Call Admission Control Scheme With Multimedia Scheduling Service in WiMAX Networks

Zhiwei Yan *zhiweiyan@mailst.xjtu.edu.cn*
The School of Electronics and Information Engineering
Xi'an Jiaotong University
Xi'an 710049, P. R. China

Guizhong Liu *liugz@mail.xjtu.edu.cn*
The School of Electronics and Information Engineering
Xi'an Jiaotong University
Xi'an 710049, P. R. China

Rui Su *ssurui@gmail.com*
Xi'an Satellite Control Center
Xi'an 710049, P. R. China

Abstract

WiMAX network introduces a multimedia data scheduling service to handle the different quality of service (QoS) requirements. The service schedules the opportunities of data transmission according to the types of traffic for the connections or users. In this paper, based on the data types defined in the service, a uniform definition of QoS level is proposed for multimedia connections. The QoS level of individual connection is only determined by the type of data of the connection and its allocated resources. By means of allocating the bandwidth resources and calculating the QoS levels, we can make a decision of call admission control (CAC) for the entry admission of a new connection, without degrading the network performance and the QoS of ongoing connections. The contribution of this scheme is to regulate the traffic of ongoing connections so that the network can work at an optimal point, especially under heavy load traffic. Simulation experiments confirm that the proposed CAC scheme can achieve better trade-off between the overall performance of network system and the QoS requirements of individual connection.

Keywords: Call Admission Control, QoS level, Multimedia, Wireless Network, WiMAX.

1. INTRODUCTION

Call admission control (CAC) is a technology that it manages call entry requests or regulates traffic volume loaded on a network. The technology is used typically in wireless voice communications or in VoIP (voice over Internet Protocol, also known as Internet telephony) networks [1] [2]. It is also used to ensure, or maintain a certain level of quality in video communications networks [3] [4] [5]. Generally, it is also considered as a methodology or a trade-off policy between limited resource supply and a larger amount of resource demand. For example, in a telephone service network, a typical real-time delay-sensitive application, call admission control will take such strategy that it can deny much more network access, not allowing traffic onto the network if the network condition is under congestion. If CAC strategy is not taken, traffic congestion of this network will deteriorate. In addition, call admission control influences the allocation of network resources: if an influx of data traffic is greedy for oversubscribing link resources in the network, the excessive network traffic of a connection will impair the quality of service of other connections. The worse thing is to bring about many connections or users who are blocked or dropped and to result in the dissatisfaction with the service among users.

In the past decades, most of CAC algorithms fall into three broad categories according to call entry management disciplines. The first category of algorithms are implemented by regulating the basic network parameters. These parameters include the total utilized bandwidth, the total number of calls, or the total number of packets or data bits, which pass a specific point per unit time. If the values of the parameters are reached or exceeded the pre-defined top or bottom threshold, new calls are likely to be prohibited from or postponed entering the network until at least one current call leaves or terminates [6] [7].

The second category of algorithms builds their discipline of service threshold through combining the utility of lower network layers (MAC and Physical Layers) together. For example, in [8], Song and Zhuang introduced a service discipline called the shortest remaining processing time. The size of this time is regulated to an appropriate threshold, which is used to judge which connection was dropped. Yilmaz and Chen did the roughly similar works [9]. They built the threshold of price for radio resource. They first made a pricing scheme by means of analyzing the features of CAC in single and multiple-class traffic, and then developed a CAC algorithm with hybrid partitioning thresholds by the pricing scheme. The scheme periodically determines an optimum price of resources in order to maximize the system revenue. Those connections or users who could not afford their required resource would be denied or dropped. In [10], [11], the problem of whether or not to accept an incoming user is transformed into a mathematical problem, which achieves the maximum resource utilization by means of establishing and analyzing a complex model of a semi-Markov decision process. Zhai and Xiang regulated the arriving traffic of the IEEE802.11 network such that the network can work at an optimal point through measuring channel busyness ratio as an indicator of the network status [12].

