

# A Two-level Resource Management Scheme in Wireless Networks Based on User-Satisfaction \*

Sourav Pal<sup>a</sup>

spal@cse.uta.edu

Mainak Chatterjee<sup>b</sup>

mainak@cpe.ucf.edu

Sajal K. Das<sup>a</sup>

das@cse.uta.edu

<sup>a</sup>Department of Computer Science and Engineering, University of Texas at Arlington, TX, USA

<sup>b</sup>Department of Electrical and Computer Engineering, University of Central Florida, FL, USA

*The success of future generation wireless data services will depend on the parameterized provisioning of quality of service (QoS) for applications whose demands and nature are highly heterogeneous. Also, user satisfaction will play a key role in the economic viability of wireless service deployments. In this paper, we present a QoS framework based on the paradigm of traffic class and user satisfaction. We address the problem of dealing with subjectiveness of user satisfaction or expectation from service providers by defining what we call user irritation factors, using Sigmoid functions. These factors reflect users' sensitivity and tolerance to delay. The proposed class-based QoS framework comprises a radio resource management scheme which considers user satisfaction based on the perceived QoS, and caters to heterogeneous applications that have diverse QoS requirements. Our resource management scheme has two components: the admission control algorithm caters to the long term user satisfaction while the session-based rate and bandwidth allocation scheme manipulates the short term user satisfaction. Soft-reservation schemes are also proposed to cater to the higher paying users. Performance metrics have been specifically defined for each traffic class. Extensive simulations using four types of traffic and three classes of users reveal that the proposed framework offers improved QoS without compromising the utilization of the system.*

## I. Introduction

The demand for wireless data services has led to the evolution of third generation (3G) wireless services which deliver a broad range of multimedia applications to mobile users. The transition from traditional voice services to data services with heterogeneous requirements necessitates a revisit of the radio resource management schemes. Resource management must consider the impact of error prone transmission medium, heterogeneity of application requirements, and issues related to fairness among users. Also, there is a need to differentiate users based on the amount of revenue they are willing to pay and their expectations from the services.

We envision that the success of wireless data services in conjunction with traditional voice services would ultimately depend on *user satisfaction*. Thus a QoS framework needs to be developed that identifies user satisfaction and also facilitates negotiation between the users and the service providers. Identifying the relevant QoS for each of the diverse services and distinguishing the variation of user satisfac-

tion with the perceived QoS is an important research challenge. User satisfaction depends on the subjective expectation of the service and hence varies with the type of services.

Several resource management schemes have been proposed which address the various requirements like heterogeneous service demands, fairness, starvation, channel transmission error and delay bounds (for example see [16] and references therein). They include power control algorithms at the physical layer, scheduling or rate-adaptation algorithms at the medium access control (MAC) layer and service admission algorithms at the network layer. However, many such schemes prioritize the voice services and allocate the residual bandwidth to non-real time applications which are deprived of any assurance on the delay bound. A user may care less about the per-packet delay and might put more emphasis on the download time of the *entire* data. This motivates us develop resource management schemes which strives to preserve the QoS requirements of different heterogeneous services, while maintaining fairness amongst different classes of users.

In this paper, we propose a radio resource management framework which tries to adhere to the QoS re-

\*This work is partially supported by NSF ITR grant IIS-0326505.

requirements of the applications by exploiting the *subjectiveness* associated with user satisfaction. We categorize users into different classes based on the revenue paid, and consider that all the classes are endowed with heterogeneous wireless services. We propose the notions of *short term irritation* and *long term irritation*, extend them to multiple traffic classes and propose service (call) admission algorithms and scheduling policies. These policies are tailored for Universal Mobile Telecommunications System (UMTS) defined traffic. Priorities are also considered among these classes. The proposed two-level resource management scheme try to improve the delay in delivering the entire non-real-time content and real-time traffic (conversational/streaming) by taking into consideration the short- and long-term effects of user irritation. More specifically, the proposed call admission control algorithm regulates the long-term irritation of the users, whereas the short-term satisfaction of the users is guaranteed by the scheduling policy. It not only provides a bound on the delay but also manipulates the resources so as to maintain the irritation of each user below a certain threshold.

The remainder of the paper is organized as follows. The network architecture along with how the user service level agreement and the policy management issues are stored in the databases are discussed in section II. The various traffic classes and the right metric for QoS characterization are presented in section III. The subjective user satisfaction model and its dependence on different service types and perceived QoS is studied in section IV. The two-level radio resource management scheme is proposed in section V while section VI presents the simulation model and the experimental results. Conclusions are drawn in the last section.

## II. Architecture and Policies

In this section, we discuss the network architecture along with the databases which contain the user profiles. These databases also hold the policies pertaining to the user classes and their respective quality of service expectations from the system.

