WC-EDD policies are known to minimize the *system* packet dropping (delay violation) probability [16], resulting in throughput maximization. Unlike a centrally controlled case in which the service deadlines or due dates would form a continuum or may be drawn from a large collection of values, only two service deadlines (due to the limited exchange of control information) are considered in the TDMA environment in this section. As a consequence, a potentially large number of packets from different sources will have identical service deadlines or due dates (one of two values); thus, significant room for dropping rate diversification may be possible without departing from the WC-EDD policies.

[2] *In this diagram* T *is chosen to be equal to* L *for illustration purposes; however, typically* T ≤ L.

If the QoS vector is not in the region of achievable QoS vectors under the WC-EDD policies, it can be concluded that such a level of QoS diversification *may* be achieved only at the expense of system throughput [17]. This might suggest that the sharing of resources by such diverse applications may need to be restricted by allowing for resource sharing by less diverse applications. However, by deriving an upper bound on the region of achievable QoS vectors under *any* policy, it can be determined whether a given QoS vector is not achievable.



■ Figure 6. *Necessary performance bounds under any policy and the sufficient bounds for the WC-EDD family of policies.*

As discussed above, the region of achievable QoS vectors induced by WC-EDD policies (denoted here $\mathcal{D}^{EDD}$) is described by the set of vectors whose components satisfy,

$$\begin{aligned} &d_g \ge b_g^{EDD} \quad \forall g \subset S, \\ &d_S = b_S^{EDD}; \end{aligned} \tag{8}$$

where b_S^{EDD} and b_g^{EDD} are lower bounds on the performance of WC-EDD policies.

It is well known that the WC-EDD policies optimize the system (S) performance by minimizing the system dropping rate. Therefore, any policy that attempts to improve (decrease) the dropping rate for a subset of sources g beyond the lower bound by relaxing the EDD condition will result in an increased system (S) dropping rate. That is, the lower bounds on the dropping rates achieved by the WC-EDD policies and the class of policies which do not necessarily satisfy the EDD condition (denoted here as an unconstrained, UC, policy) satisfy the following conditions:

$$\begin{aligned} &b_g^{EDD} \ge b_g^{UC} \quad \forall g \subset S, \\ &b_S^{UC} \ge b_S^{EDD}. \end{aligned} \tag{9}$$

b_g^{UC} is the *unconstrained* lower bound on the dropping rate for the sources in set g. This bound is achieved by considering that packets from sources in set g *only* are present and serviced under a WC-EDD policy; sources in $\{S - g\}$ are considered absent.

Since *no* policy can do better for sources in set g than b_g^{UC} when all sources in S are present, the following necessary conditions must be satisfied by *any* QoS vector $\mathbf{d}$ which is achieved under some policy,

$$\begin{aligned} &d_g \ge b_g^{UC} \quad \forall_g \subset S, \\ &d_S \ge b_S^{EDD}. \end{aligned} \tag{10}$$

The inequalities in Eq. 10 are only necessary and not sufficient in order for a QoS vector $\mathbf{d}$ to be achieved by some policy. Therefore, the inequalities in Eq. 10 provide for an upper bound on the region of QoS vectors achieved under *any* policy, denoted as $\mathcal{D}^{UC,u}$, and contains $\mathcal{A}^{EDD}$ — the call admission region under the EDD family of policies. Figure 6 depicts the regions $\mathcal{D}^{EDD}$, $\mathcal{A}^{EDD}$, and $\mathcal{D}^{UC,u}$ for the case of two sources.

The upper bound, $\mathcal{D}^{UC,u}$, depicted in Fig. 6 contains all vectors that can be achieved under *any* policy; therefore, the *call admission region* (CA) is a subset of $\mathcal{D}^{UC,u}$. That is, the region $\mathcal{D}^{UC,u}$ contains the set of all vectors whose performance is induced by any policy.