The third category of algorithms is called resource reservation algorithms. In the algorithms, researchers made a reservation of radio resources for new incoming connections. This topic has been extensively examined in the last decade. For example, in [13], [14], preferential admission is given to high priority calls by preserving an interference guard margin. The guard margin is is dynamically adjusted by referencing traffic conditions in neighboring cells based upon users' quality of service. El-Kadi presented such a reservation method by borrowing bandwidth from other calls in a cell for an incoming call [15]. The amount of bandwidth borrowed from ongoing calls is proportional to its tolerance to bandwidth loss in the scheme. In [16], Elsayed evaluated the three resource reservation policies: the uniform reservation policy, the capacity proportional policy, and the remaining capacity proportional policy. The policies are used to map the end-to-end delay requirement into a local rate to be reserved at each switch.

However, these CAC algorithms above consider less about the features of connections from the view of application layer. Few literatures address this feature in CAC schemes. Therefore, we construct a uniform QoS metric of connection in accordance with the types of multimedia applications and its allocated resources of connections in the paper. And then using the metric, we propose a CAC scheme to make the network perform at an optimal point.

The rest of this paper is organized as follows: Section 2 introduces the five types of application traffic defined in WiMAX scheduling services. Then we develop a uniform metric of QoS level for the different traffic. Especially, we describe in detail the QoS of video application based on an idea of rate distortion optimization (RDO) model in Section 3. Next, we prove mathematically that there is an optimum solution of bandwidth allocation if the bandwidth of connections can be adjusted dynamically in Section 4. In Section5, CAC algorithm is proposed to determine according to the optimal solution of bandwidth allocation of connections. We report the results of simulation experiments in Section 6. Finally, we conclude the paper in Section 7.

2. WIMAX SCHEDULING SERVICES

One of the key functions of the WiMAX (Worldwide Interoperability for Microwave Access) MAC layer is to ensure that QoS requirements for MAC packet data units are met as reliably as possible given the loading conditions of the system. These packet data belong to the five different service flows. This implies that various negotiated performance indicators that are tied to the overall QoS, such as latency, jitter, data rate, packet error rate, and system availability, must be met for each connection [17] [18]. The WiMAX MAC layer designs a scheme of scheduling service to deliver and handle MAC packet data with different QoS requirements. A scheduling service uniquely determines the mechanism the network uses to allocate upload and download transmission opportunities for the packet data. WiMAX standard define five scheduling services.

• The unsolicited grant service (UGS) is designed to support real-time service flows. The flow data include fixed-size data packets with a periodic basis, such as T1/E1 and VoIP. So UGS offers fixed-size grants on a real-time periodic basis. The features of traffic implies that the resources the UGS requests are constant during the whole communication.

• The real-time polling services (rtPS) is designed to support real-time services that generate variable- size data packets on a periodic basis, such as MPEG (Motion Pictures Experts Group) or H.264 video. In this service class, the base station provides unicast polling opportunities for the MS (Mobile Station) to request bandwidth. The unicast polling opportunities are frequent enough to ensure that latency requirements of real-time services are met. This service requires more request overhead than UGS does but is more efficient for service that generates variable-size data packets or has a duty cycle less than 100 percent.

• The non-real-time polling services (nrtPS) is very similar to rtPS except that the MS can also use contention-based polling in the uplink to request bandwidth. In nrtPS, it is allowable to have unicast polling opportunities, but the average duration between two such opportunities is in the order of few seconds, which is large compared to rtPS. All the MSs belonging to the group can also request resources during the contention-based polling opportunity, which can often result in collisions and additional attempts.

• The best-effort service (BE) provides very little QoS support and is applicable only for services that do not have strict QoS requirements. Data is sent whenever resources are available and not required by any other scheduling-service classes. The MS uses only the contention-based polling opportunity to request bandwidth.