### II.A. Network Architecture

The resource management framework on which our proposed scheduling algorithm operates is based on the architecture components proposed in the IETF Policy Information Base (PIB) [6] for differentiated services. Figure 1 shows the architectural overview of a 3G wireless cellular network that consists of three

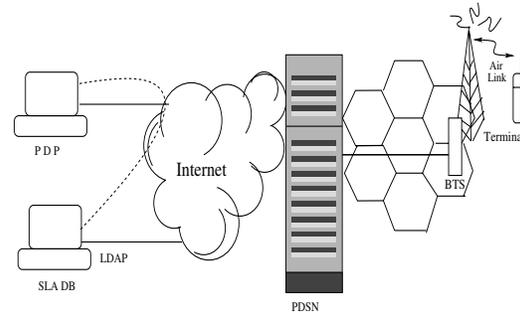


Figure 1: Network Architecture

important network elements : (1) the airlink and terminal, (2) the base station transceiver system (BTS) and (3) packet data service node (PDSN). Each cell executes a unique copy of the proposed scheduling algorithm for handling requests generated in that cell. In the 3G system, the PDSN includes the gateway functions that interconnect the Internet domain. The PIB framework includes components like (1) Service Level Agreement (SLA) database, (2) policy decision point (PDP), and (3) policy execution point (PEP). The air link supports *uplink* and *downlink* channels. The uplink channel transmits the request of the clients to the server, where the scheduler schedules the data to the clients through the downlink channels. Though the uplink bandwidth is smaller, we assume that there exist no uplink channel contention between different clients sending requests to the server. Location dependent channel transmission errors which are bursty in nature needs to be considered for the scheduler design to achieve accuracy and efficiency of the system.

### II.B. Policy Management and SLA Issues

It is to be noted that many wireless carriers are already adopting what can be called *differentiated service contracts* for voice services to mobile users. However, no similar effort has been made for data services. The objective here is to create different classes of customers based on the selected service packages for data services. The distinction in QoS levels lies in the bandwidth provided and hence the delay and throughput offered to the different classes of users. The policy management and the customer service agreements are stored in the SLA database as a set of rules. The policy rules are created using the IntServ/DiffServ QPIM (quality policy information modeling) technique as reported in most IETF drafts, for example, in [13]. The PDP contains all the PIBs [6] in addition to the MIBs (management information bases) required for policy management. The PDP function for the differentiated

services can be located in the PDSN or mobile switching center(MSC). The policy execution function in this case remains within the radio network controller (RNC). We consider three service classes - *Gold*, *Silver*, and *Bronze*, where each user class supports all the services. We propose that the PIB support all three user classes but with different commitment levels. The classification being dependent on the QoS they expect from the network and the revenue they are willing to pay. The Gold class pays the highest revenue and the Bronze class pays the least. The delay suffered for the same service is thus highest for an Bronze class client and least for a Gold Class client. The fairness of service in this context is relative to the price paid by the clients. The scheduler internally classifies the users on the basis of the *user irritation factor (UIF)*, assigning higher UIF to a user with higher priority. This is modeled in the next section.

The PIB also specifies that service to client requests would be processed in the following manner. *Guaranteed QoS* mode of service implies that the system will be able to honor the bounded delay. *Negotiable QoS* is when the system possesses insufficient or no resource at all for a request made. However, if the system is optimistic about being able to serve the request within the bounded delay with the anticipation that on-going transmissions might release resources in near future, then the request is admitted. Nonetheless, the admittance is not strict in nature and the delay restrictions might be violated. The third scenario is when the system is well aware that under no circumstances it can honor the client request within the specific deadline. It simply rejects the request.

### III. Traffic Classes and QoS Metrics

The proposed QoS framework caters to the diverse multimedia applications as well as the traditional voice calls in wireless networks. The framework supports multiple classes of users with different priorities and enables fair sharing of the radio resource based on the user class and subjective satisfaction. Moreover, service negotiation between users and service providers as in [12] is flexible in the sense that service classes can be pre-configured with the user's applications, or explicitly selected at the time of initiation of the applications.

For the purpose of illustration and simplicity, we consider the following four QoS classes as proposed for UMTS networks [1, 2]: *conversational*, *streaming*, *interactive*, and *background*. However, the pro-

posed framework is generic enough and can be extended to any number of traffic classes and services as desired by the network operator. These heterogeneous traffic/service classes have specific QoS requirements. The main distinguishing factor is the delay sensitivity of each class. We proceed to outline the different attributes of each traffic class and use them to model the user satisfaction.

#### III.A. Conversation Class

This class is mainly intended to be used to carry real-time traffic flows. Time relation (variation) between information entities of the flow is preserved in conversation class. The other fundamental characteristic is that it has extremely stringent and low delay requirement. Voice and video telephony are the target applications that falls within the domain of this class.

We use *time-hysteresis outage probability* [17] as the QoS metric for the conversational class. Outage probability is a classical metric in cellular systems which is defined as the probability that the received signal to noise ratio (SINR) will drop below a specified  $E_b/N_o$ , where  $E_b$  is the energy per bit and  $N_o$  is the noise power. The assumption is that the bit rate requirement and the bit error rates can be mapped onto an equivalent  $E_b/N_o$ .

#### III.B. Streaming Class

The streaming class is very similar to the conversation class but is less delay sensitive. The other distinguishing factor is that applications in this class, such as streaming video, are uni-directional (i.e. one-way transport). However, the fundamental criterion of time relation being preserved between information entities remains the same.

Though delay and bit error rate affect streaming sessions, *rate jitter* is the most important parameter influencing the quality of such traffic. Hence we employ rate jitter as the QoS metric for modeling user satisfaction, or in other words, for quantifying user irritation for streaming class applications.