Effects of the Wireless Channel and an Error Control Scheme on the Region of Achievable QoS and the Call Admission Region

Many error control architectures have been proposed to enable wireless ATM [18, 19]. The primary focus of these works is to design a forward error control (FEC) code and an ATM header error correction (HEC), producing a concatenated code which protects the ATM cell header against undetected errors; undetected errors in an ATM cell header can lead to misrouting and other impairments. In [18], the architecture also involves a traditional data link automatic repeat request (ARQ) protocol applied only to traffic that is not sensitive to increased delays. Recently, an ARQ scheme has been developed to accommodate real-time service [20]. The result is achieved by scheduling the feedback — acknowledgments (ACKs) and negative acknowledgments (NAKs) based on the relative urgency of the transmitted packets; in addition, if the due date of the packet has expired, the packet is discarded and the receiver notified. In this manner, the number of retransmissions of each packet is controlled. In [10] and in this work, the probability of correct reception of a packet is increased (or the probability of dropping reduced) by transmitting multiple copies of certain packets (before the packet has expired) without any feedback. In this way, the probability of packet dropping can be controlled by the scheduling policy which controls the number of copies transmitted from each source based on the QoS required by the applications. The focus of this section is to examine the impact of the wireless channel and a simple QoS-sensitive FEC scheme on the region of achievable QoS.

Effects of the Wireless Channel

Although the necessary resources may become available on time, packets may be corrupted due to channel errors and be dropped at the receiver. Under these conditions, the performance is limited by the interference introduced in the channel (interference-limited). Such packet discarding may occur with a frequency comparable to that of packet discards at the transmitter due to resource limitations. As a consequence, the region of achievable QoS vectors is shaped by the packet discarding process at both the transmitter and the receiver due to resource and interference limitations, respectively. In this section the effect of the wireless channel on the region of achievable QoS established earlier is studied. In that system, packets from a source which do not receive service over a frame following their arrival are considered to have excess delay and are dropped at the source.

The effects of the wireless channel are modeled as in [18], where a Gaussian noise channel with random bit erasure interference is considered. The erasure process might be produced by a burst noise process which produces bursts of erasures. However, in [18], an interleaver/deinterleaver is employed to turn the erasure bursts into statistically independent bit erasures. Therefore, in this system packet erasures are considered to be statistically independent and occur when the interference in the channel is such that the packet is corrupted beyond correction. The corrupted transmitted packets are discarded (dropped) at the receiver.

The event that the packet is corrupted and therefore dropped is a function of interference in the channel, the transmitted power, the coding scheme, and the packet length. Let Z be an indicator function of a packet erasure. That is,

$$Z=\begin{cases}1 & \text{if the packet is corrupted}\\ 0 & \text{otherwise}\end{cases}. \tag{11}$$

Therefore, the expected value of Z, $E[Z] = \beta$, is the probability that a packet is corrupted by the channel and dropped.

Considering the combined impact of scheduling policy f and the physical channel, the number of packets from source i dropped in frame n due to the above competition for the resources and the interference in the channel is given by

$$d_i^f(n)=\begin{cases}\sum_{m=1}^{\lambda_i(n)} Z_m & \text{if } \sum_{j=1}^{N}\lambda_j(n)\leq T\\ \lambda_i(n)-a_i^f(n)+\sum_{m=1}^{a_i^f(n)} Z_m & \text{if} \sum_{j=1}^{N}\lambda_j(n)>T\end{cases}. \tag{12}$$

Z_m is an indicator function associated with the transmission of the mth packet from source i. Considering the effects of the wireless channel, it must be determined whether *under given channel conditions* a QoS vector is achievable under some policy f.

The earlier results can be extended to account for channel quality. The system dropping rate is derived in [10] and is given by

$$\begin{aligned}d_S^f = {}& \{E[\lambda_S(n)\,|\,\lambda_S(n) > T] - T(1-\beta)\}P(\lambda_S(n) > T)\\ &+\beta\{E[\lambda_S(n)\,|\,\lambda_S(n) \leq T]\}P(\lambda_S(n) \leq T).\end{aligned} \tag{13}$$

As can be seen from Eq. 13, d_S^f is independent from policy f; it only depends on the aggregate arrival process, the number of resources T, and the channel characteristics β. Therefore, the system dropping rate, d_S^f, is conserved under any work-conserving policies f and is denoted b_S.