• The extended real-time polling service (ertPS), a new scheduling service introduced with the IEEE 802.16e standard, builds on the efficiencies of UGS and rtPS. In this case, periodic upload allocations provided for a particular MS can be used either for data transmission or for requesting additional bandwidth. This features allow ertPS to accommodate data services whose bandwidth requirements change with time. Note that in the case of UGS, unlike ertPS, the MS is allowed to request additional bandwidth during the upload allocation for only non-UGS-related connections.

3. CAC SCHEDULER MODEL, SERVICE FLOW AND QOS OPERATIONS

Since the QoS requirements of different data services can vary greatly, WiMAX needs various handling and transporting mechanisms to meet that variety. The scheduling mechanisms of the services, however, are not implemented in WiMAX standards. The WiMAX standard explains the reason that the scheduling mechanisms or algorithms will be implemented independently by various vendors. In the paper, our work focuses on the issue. The key idea is that we combine the data scheduling mechanism with call admission control mechanism to obtain the more effective results than before.

3.1. CAC Scheduler Model

Just as we said above, CAC methodology is to determine which connection or user ought to be dropped or blocked, averting network congestion. In the paper, we consider a general CAC model of WiMAX network, which consists of the K connections in one base station (BS), illustrated in Figure 1. These connections include $K - 1$ ongoing connections and one new incoming connection. All of these connections fall into five different application traffic categories, belonging to the corresponding WiMAX services. Before the K th connection is admitted into the BS, all $(K - 1)$ connections share all available bandwidth by provided the BS, denoted by B_{total}. The CAC process is triggered by the K th connection when a mobile station sends its new connection request and bandwidth allocation request. A vector or scheme of bandwidth allocation b is provided by the our proposed QoS level evaluator, where b_i denotes the assigned bandwidth for the i th connection, $b = [b_1; b_2; b_3; \ldots; b_K]$ and $\sum_{i=1}^{K} b_i \le B_{total}$. Based on the allocated bandwidth b_i of every connection, the CAC scheduler examines the QoS level of each connection and the whole QoS of the BS. After that, the CAC scheduler makes the decision whether or not the K th connection will be blocked.

3.2. QoS Metric And Bandwidth Allocation

The traditional QoS metrics of connections are developed by most of the existed CAC algorithms. They are directly taken from the measurement of MAC layer or physical layer, such as bandwidth utilization ratio, drop call rate and so on. Our CAC method emphasizes the QoS metric from the view of application layer for the connections, which depends on from the bandwidth allocation scheme **b** and the characteristic of applications.

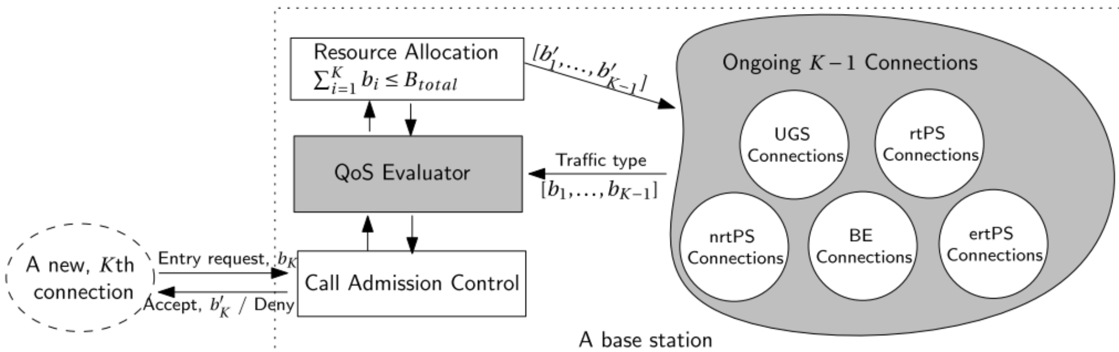

FIGURE 1: System model of CAC in WiMAX networks

3.2.1. QoS Metric of UGS Traffic.

Pkt-Loss(%):	0	0.1	0.3	0.5	0.7	0.9	1	2
R-Factor:	93.2	84.6	71.3	61.5	54.1	48.2	45.2	29.9