#### III.C. Interactive & Background Classes

Interactive and background classes are mainly meant to be used by traditional Internet applications like WWW, Email, Telnet, FTP and News. Since the delay requirements of these classes are more slack compared to conversational and streaming classes, they provide better error rate by means of channel coding and retransmission. The main difference between Interactive and background classes is that the interac-

tive class is mainly used by interactive applications like network games and chats, while the background class is meant for background traffic such as emails or web downloading. Responsiveness of the interactive applications is ensured by separating interactive and background applications. Traffic in the interactive class has higher priority in scheduling than background class traffic, so background applications use transmission resources only when interactive applications do not need them. This is very important in wireless environment where the bandwidth is low compared to wire-line networks.

We observe that the delay a user is willing to tolerate before canceling a session can be considered as a suitable metric for these classes. Though the amount of acceptable delay depends on the particular user, it can be treated as a tunable parameter which varies with the user class itself. An additional constraint is that there should be no data loss.

#### IV. Modeling User Irritation Factor

The success of our scheduling algorithm lies in modeling the irritation/satisfaction of the user. A basic understanding of the client irritation, i.e., what amount of performance degradation the customer is ready to suffer without complaining, will enable the scheduler to estimate the resources (i.e., number of channels) that has to be allocated to a particular request. This will directly help in maintaining the delay bound and indirectly help the service provider to control the churn factor [8].

In the following, we propose a method to model the user irritation and present two new metrics: *short term user irritation factor (SUIF)* and *long term user irritation factor (LUIF)*. Each factor signifies different levels of user satisfaction as described below. Qualitatively, *SUIF* measures the delay that the user is ready to suffer prior to which the user decides to change or cancel the particular request. *LUIF* determines the tolerance or irritation of the user resulting from continued degradation of service after which the client decides to cancel service completely. For different classes as specified in the SLA, the *User Irritation Factor (UIF)* will vary. A high priority user paying higher revenue expecting lesser delay will be assigned a higher UIF. The goal of the scheduler will be to schedule requests for each client such that the UIF is not violated.

A Sigmoid function has been used in the literature to approximate the user's satisfaction with respect to service qualities or resource allocations [14, 15]. For

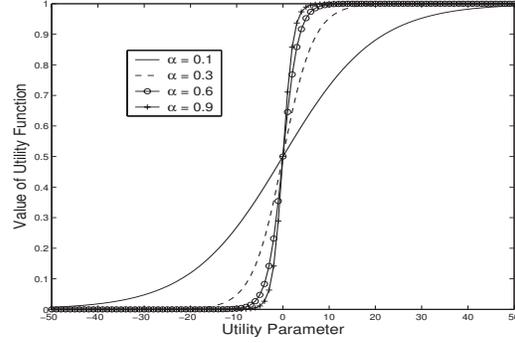


Figure 2: Examples of Utility Functions

modeling the satisfaction/dissatisfaction of users, we also use the Sigmoid function and correlate it with the proposed metrics, SUIF and LUIF. For a random variable  $x$  representing a service parameter like coverage or reliability, the corresponding satisfaction,  $U(x)$ , is given by

$$U(x) = \frac{1}{1 + e^{-\alpha(x-\beta)}}. \quad (1)$$

Here  $\alpha$  and  $\beta$ , determining the steepness and the center of the curve respectively, can be tuned to customize the function for different users. The plots for Equation (1), for different values of  $\alpha$  are shown in Figure 2. From the figure we observe that the satisfaction (utility function) increases with increasing  $x$ . But for parameters like price or delay, the satisfaction decreases with increasing  $x$ . In such cases we can model satisfaction as

$$U(x) = 1 - \frac{1}{1 + e^{-\alpha(x-\beta)}}. \quad (2)$$

The value of  $\alpha$  indicates user's sensitivity to the QoS degradation while  $\beta$  indicates the "acceptable" region of operation. We use Equation (2) to model the UIF. In the following section we determine SUIF and LUIF analytically for each class of traffic.

##### IV.A. Short Term User Irritation Factor

The SUIF is measured on a per-session per-user basis. It is also responsible for distinguishing between a call type - new or handoff call and accordingly model the user irritation. An in-session user if deprived of service due to handoff, would suffer from greater irritation than a user whose request is blocked. Hence, we propose a simple mechanism to assign  $\tau_1$  and  $\tau_2$  signifying the quantitative factors associated with irritation suffered due to a new and handoff call, respectively, where  $\tau_1 < \tau_2 < 1$ . In the rest of the derivations, we shall use  $\tau = \tau_1$  or  $\tau_2$  depending on the request being

a handoff or a new call. Also, all the random variables  $x_{i,j}$  are normalized with the best possible value being 0 (representing zero delay and jitter) and the worst being 1. Equation (2) utilizes the  $x_{i,j}$ s defined later to measure the SUIFs.