The lower bound on the aggregate packet dropping rate (with channel quality β) for sources in set g is given by

$$\begin{aligned}b_g = {}& \{E[\lambda_g(n)\,|\,\lambda_g(n) > T] - T(1-\beta)\}P(\lambda_g(n) > T)\\ &+\beta\{E[\lambda_g(n)\,|\,\lambda_g(n) \leq T]\}P(\lambda_g(n) \leq T).\end{aligned} \tag{14}$$

The region of achievable QoS is shaped by both the amount of available resources and the level of interference in the wireless channel. To illustrate this, consider the following example of two sources competing for T slots in a TDMA frame. The source packet arrival processes are assumed to be mutually independent. Each arrival process is embedded at the frame boundaries. The number of packets generated (and requesting service) by a source in the current frame is (probabilistically) determined by the present state of the underlying arrival process.

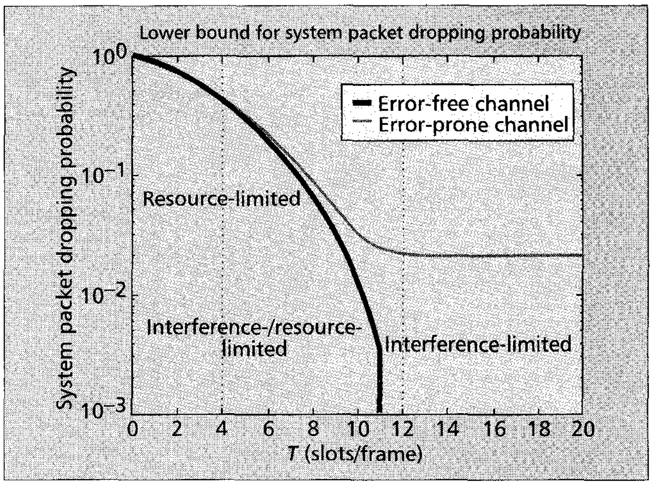
In this example, each VBR source is modeled by discrete-time batch Markov arrival process embedded at frame boundaries, with mean rate of 3.6 and 3.2 packets per frame, and variance of 2.04 and 3.36 packets per frame, respectively. Such an arrival process could be used to model a low rate video source, such as H.263.

In Fig. 7 the conserved system packet dropping probability

$$p_S=\frac{b_S}{E[\lambda_S(n)]}$$

is plotted as a function of available resources T (time slots) for an error-free channel (i.e., $\beta = 0$) and a wireless channel with channel quality $\beta = 0.02$.

As can be seen in this figure, there are three distinct regions of operations: *resource-limited*, *interference-/resource-limited*, and *interference-limited*. In the resource-limited region, the performance is primarily determined by the amount of available resources. This result is evident since the packet dropping probabilities for the systems with the error-free channel and the (nonideal) wireless channel are almost identical. In the interference-limited region, the dropping probability in the error-free channel is zero, while the performance in the wireless system is limited by the interference and given by $\beta = 0.02$. The performance in the interference-/resource-limited region is determined by both the available resources and the level of interference in the channel. In this example, the system packet dropping probability in this region ranges from 10^{-1} to 10^{-2}, an operation region of interest for real-time applications. It is important to note (as shown earlier) that, in general, satisfying the system packet dropping rate (probability) is only necessary and not sufficient to guarantee that the target QoS vector is achievable.



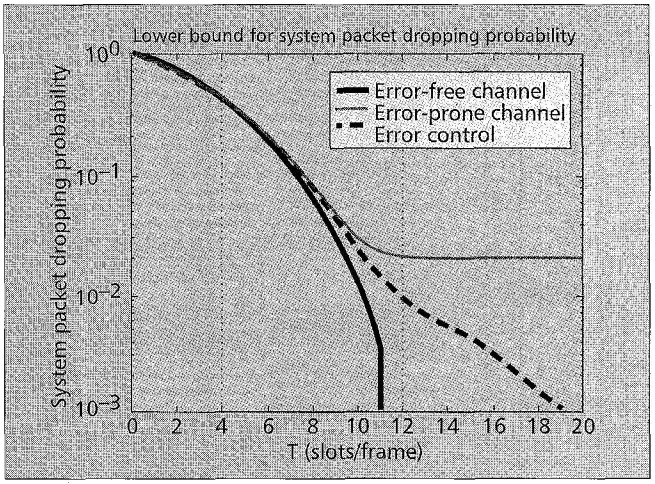
■ Figure 7. *System packet dropping probability in an error-free channel and a (nonideal) wireless channel with channel conditions* β *= 0.02.*

The Impact of QoS-Sensitive Error Control

An error control scheme can be added to combat the effects of the wireless channel in an interference-/resource-limited or interference-limited system. Due to the real-time constraints of the supported applications, traditional ARQ strategies are not possible. In this section, a simple FEC scheme is described as in [10] to illustrate the impact of this layer on the region of achievable QoS and the CA region.