TABLE 1: R-Factor vs. Packet Loss

For VoIP application as a typical UGS traffic, the E- Model provides an objective method of assessing the transmission quality of a telephone connection. The model defines an analytic model for prediction of voice quality based on network impairment parameters, such as packet loss and delay [19]. The E-Model results in an R- factor ranging from a best case of 100 to a worst case of 0. The R- factor uniquely is reduced into another well-known metric: the Mean Opinion Score (MOS), which provides a numerical indication of the perceived quality of received media after compression and/or transmission. The number is in a scale of 1 to 5, where 1 is the lowest perceived quality, and 5 is the highest perceived quality [20].

The R-factor can be obtained by the following expression:

$$R = R_0 - I_s - I_d - I_e + A \qquad (1)$$

where R_0 represents the basic signal-to-noise ratio. I_s represents the combination of all impairments which occur more or less simultaneously with voice signal; I_d stands for the impairments caused by delay. I_e denotes impairments caused by low bit rate codecs and A is the advantage factor, that corresponds to the user allowance due to the convenience in using a given technology. Eqn.(1) can be reduced to:

$$R = 93.4 - I_d(T_a) - I_e(codec, loss) \qquad (2)$$

where I_d is a function of the absolute one-way delay and I_e rests on the sound codec type and the packet loss rate. (Here, we assume the delay is very small and can be omitted). We can obtain the results of voice QoS level for different packet loss rate by the E-Model calculator [21]. The results, in Table I, suggest that the voice applications hardly tolerate packet loss because R-Factor $<$ 50 means poor quality (Nearly all users dissatisfied). Therefore, we have the QoS metric of UGS(voice) traffic, defined by Eqn.(3)

$$\alpha^{UGS} = \begin{cases} 1 & \text{if } b = B, \\ 0 & \text{others} \end{cases} \qquad (3)$$

where α^{UGS} represents the QoS level of UGS traffic. b is the allocated bandwidth and B is the required bandwidth for voice applications. Furthermore, we express the metric by the Dirac measure , shown in Eqn(4).

$$\alpha^{UGS} = \delta_b(B) = \begin{cases} 1 & \text{if } b = B, \\ 0 & \text{others} \end{cases} \qquad (4)$$

3.2.2. QoS Metric of rtPS Traffic.

Video traffic is a typical rtPS application. In video applications, the most popular objective criterion for image quality is peak signal-to-noise ratio (PSNR), which measures the difference between the corrupted video image K and original video one I , illustrated in Eqn.(5).

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$

$$MSE = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}[I(i,j) - K(i,j)]^2 \qquad (5)$$

where MAX_I is the maximum possible pixel value of the image. When the pixels are represented using 8 bits per sample, this is 255. More generally, when samples are represented using linear PCM with z bits per sample, MAX_I is $2^z - 1$. The mean squared error (MSE) which for two $M \times N$ monochrome images I and K. The image I means the reference image and the image K is a corrupted image.

The PSNR, however, is not directly applied in CAC processes for the measurement of QoS level of connections, because the value of the PSNR of a video image can be calculated correctly only after the video packets must be decoded into the YUV images in the application layer. Moreover, the reference video images are hardly obtained by CAC schedulers in BS in advance. What we have in CAC scheduler is bandwidth requirement B from the application layer and bandwidth allocation b provided by the CAC scheduler. Thus, similar to the ratio of network throughput, the ratio of the allocated bandwidth b to the required bandwidth B sometimes is used to evaluate the quality of connection. Unfortunately, it is also not suitable for the video traffic, because video quality is not linear relationship with the ratio [22] . We should seek out another alternative measurement solution, which is closely akin to the PSNR and particularly suitable for the CAC process.