**SUIF for Class 1:** A classical performance metric in cellular systems is the *outage probability* which is usually defined as the probability that the QoS provided to an existing connection will drop below a certain threshold, or it is the probability that the received SINR will drop below a specified  $E_b/N_o$ , where  $E_b$  is the energy per bit and  $N_o$  is the noise power. The assumption is that the bit rate requirement and the bit error rates can be mapped onto an equivalent  $E_b/N_o$ . We argue that the SINR translates to the QoS estimation at the physical layer [9]. Classical definitions of outage probability ( $P_{op}$ ) based on marginal statistics fail to capture the true characterization of the dynamism at the physical layer. It has been shown that higher-order statistics of the wireless channel errors affects the performance of the upper layers of the protocol stack. Thus a more general definition of outage probability which considers the *time dependencies* and durations of the unpredictable events is needed. To this end, the concept of *time-hysteresis outage probability* [17] is a more relevant performance metric for voice and data communications. We also feel that time-hysteresis outage probability is a good representation and captures the system performance with respect to QoS as well. Thus, we model the SUIF for class 1 on time-hysteresis outage probability as the metric for the system performance in our context. If  $\mathbf{x}_{1,j}$  denotes the random variable representing the SUIF for class 1 for the  $j^{th}$  user, then

$$\mathbf{x}_{1,j} = \tau \times P_{out,j} \quad (3)$$

where  $P_{out,j}$  is the time-hysteresis outage probability for the  $j^{th}$  user. Also  $SUIF_{max,V}$  is defined as the threshold SUIF crossing which the voice call is dropped.

**SUIF for Class 2:** We utilize *rate jitter* as the QoS parameter to model the SUIF for streaming traffic. Jitter, quantified in two ways – *delay jitter* and *rate jitter* is introduced due to variable queuing and propagation delays. Rate jitter [10] which measures the difference between the minimal and maximal inter-arrival packet times is more appropriate for streaming services than delay jitter. It actually bounds the difference in packet delivery rates during the entire period of service for that particular session and thus is an ideal metric for quantifying user irritation. The higher the rate jitter, the higher the user irritation and vice versa. Thus if

$\eta_{max,i,j}$ ,  $\eta_{min,i,j}$ ,  $\psi_{i,j}$  and  $\mathbf{x}_{2,j}$  respectively denote the maximum inter-arrival time, minimum inter-arrival, rate jitter and the random variable representing the SUIF for the streaming class respectively for the  $i^{th}$  session of the  $j^{th}$  user, then

$$\mathbf{x}_{2,j} = \tau \times \frac{\psi_{i,j}}{\eta_{max,i,j}} \quad (4)$$

where  $\psi_{i,j} = \eta_{max,i} - \eta_{min,i}$  is defined as the rate jitter. Again,  $SUIF_{max,S}$  is defined as the threshold beyond which the call is dropped.

**SUIF for Classes 3 and 4:** We define the SUIF for the interactive and background classes as the delay that an user is ready to endure before he decides to cancel his request is a measure of his irritation. Additional constraint specific to this class is that there should be absolutely *no* data loss. The ideal transfer time ( $\Delta$ ) for a file of size  $S$  Kbytes is  $\Delta = \frac{S}{BW}$ , where  $BW$  Kbps is the ideal bandwidth supported by the system for that user. However, due to congestion, an admitted new request might suffer a delay even before it is scheduled for service. The system can afford to assign a greater delay ( $\delta_{ext}$ ) to class 4 traffic than class 3 since the delay requirements for background traffic is much less strict than interactive. Let the available bandwidth to the new request be denoted by  $BW_r$ . Hence the actual delay  $\delta$ , suffered by the user is given by  $\delta = \delta_{ext} + \frac{S_{eff}}{BW_r}$  where  $BW_r$  is the bandwidth assigned in reality to the user and  $S_{eff}$  is the effective data size [11] that needs to be transmitted due to retransmission on account of frame error rate (FER).

The scheduler is designed to exploit the sensitivity of human nature to delay by transmitting the main page (for Web traffic) and some initial data for class 4 at the earliest possible time. Scheduling the intermediate packets on a regular basis will keep the user satisfied and also provide the scheduler more time to transmit the entire data. We assume that the maximum delay that any user would be ready to tolerate will be  $n$  (some multiple) times the ideal time needed to deliver the data, i.e.,  $n \times \Delta$ . Thus, if  $\delta > n\Delta$ , the request is serviced in the negotiated mode. We define the random variable  $\mathbf{x}_{(3,4),j}$  denoting the SUIF of the  $j^{th}$  user for class 3 and 4 traffic as

$$\mathbf{x}_{(3,4),j} = \tau \times \frac{\delta_{i,j} - (n-1)\Delta_{i,j}}{\Delta_{i,j}} \quad (5)$$

where  $\delta_{i,j}$  and  $\Delta_{i,j}$  are the ideal and actual delay for the  $i^{th}$  session of the  $j^{th}$  user. The worst case bounded delay is  $\delta = n\Delta$ . The corresponding SUIF is termed as  $SUIF_{max,BI}$ .

For scheduling purpose, we use the *stretch* metric [7], which measures the tolerance for the user, i.e., the delay in excess which the user can be made to suffer than what he is actually going to suffer without crossing the SUIF. We define stretch as the difference between the actual SUIF and  $SUIF_{max}$  for all the different traffic types.