To combat the effects of interference, the error control scheme considered in this work will generate multiple copies[3] of certain packets for transmission over the current frame. This strategy will improve the probability of correct reception (or reduce the probability of packet dropping at the receiver) while meeting the real-time service constraint. Copies are transmitted only during underloaded frames utilizing the remaining resources ($T - \lambda_S(n)$). Transmitting a copy from a set g during an overloaded frame would reduce the probability of packet dropping at the receiver for set g, but would force an original packet from the complement set $\{S - g\}$ to be dropped at the source. The effect would be that the aggregate dropping rate for subset g of sources is reduced, but the aggregate dropping rate for complement set $\{S - g\}$ is increased by an amount greater (in realistic systems) than the decrease attained for subset g, causing the overall system dropping rate to increase. In view of the above discussion, if

[3] *The copies are not used in an error correction scheme.*



■ Figure 8. *The impact of the QoS sensitive error control scheme on the system packet dropping probability in a wireless channel with channel conditions* β = *0.02.*

the objective of the error control protocol is to minimize the system packet dropping probability (or, equivalently, packet dropping rate), and therefore maximize system throughput, then multiple copies of packets can be sent only during underloaded frames — utilizing the remaining resources.

During underloaded frames, the sender generates multiple copies of certain packets for transmission over the current frame. The number of copies generated by the sender is a function of the scheduling policy and the amount of remaining resources. It is shown in [10] that a policy which attempts to "fairly" allocate the remaining resources among sources will result in the minimum system dropping rate. The policy is fair in the sense that the number of (re)transmissions allocated to source i is proportional to the number of packets requiring service in the frame,

$$\frac{T}{\lambda_S(n)}\lambda_i(n).$$

The results found earlier can be modified to account for the impact of the QoS-sensitive error control protocol. Considering the effects of the physical channel, the scheduling policy and the error control protocol, the number of packets from source i dropped in frame n is given by

$$d_i^f(n) = \begin{cases} \sum_{m=1}^{\lambda_i(n)} \prod_{q=1}^{R+1_m^f} Z_m^q & \text{if } \lambda_S(n) \le T \\ \lambda_i(n) - a_i^f(n) + \sum_{m=1}^{a_i^f(n)} Z_m & \text{if } \lambda_S(n) > T \end{cases}, \quad 1 \le i \le N. \quad (15)$$

Due to the granularity in the system (i.e., resources can be allocated only in integer multiples), all packets are transmitted

$$R \overset{\Delta}{=} \left\lfloor \frac{T}{\lambda_S(n)} \right\rfloor, \quad R \ge 1$$

number of times, where $\lfloor . \rfloor$ denotes the integer part, and a subset of packets $X \overset{\Delta}{=} (T - \lambda_S(n)R)$ can be transmitted one additional time, $(R + 1)$. $1_m^f \in \{0, 1\}$ and indicates the dependency of the additional transmission of a copy of packet m on the fair policy f, where

$$\sum_{m=1}^{\lambda_S(n)} 1_m^f = X, \ \forall f.$$

Z_m^q is an indicator function associated with the qth transmission of the mth packet from source i. The system dropping rate is conserved regardless of to which set of sources the packets that are transmitted one additional time belong. It is easy to show that the system dropping rate is given by

$$\begin{aligned} b_S = {} & \{E[\lambda_S(n) | \lambda_S(n) > T] - T(1 - \beta)\} P(\lambda_S(n) > T) \\ & + E[X\beta^{R+1} | \lambda_S(n) \le T] P(\lambda_S(n) \le T) \\ & + E[(\lambda_S(n) - X)\beta^R \mid \lambda_S(n) \le T] P(\lambda_S(n) \le T). \end{aligned} \quad (16)$$

Therefore, under any *fair work-conserving policy*, the system dropping rate (given in Eq. 16) is minimum and also conserved.