We find that we can utilize the idea of rate distortion optimization (RDO) in our case. RDO is a method of improving video quality in image compression and video encoding[23][24][25]. It aims to establish the relationship between the amount of distortions (loss of video quality) and the amount of data required to encode the video. The RDO method bridges between the video QoS measurement and the bandwidth allocation in the CAC problem. In order to find a suitable metric for video traffic, we investigated some typical fitting models for the relationship between encoding rate and quality of video.

The Table II shows the results of fitting operations as an example, for a video sequence called "Forman". We choose the last two fitting modes (labeled by 7 and 8) as the candidates of QoS evaluation. If we had considered only mathematical results not practical conditions, we might have chosen the model $p_1x^4 + p_2x^3 + p_3x^2 + p_4x + p_5$ just according to the goodness of fit (SSE and Adjusted R-square). Here, we choose the fitting model (labeled by 7), $a(1 - e^{\rho \frac{x}{max(x)}})$. There are three reasons for the choice: First, from the view of complexity, the polynomial model contains five parameters and is more complex than the exponential one. What we just need is a simple and robust QoS metric in the CAC process. Second, the exponential model also denotes the relationship better than other six models and make a good balance between the simplicity and accuracy. Third, the model also cover the other type of traffic, detail in the later Section 3.3.

We normalized the relationship and proposed a formula shown in Eqn(6), which denotes the relationship between quality of video or rtPS application α and allocated bandwidth b for the rtPS connection.

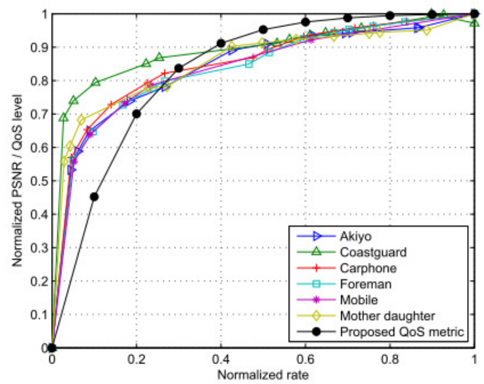

FIGURE2: Comparison of the proposed metric of QoS level ($\rho = 4$) and the normalized PSNR of video clips, which are Akiyo, Coastguard, Carphone, Foreman, Mobile, and Mother &daughter.

Id	Fitting model: $f(x)$	SSE	Adjusted R-square
1	ae^{bx}	594.8	0.4233
2	$p_1x + p_2$	505.7	0.5096
3	$a_1e^{-((x-b1)/c1)^2}$	339.1	0.6243
4	$a_0 + a_1\cos(xw) + b_1\sin(xw)$	217.5	0.7317
5	$p_1x^2 + p_2x + p_3$	207.5	0.7701
6	$p_1x^3 + p_2x^2 + p_3x + p_4$	62.04	0.9465
7	$a(1 - e^{\rho \frac{x}{\max(x)}})$	19.83	0.9808
8	$p_1x^4 + p_2x^3 + p_3x^2 + p_4x + p_5$	14.93	0.9871

TABLE 2: Fitting Model Types

$$\alpha^{rtPS} = \frac{PSNR_b}{PSNR_B} = \frac{1 - e^{-\rho\frac{b}{B}}}{1 - e^{-\rho}}; \rho > 0 \quad (6)$$

where $PSNR_b$ and $PSNR_B$ are the qualities of video with the encoding rate b and B, respectively. α^{rtPS} indicates the normalized QoS level of a rpPS or video stream. ρ stands for the characteristic of the rtPS or video connection. Every rtPS connection has its unique value ρ, which can be determined before data transmission. $e^{-\rho\frac{b}{B}}$ denotes the distortion, $1 - e^{-\rho\frac{b}{B}}$ denotes the QoS level of the rtPS stream and the denominator $1 - e^{-\rho}$ is introduced for normalization. The b denotes the amount of allocated bandwidth of the connection, which is larger than the minimum bandwidth B_{min}, and B

denotes the amount of desired bandwidth of the connection. In order to further validate the formula, we encode the standard video sequences with the different traffic rate and compare the results with the proposed metric with $\rho = 4$. The results are shown in Fig.3. As expected, the proposed metric is a proper approximation for the QoS of video streams.