#### IV.B. Long Term User Irritation Factor

LUIF is a quantitative measure of a user's tolerance to continuous degradation of the service provided, after which the user decides to cancel the service and churn out of the network. Thus, LUIF keeps track of the long term QoS being provided to each user. The service providers would be able to control the churn rate by judicious manipulation of the LUIF. Since LUIF is calculated on a per-user basis based on all the SUIFs perceived by that user, maintaining SUIF for each and every request for all users becomes memory and cost extensive. We argue that the QoS received in the distant past would have less significant impact (though not zero) on the users' overall long time irritation than what he received recently. An exponentially weighted moving average (EWMA) mechanism is used to maintain continuous measure of the SUIFs' for each user. Let the stored LUIF be  $U(\kappa_{n-1})$  and the LUIF to be computed be  $U(\kappa_n)$ ; the input to the system, i.e., the current SUIF be  $U(x_i)$ . The value of  $U(x_i)$  can be computed using any one of the following Equations (3), (4) and (5) depending on the type of request. Then  $\kappa_n$  is calculated as follows:

$$\kappa_n = \rho \times \kappa_{n-1} + (1 - \rho) \times U(x_i) \quad (6)$$

where  $\rho$  is the weightage given to the cumulative SUIF and  $\kappa_n$  denotes the random variable which is used to measure the LUIF at the  $n^{th}$  request using Equation (2). The value of  $\rho$  needs to be experimentally determined. However, in EWMA mechanisms,  $\rho = 0.2$  or  $0.3$  is generally chosen. We also define a threshold LUIF signifying that if the LUIF of a particular user exceeds that threshold value, he or she may cancel the service. This value basically determines the amount of churn in the system. We define another parameter, *tolerance factor*, which measures the amount of patience the particular user has to continue with the current service. Quantitatively, it would be the difference between the threshold LUIF and the actual LUIF. Thus, the call admission control algorithm must always take into consideration that the LUIF of any user does not exceed its threshold.

## V. Radio Resource Management

Let us now present the radio resource management scheme offering class-based QoS to heterogeneous services. The proposed scheme consists of two phases: the first phase is admission control and the second phase deals with bandwidth reservation and allocation to the admitted requests. The motivation of decoupling the process of admission control and dynamic bandwidth allocation lies in the premise that admission control is designed to monitor the LUIF of the users, whereas the second phase concerned with the SUIF determines whether guaranteed or negotiated mode of QoS will be provided. Thus, the QoS framework achieves differentiated service by offering different levels of satisfaction to different classes. Since the framework strives to adhere to the SUIF and LUIF of different classes, user satisfaction is maximized and hence, churn rate is controlled. The delay bound for a session is computed based on the type of traffic and user class.

The proposed admission control not only admits a higher number of requests by exploiting the delay tolerant nature of elastic traffic, but also endeavors to honor the LUIF of requesting users. The scheme is intelligent enough to judiciously choose and preempt users whose SUIF can be manipulated, or select in-session users who can be delayed/preempted without violating their SUIFs. At the time of a session establishment, the user application specifies its requirements in the form of *Service Request Tuple*. This tuple differs for each class of traffic and is summarized below:

#### Conversational Class:

$\langle New/Handoff, BW \text{ Required} \rangle$

#### Streaming Class:

$\langle New/Handoff, Min/Max \text{ BW}, Size \rangle$

#### Interactive Class:

$\langle Size \rangle$

#### Background Class:

$\langle Size \rangle$

Although the duration for voice calls is not known priori for the other request types the ideal delay is computed for making admission decisions. For elastic traffic, in scenarios when the delay bound is violated, the requests are served in negotiated mode with priority given to higher class users. The admission control algorithm is illustrated in Figure 3.

Once a request is admitted as per the admission control algorithm, the QoS framework thereafter allocates or reserves bandwidth for that session. We do not allow the entire bandwidth to be accessible to all

### Admission Control Algorithm

```
1: Identify Class, Traffic type of Request
2: Compute SUIF, Retrieve LUIF
3: if Bandwidth Available then
4:   Admit Call
5: else
6:   if Voice/Streaming Request then
7:     Preempt Active Interactive/Background Sessions
8:   if preempted sessions SUIF violated then
9:     Compare LUIFs of requesting and in-session call
10:    if LUIF for requesting call greater then
11:      Drop Call, Update LUIF
12:    else
13:      Admit Call, preempt in-session call
14:    end if
15:  else
16:    Admit Call, preempt in-session call
17:  end if
18: else
19:   if  $D_r \geq n \times D_i$  then
20:     Admit Call, guaranteed QoS
21:   else
22:     Compare LUIFs of requesting and in-session call
23:     //only Background sessions
24:     if LUIF for requesting call greater then
25:       Admit Call, negotiated QoS
26:     else
27:       Admit Call, preempt in-session call
28:     end if
29:   end if
30: end if
31: end if
```

Figure 3: Service admission control algorithm

the classes. This is done to allow Gold class customers to have their deadlines met. Usually, in case of non-availability of bandwidth, new requests are simply dropped, leading to higher blocking probability. To prevent this situation, a hybrid mechanism involving both preemption and reservation is adopted. We term this as pseudo preemptive service or soft bandwidth reservation mechanism. The mechanism is inspired due to the following two reasons. Only preemption of lower class clients leads to longer delay for them at the cost of lower blocking probability of higher class clients. Whereas, in case of exclusive reservation of the bandwidth for Gold class, the system utilization is low since the reserved bandwidth might be idle at