With the addition of this error control protocol, the lower bound b_g for the aggregate packet dropping rate for sources in g under any fair work-conserving policy f is determined to be

$$\begin{aligned} b_g = {} & \{E[\lambda_g(n) | \lambda_g(n) > T] - T(1 - \beta)\} P(\lambda_g(n) > T) \\ & + E[\min[\lambda_g(n), X]\beta^{R+1} \mid \lambda_g \le T] P(\lambda_g(n) \le T) \\ & + E[\max[0, \lambda_g(n) - X]\beta^R \mid \lambda_g(n) \le T] P(\lambda_g(n) \le T). \end{aligned} \quad (17)$$

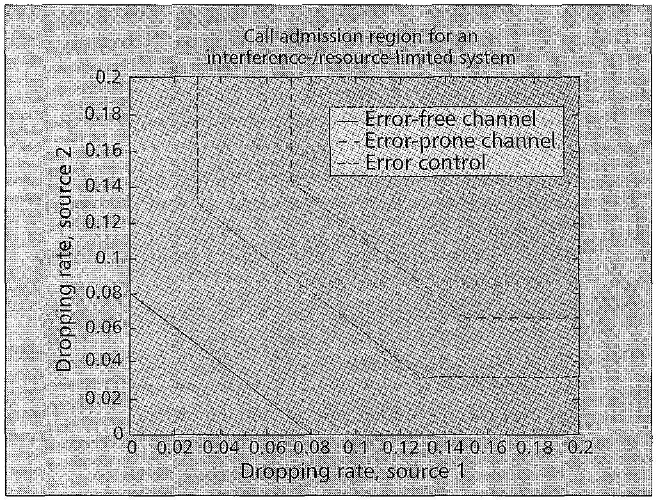
The expected value in Eq. 17 is with respect to $\{\lambda_g(n), \lambda_S(n)\}$, which is easily computed since $P(\lambda_g(n) = i, \lambda_S(n) = j) = P(\lambda_g(n) = i)P(\lambda_{\{S-g\}}(n) = j - i)$. The expressions in Eqs. 16 and 17 reduce to Eqs. 13 and 14, respectively, for R fixed and equal to 0. The extreme points in this case correspond to Fair-Ordered-HoL (F-O-HoL) priority service policies. An F-O-HoL service policy is an O-HoL service policy $\pi = (\pi_1, \pi_2, \ldots, \pi_N)$ in which the additional retransmissions are also allocated according to π.

With the addition of the QoS-sensitive FEC scheme, the region of achievable QoS for an interference-/resource-limited or interference-limited system can be improved compared to the system not employing the error control protocol. The impact this protocol has on the system packet dropping probability for the example described earlier is shown in Fig. 8.

In this figure, the performance of the two systems is compared to that in an error-free environment. As can be seen in this figure, the system employing the error control scheme induces a lower packet dropping probability (and therefore

Theoretical results			
	$b_{\{1\}}$	$b_{\{2\}}$	$b_{\{1,2\}}$
Error-free	0.0	0.0	0.08
Error-prone	0.0720	0.0640	0.2144
Error control	0.0303	0.0316	0.1606

■ Table 1. *The impact of channel quality and error control on the region of achievable QoS.*



■ Figure 9. *The call admission region 𝒜 for the two-source system given in the third section, in an error-free channel, and in a wireless channel (*β = *0.02) with and without QoS-sensitive error control.*

higher throughput) than the system without it. The impact is most significant in the interference-/resource-limited and interference-limited regions. In these regions the system with the error control scheme takes advantage of the remaining resources and can reduce the packet dropping probability. In the interference-limited region, the packet dropping probability can be made arbitrarily small if resources are abundant.

The impact the QoS-sensitive error control has on the region of achievable QoS $\mathcal{D}$ and the call admission region $\mathcal{A}$ is illustrated in Table 1 and Fig. 9, respectively. In this figure, $\mathcal{A}$ is derived for the system of sources given earlier and with $T = 10$ slots. Thus, according to Fig. 7, the system is interference/resource-limited. As it can be seen, the proposed error control protocol moves the call admission region toward lower dropping rates. This implies that a larger collection of QoS vectors can be accommodated.