3.2.3. QoS Metric of Other Traffic.

For the BE traffic, FTP and HTTP applications are popular examples. File Transfer Protocol (FTP) is a standard network protocol used to transfer files from one host to another host over a TCP-based network. FTP is usually built on a client-server architecture and uses separate control and data connections between the client and the server [25]. For instance, we use it to upload web pages and other documents from a private development machine to a public web-hosting server. The Hypertext Transfer Protocol (HTTP) is an application protocol for distributed, collaborative, hypermedia information systems. HTTP is the foundation of data communication for the World Wide Web [26]. The same point for the two applications is that their quality just depends on the ratio between the volume of data received by the receiver end and the volume of data sent by the sender end. Therefore, we can simply define the QoS metric, shown in Eqn.(7).

$$\alpha^{others} = \frac{b}{B} \tag{7}$$

where b is the allocated bandwidth and B is the required bandwidth for BE applications. We think the ratio represents the QoS of BE traffic. For the last two type of traffic, nrtPS and ertPS, the relationship between their data and QoS level are also similar to the BE applications. We use the Eqn.(7) to denote it too.

3.2.4 Uniform QoS Metric With Allocated Bandwidth Resources

Because the forms of QoS level metric of the five services traffic look like different, we summarize them into one form. For BE, nrtPS or ertPS, we notice that the limit of Eqn.(6) when ρ approaches 0, illustrated in Eqn(8).

$$\lim_{\rho \to 0} \frac{1 - e^{-\rho \frac{b}{B}}}{1 - e^{-\rho}} = \frac{b}{B} \tag{8}$$

Therefore, we rewrite the QoS metric of Eqn(9) as

$$\alpha^{others} = \frac{b}{B} = \frac{1 - e^{-\rho \frac{b}{B}}}{1 - e^{-\rho}} \quad \text{if} \quad \rho \to 0 \tag{9}$$

Finally, we sum up the equations of QoS level for all traffic and we have

$$\alpha = \begin{cases} \delta_b(B); & \text{for UGS traffic} \\ \frac{1 - e^{-\rho \frac{b}{B}}}{1 - e^{-\rho}}; \rho > 0 & \text{for rtPS traffic} \\ \frac{1 - e^{-\rho \frac{b}{B}}}{1 - e^{-\rho}}; \rho > 0; \rho \to 0; & \text{for BE, nrtPS, ertPS traffic} \end{cases} \tag{10}$$

4. OPTIMUM RESOURCE ALLOCATION SOLUTION

The target of CAC schemes is to maximize the QoS level demand of all connections as it can as possible. At the same time, bandwidth resource can be utilized efficiently. We formulate the target as an optimum issue about the sum of QoS level of all connections. We assume that there are M UGS connections, N rtPS connections and Q other connections in the base station. These connections are independent with each other. The sum of QoS levels of all the $K = M + N + Q$ connections in the BS, is formulated by Eqn.(11).

$$
\begin{aligned}
U_{BS} &= \sum_{i=1}^{K} \alpha_i \\
&= \sum_{i=1}^{M} \alpha_i^{UGS} + \sum_{i=1}^{N} \alpha_i^{rtPS} + \sum_{i=1}^{Q} \alpha_i^{others} \\
&= M' + \sum_{i=1}^{N} \left(\frac{1 - e^{-\rho_r \frac{b_i}{B_i}}}{1 - e^{-\rho_r}} \right) \\
&\quad + \sum_{i=1}^{Q} \left(\frac{1 - e^{-\rho_o \frac{b_i}{B_i}}}{1 - e^{-\rho_o}} \right); \qquad (11)
\end{aligned}
$$