### Bandwidth Allocation Algorithm

```
1: if Bandwidth Available then
2:   Admit Call
3: else
4:   if Gold Class then
5:     if Reserved Bandwidth Available then
6:       Serve Call
7:     else
8:       if Lower Class occupies Reserved Bandwidth then
9:         preempt lower class
10:        Serve Call
11:       else
12:        Drop Call, Update LUIF
13:       end if
14:     end if
15:   else
16:     if Reserved Bandwidth Available then
17:       Serve Call
18:       Preempt for Gold class
19:     else
20:       Drop Call, Update LUIF
21:     end if
22:   end if
23: end if
```

Figure 4: Bandwidth allocation algorithm

times. Hence the hybrid approach. We assume a certain fraction of the entire available bandwidth is reserved for the Gold class. The reservation scheme can be also extended for Silver class users, but of course the reserved bandwidth should be less than the Gold class. If the available bandwidth is used up by the currently admitted clients, then a lower class client is admitted using the reserved bandwidth. But of course, the arrival of a higher class request will lead to the preemption of the lower class client from the reserved bandwidth. The detailed bandwidth allocation and reservation algorithm is given in Figure 4.

The proposed framework enforces a probabilistic bound on the delay which varies according to the user class. Fairness on the basis of the revenue paid is enforced since the resource allocation is performed on the basis of the user satisfaction as modeled in section IV. The lower class users do not suffer from the starvation since the admission control scheme takes care of the LUIF of each user. When a request for a user is dropped repetitively, the LUIF of that particular user is updated by a factor greater than its SUIF. This enforces the LUIF to increase towards its thresh-

old value and the scheduler then allocates resources so as to prevent the LUIF reaching the threshold. The algorithm takes care of the channel transmission error by considering the effective data size to be transmitted.

## VI. Simulation Model and Results

To validate the proposed two-level resource management scheme, we conducted extensive simulation experiments. We evaluate the performance of the proposed class-based QoS framework for heterogeneous wireless services. For simplicity, the simulation considers only a single cell. It was assumed that all 300 users subscribe to heterogeneous services. At any instant, the base station could provide 200 Kbps bandwidth for data services (streaming, interactive, background). The bandwidth reserved for Gold users was set to 10%. Depending on the class of the user, for data services the bandwidth was allocated according to the ratio 4 : 2 : 1. Gold, Silver and Bronze users were assigned 38.4 Kbps, 19.2 Kbps and 9.6 Kbps, respectively. Also, the value of  $\alpha$ , the parameter which distinguishes the user class is set to 0.1, 0.3 and 0.9 for Gold, Silver and Bronze users, respectively. For varying system loads, we measured the *blocking probability* for voice calls and streaming requests, *average rate jitter* for streaming requests, *average bounded* and *negotiated delay* for interactive data and background traffic. The effect of frame loss at the link layer and the traffic models under consideration have been described next.

### VI.A. FER modeling

Here we describe the modeling of the FER and its impact on our scheduling algorithm. The dynamically varying capacity of the wireless channel should be sensed by the scheduler and thereafter to take appropriate actions. Although in real life, the packet size varies (if we assume TCP segments), in our design we will assume that the packet size handled by the scheduler is of constant size. This is made possible using the radio link protocol (RLP) which would fragment a transport layer segment into equal size RLP frames [1]. In order to perform scheduling, the channels are modeled as common pool of bandwidth available for sharing. However, the scheduler must capture the different channel quality (or FER) of each user and the corresponding perceived bandwidth. Due to the nature of wireless medium, packets get lost or damaged, resulting in retransmissions which lead to the increased bandwidth demand. The effective data that needs to

be transmitted is given by  $S_{eff} = S/(1 - p)$ , where  $S$  is the original data size and  $p$  is the packet loss rate of the channel. (This equality holds true if we do not restrict the number of possible retransmissions.) If we denote the FER by  $p$ , then for transmitting  $n$  packets, the expected number of retransmissions would be  $pn$ . But for the  $pn$  retransmitted packets,  $p^2n$  more packets are expected to undergo loss or corruption. Thus, due to the successive nature of the retransmissions, the expected number of packets to be transmitted in a recursive manner is

$$(1 + p + p^2 + p^3 + \dots)n = \frac{n}{1 - p} \quad (7)$$

since  $p < 1$ . If the original data size be  $S$ , the effective size becomes  $S_{eff} = S/(1 - p)$ . However, for practical systems, the number of retransmissions allowed is finite (usually 3 as reported in [4]). Thus Equation (7) can be appropriately modified by considering the required number of finite terms.