Scheduling Achievable Performance

The development of the region of achievable QoS $\mathcal{D}$ presented above leads to the identification of a class of scheduling polices capable of delivering any achievable performance. This result can be derived (with minor modifications) for any of the regions present in this article.

The result follows from the fact that $\mathcal{D}$ can be written as a convex combination of the extreme points (vertices) $\mathbf{d}_{\text{ext-}i}$ of $\mathcal{D}$. That is, if $\mathbf{d} \in \mathcal{D}$, then

$$\mathbf{d} = \sum_{i=1}^{N!} \alpha_i \mathbf{d}_{\text{ext}-1}$$

for some $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{N!})$ where

$$\alpha_i \geq 0,\ 1 \leq i \leq N!,\ \sum_{i=1}^{N!} \alpha_i = 1.$$

Therefore, by selecting the (O-HoL, DSO-HoL, or F-O-HoL, depending on the environment under consideration) priority policy that induces the extreme point $\mathbf{d}_{\text{ext}-i}$ of $\mathcal{D}$ with probability α_i, any QoS vector in $\mathcal{D}$ can be delivered. This class of policy is referred to as a *mixing* (O-HoL, DSO-HoL, or F-O-HoL) priority policy. In most systems, several mixing priority policies exist that can deliver the target dropping rate vector. This allows for the incorporation of additional constraints representing other desirable qualities of the policies which are important in wireless ATM. For instance, among all the mixing policies inducing $\mathbf{d}$, the one which minimizes the variance of the service provided to certain sources may be selected.

Conclusion

In this article the call admission region is precisely described for a system of real-time heterogeneous VBR sources competing for an unreliable wireless channel (slots of a TDMA frame). The QoS has been defined in terms of a maximum tolerable packet delay and packet dropping probability (or, equivalently, packet dropping rate). Packets from sources in the system were dropped as a result of delay violations (slots may not become available on time and a source is forced to drop the packets with excess delay) and channel-induced errors (corrupted packets are dropped at the receiver). As a consequence, it has been shown that the call admission region is shaped by both the interference in the physical channel and the amount of available resources, and it can be impacted by a QoS-sensitive error control scheme.

References

[1] D. Raychaudhuri *et al.*, "WATMnet: A Prototype Wireless ATM System for Multimedia Personal Communications," *IEEE JSAC*, vol. 15, no. 1, 1997, pp. 83–95.

[2] D. Petras *et al.*, "Medium Access Control Protocol for Wireless, Transparent ATM Access," *Proc. 1995 Wireless Commun. Sys. Symp.*, Long Island, NY, Nov. 1995.

[3] M. Karol, Z. Liu, and K. Eng, "An Efficient Demand-Assignment Multiple Access Protocol for Wireless Packet (ATM) Networks," *ACM Wireless Networks*, vol. 1, no. 4, Dec. 1995, pp. 267–79.

[4] A. Mahmoud, D. Falconer, and S. Mahmoud, "A Multiple Access Scheme for Wireless Access to a Broadband ATM LAN Based on Polling and Sectored Antennas," *IEEE JSAC*, vol. 14, no. 4, May 1996.

[5] C. Chang *et al.*, "Guaranteed Quality-of-Service Wireless Access to ATM Networks," *IEEE JSAC*, vol. 15, no. 1, Jan. 1997.

[6] B. Walke *et al.*, "Wireless ATM: Air Interface and Network Protocols of the Mobile Broadband System," *IEEE Pers. Commun.*, vol. 3, Aug. 1996.

[7] G. Anastasi, D. Grillo, and L. Lenzini, "An Access Protocol for Speech/Data/Video Integration in TDMA-Based Advanced Mobile Systems," *IEEE JSAC*, vol. 15, no. 8, Oct. 1997.

[8] J. Capone and I. Stavrakakis, "Achievable QoS and Scheduling Policies in Integrated Services Wireless Networks," *Perf. Eval.*, vols. 26 and 27, no. 1, Oct. 1996.

[9] J. Capone and I. Stavrakakis, "Delivering Diverse Delay/Dropping QoS Requirements in a TDMA Environment," *Proc. ACM MobiCom*, Budapest, Hungary, Sept. 1997.

[10] J. Capone and I. Stavrakakis, "Achievable QoS in an Interference/Resource-Limited Shared Wireless Channel," submitted to *IEEE JSAC*.

[11] P. T. Brady, "A Model for On-Off Speech Patterns in Two-way Conversation," *Bell Sys. Tech. J.*, vol. 48, Sept. 1969, pp. 885–90.

[12] J. De Vile, "A Reservation Multiple Access Scheme for Adaptive TDMA Air Interface, *Proc. 4th WINLAB Wksp Third Generation Wireless Info. Networks*, Oct. 1993.

[13] F. Li and L. F. Merakos, "Voice Data Integration in Digital TDMA Cellular System," *Proc. Int'l. Zurich Sem. Mobile Commun.*, Zurich, Switzerland, Mar. 1994.

[14] P. Narasimban and R. Yates, "A New Protocol for the Integration of Voice and Data Traffic over PRMA," *IEEE JSAC*, vol. 14, no. 4, May 1996.

[15] A. Brøndsted, *An Introduction to Convex Polytopes*, Springer-Verlag, 1983.

[16] S. S. Panwar, D. Towsley, and J. K. Wolf, "Optimal Scheduling Policies for a Class of Queues with Customer Deadlines to the Beginning of Service," *J. ACM*, 1988, pp. 832–44.

[17] G. Chen and I. Stavrakakis, "ATM Traffic Management with Diversified Loss and Delay Requirements," *Proc. IEEE INFOCOM*, San Francisco, CA, Mar. 1996.

[18] J. B. Cain and D. McGregor, "A Recommended Error Control Architecture for ATM Networks with Wireless Links," *IEEE JSAC*, vol. 15, no. 1, Jan. 1997.

[19] Y. Nakayama and S. Aikawa, "Cell Discard and TDMA Synchronization Using FEC in Wireless ATM Systems," *IEEE JSAC*, vol. 15, no. 1, Jan. 1997.

[20] D. Petras and A. Hettich, "Performance Evaluation of a Logical Link Control Protocol for an ATM Air Interface," To appear, *Int'l. J. Wireless Info. Networks*, 1997.

Additional Reading

[1] E. G. Coffman and I. Mitrani, "A Characterization of Waiting Time Performance Realizable by Single-Server Queues," *Ops. Res.*, vol. 28, no. 3, May-June 1980, pp. 810–82.

[2] W.W Sharkey, "Cooperative Games with Large Cores," *Int'l. J. Game Theory*, vol. 11, 1982, pp. 175–82.

[3] D. J. A. Welsh, *Matroid Theory*, Academic Press, 1976.

Biographies

JEFFREY M. CAPONE [M] (jcapone@asu.edu) received a B.S.E.E. degree from the University of Vermont, Burlington, in 1992, and M.S.E.E. and Ph.D. degrees from Northeastern University, Boston, Massachusetts, in 1995 and 1997, respectively. In 1997, he joined the faculty of Electrical Engineering at Arizona State University. His primary research interest is in the design and analysis of controlling policies for bandwidth management in integrated services wireless communication networks, multihop packet radio networks and large-scale network models. He is a member of the Technical Committee on Computer Communications. He has served on the technical committee for GLOBECOM, and the program committee for MobiCom and ICASSP.

IOANNIS STAVRAKAKIS (ioannis@cdsp.neu.edu) received a Diploma in electrical engineering from the Aristotelian University of Thessaloniki, Greece, 1983, and a Ph.D. degree in electrical engineering from the University of Virginia, 1988. In 1988, he joined the faculty of Computer Science and Electrical Engineering at the University of Vermont as an assistant and then associate professor. Since 1994, he has been an associate professor of Electrical and Computer Engineering at Northeastern University, Boston. His research interests are in stochastic system modeling, teletraffic analysis and discrete-time queueing theory, with primary focus on the design and performance evaluation of Broadband Integrated Services Digital Networks (B-ISDN). Dr. Stavrakakis is a senior member of IEEE and a member of the IEEE communications Society, Technical Committee on Computer Communications.