### VI.B. Traffic Models

The voice calls are modeled as Poisson process where the inter-arrival time and call holding time being modeled as negative exponential distribution. We modeled the streaming traffic (video) as an *on-off* traffic source, where the “on” time and the “off” time were exponentially distributed with mean value of 30 seconds and 120 seconds, respectively. The traffic for the interactive class was modeled as HTTP [5]. Instead of investigating the nature of HTTP traffic, we synthetically generate such traffic by using the results obtained in [5]. The basic model of HTTP is shown in Figure 5 in which a packet call represents the download of a web page requested by a user. It usually has a main page followed by some embedded objects. A new request (packet call) is immediately generated after the expiration of the viewing period. The model is similar to an ON/OFF source, where the ON state represents the activity of a page request and the OFF

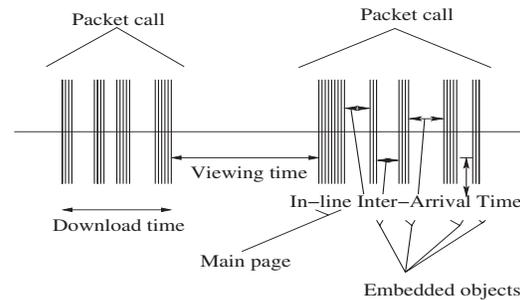


Figure 5: Web page traffic scenario

Table 1: Statistics for HTTP

Component	Distribution	Mean
Main page size	Lognormal	10710 bytes
Embedded object size	Lognormal	7758 bytes
No. of embedded objects	Pareto	5.55
Viewing time	Weibull	40 ms

state represents a silent period after all objects in that page are retrieved. The download time of a page follows Weibull distribution, the mean of which depends on the underlying bandwidth of the wireless channel. Each object (main page and embedded objects) of the HTTP traffic is fragmented into multiple equal-sized frames so as to fit into a packet. Other statistics and parameters used to generate the HTTP traffic are shown in Table 1. The FTP requests are similar to the web traffic but has only one embedded object and the packet size modeled as a Pareto distribution having different scale and shape parameters.

### VI.C. Simulation Results

The blocking and dropping probability for voice calls belonging to all 3 user classes are shown in Figures 6

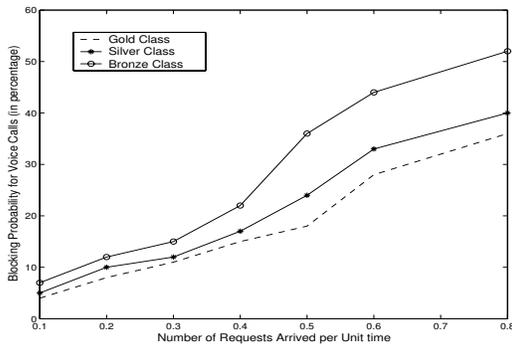


Figure 6: Block probability for voice Calls

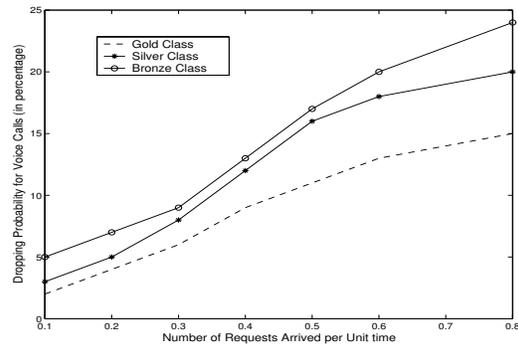


Figure 7: Dropping probability for voice calls

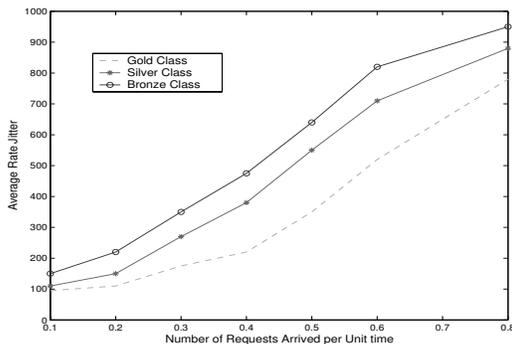


Figure 8: Average rate jitter for streaming class

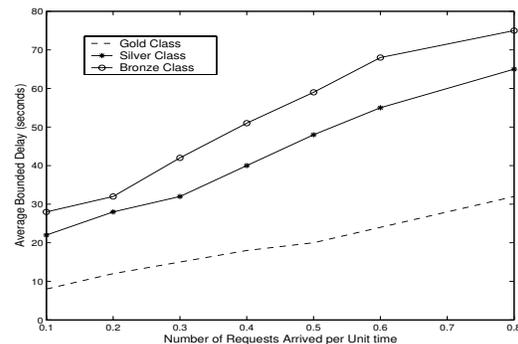


Figure 9: Average bounded delay for elastic traffic

and 7, respectively. Simulation results prove that the QoS received for Gold class is the best followed by Silver and Bronze classes. Thus, the notion of fairness based on revenue paid is adhered to. The rate jitter shown in Figure 8 was averaged for the particular class of users so as to measure the average rate jitter for each class. Here also, the Gold class suffers from the least jitter since the bandwidth assigned is maximum for these users. The bounded delay suffered by the background traffic for each of the classes is shown in Figure 9. The delay suffered in negotiated QoS by the users when the scheduler optimistically admits the user, is presented in Figure 10. It is to be noted that the delay suffered during negotiation is higher than when served in guaranteed QoS. The difference between the delays for the Bronze and higher classes is much higher in the negotiated mode since the background class is penalized more when the system is more overloaded. The Bronze class suffers the maximum delay whereas the Gold Class suffers the least, which proves the validity of the proposed class-based QoS framework .

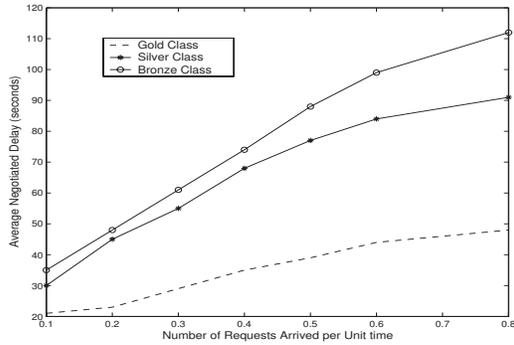


Figure 10: Average negotiated delay for elastic traffic

## VII. Conclusions

In this paper, a QoS framework for both traditional voice communications and rapidly emerging data services has been proposed. The framework is based both on the traffic class and the user satisfaction. Though non-real time wireless data services have elastic requirements, the basic form of QoS or SLA is still non-existent. We provided an insight to what might be the possible QoS parameters for these data services. Variation of user satisfaction to the different services received has been modeled which forms the basis for the QoS framework. The proposed radio resource management framework comprises call admission control algorithm and bandwidth allocation policies. The admission control algorithm admits sessions based on not only the resource requirements but also the irritation level of the requesting user. In addition, the framework judiciously penalizes users such that the user irritation factors remain bounded. Soft reservation schemes have been proposed to guarantee QoS for higher class users. Simulation results prove that the framework successfully exploits the flexibility in user tolerance with respect to the perceived QoS and provides assurance for non-real time applications. Thus, the results reveal that the proposed algorithms provide bounded delay guarantees and acceptable call blocking and dropping probabilities.

## References

- [1] 3GPP TR 25.858 V5.0.0, "High Speed Downlink Packet Access: Physical Layer Aspects (Release5)," Mar. 2002.
- [2] 3GPP2 C.S0024 Ver 3.0, "cdma2000 High Rate Packet Data Air Interface Spec." Dec. 5, 2001.
- [3] L. Badia, M. Lindström, J. Zander, M. Zorzi, "Demand and Pricing Effects on the Radio Resource Allocation of Multimedia Communication Systems", *Proceedings IEEE Globecom 2003*, San Francisco, CA, vol. 7, pp. 4116–4121, Dec. 2003.
- [4] G. Bao, "Performance evaluation of TCP/RLP protocol stack over CDMA wireless links", *ACM Wireless Networks Journal*, Vol. 2, 1996, pp. 229-237.
- [5] H.K. Choi and J.O. Limb, "A Behavioral Model of Web Traffic", *International Conf. of Network Protocols (ICNP)*, 1999, pp.327-334.
- [6] M. Fine, K. McCloghrie, J. Seligson, K. Chan, S. Hahn, R. Sahita, A. Smith and F. Reichmeyer, *Framework Policy Information Base*, IETF-rap-framework/pib-04.
- [7] M. Bender, S. Chakrabarti and S. Muthukrishnan, "Flow and stretch metrics for scheduling continuous job streams," *Proc. ACM Symposium on Discrete Algorithms (SODA)*, 1998, pp. 270-279.
- [8] H. Lin, M. Chatterjee, S. K. Das and K. Basu, "ARC: An Integrated Admission and Rate Control Framework for CDMA Data Networks Based on Non-Cooperative Games," *Intl. Conference on Mobile Computing and Networking (MobiCom)*, 2003.
- [9] N.B. Mandayam, P.-C. Chen and J.M. Holtzman, Minimum duration outage for cellular systems: a level crossing analysis, *Proc. of IEEE VTC*, 1996, pp. 879-883.
- [10] Y. Mansour, B. Patt-Shamir, "Jitter Control in QoS Networks," *IEEE/ACM Transactions on Networking*, Vol. 9, No. 4, August 2001, 1998.
- [11] S. Pal, M. Chatterjee and S. K. Das, "Improving Guarantees on Delivery Time for Wireless Data Services," *IEEE WCNC*, Volume: 4, pp. 2539-2544, 2004.
- [12] S. Pal, M. Chatterjee and S. K. Das, "User-Satisfaction based Differentiated Services for Wireless Data Networks," *IEEE International Conference on Communications (ICC)*, May 2005.
- [13] Y. Snir, Y. Ramberg, J. Strassner, R. Cohen, *Policy Framework QoS Information Model*, IETF-policy-qos-info-model-02.
- [14] G.D. Stamoulis, D. Kalopsikakis and A. Kyriakoglou, "Efficient agent-based negotiation for telecommunications services", *Global Telecommunications Conference (GLOBECOM)*, Vol 3. pp. 1989-1996, 1999.
- [15] M. Xiao, N.B. Shroff, E.K.P Chong, "Utility-based power control in cellular wireless systems", *Proceedings of IEEE INFOCOM 2001*, Vol.1, pp. 412-421.

- [16] D. Zhao, X. Shen and J. W. Mark, "Radio Resource Management for Cellular CDMA Systems Supporting Heterogeneous Services", *IEEE Transactions on Mobile Computing*, Vol. 2, No. 2, April-June 2003.
- [17] M. Zorzi, "Outage and error events in bursty channels", *IEEE Transactions on Communications*, Vol. 46 No. 3, pp. 349-356, Mar 1998.