$M'(\leq M)$ stands for the M' UGS connections are allocated their required bandwidth. ρ_r denotes the coefficient of the ith rtPS connection. ρ_o is the coefficient of the BE traffic whose value approaches zero. We will use the uniform symbol ρ to replace the ρ_r and ρ_o in the following sections because their QoS level metrics have the same form. U_{BS} also denotes the utility of the whole BS when it provides services for the K connections. In order to maximize U_{BS}, we should find an optimum solution of bandwidth allocation b for the connections. The solution will affect the judgment of CAC scheduler. The optimum problem is defined by

$$
\begin{aligned}
b^* &= \arg\max \{ U_{BS}(b) \mid b = [b_1; b_2; \ldots; b_K] \\
&\quad \text{subject to:} \\
&\quad \sum_{i=1}^{K} b_i + B_{ava} = B_{total} \qquad (12)
\end{aligned}
$$

where b^* is an optimum solution of bandwidth allocation, and B_{total} is the total amount of bandwidth resources provide by the BS and B_{ava} is the amount of available bandwidth.

Here, we make an assumption for the sake of simplicity that all UGS connections can be allocated their desired bandwidth. Meanwhile, if we only consider a heavy traffic scenario, the value of the B_{ava} is zero. Thus, we have the final formulated equation for the optimum problem, Eqn.(13).

$$
\begin{aligned}
b^* &= \arg\max \{ U_{BS}(b) \mid b = [B_1; B_2; \ldots; B_M; \\
&\qquad b_{M+1}; \ldots; b_K] \} \\
&\quad \text{subject to:} \\
&\quad \sum_{i=1}^{M} B_i + \sum_{i=M+1}^{K} b_i = B_{total} \qquad (13)
\end{aligned}
$$

Theorem:

There is a unique optimum solution of bandwidth allocation b^* to make the U_{BS} reach up to its maximum, if the optimum problem is formulated as Eqn.(13).

Proof: For convenience in the following proof procedures, we introduce a temporary variable $x_i \in [0; 1]$.

$$
x_i = \frac{b_i}{B_i}; i \in [M+1; K]:
$$

To prove the theorem, we start with introducing a new variable λ, called a Lagrange multiplier according to the Lagrange Method. Since the connections of UGS traffic will be allocated to its required bandwidth, we examine the Lagrange function defined by Eqn.(14)

$$
\begin{aligned}
F(x_{M+1}, x_{M+2}, \ldots, x_K, \lambda) &= \sum_{i=M+1}^{M+N} \left(\frac{1 - e^{-\rho_i x_i}}{1 - e^{-\rho_i}} \right) \\
&+ \sum_{i=M+N+1}^{K} \left(\frac{1 - e^{-\rho_i x_i}}{1 - e^{-\rho_i}} \right) \\
&+ \lambda \left(B_{total} - \sum_{i=1}^{M} B_i - \sum_{i=M+1}^{K} x_i B_i \right). \qquad (14)
\end{aligned}
$$

Then, we have the $\frac{\partial^2 F}{\partial x_i \partial x_j}$ Hermitian matrix M.

$$
\begin{aligned}
M &= \begin{bmatrix}
\frac{\partial^2 F}{\partial x_{M+1}^2} & \frac{\partial^2 F}{\partial x_{M+1}\partial x_{M+2}} & \cdots & \frac{\partial^2 F}{\partial x_{M+1}\partial x_K} \\
\frac{\partial^2 F}{\partial x_{M+2}\partial x_{M+1}} & \frac{\partial^2 F}{\partial x_{M+2}^2} & \cdots & \frac{\partial^2 F}{\partial x_{M+2}\partial x_K} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 F}{\partial x_K \partial x_{M+1}} & \frac{\partial^2 F}{\partial x_K \partial x_{M+2}} & \cdots & \frac{\partial^2 F}{\partial x_K^2}
\end{bmatrix} \\
&= \begin{bmatrix}
\frac{-\rho_{M+1}^2 e^{-\rho_{M+1} x_{M+1}}}{(1 - e^{-\rho_{M+1}})} & & \mathbf{0} \\
& \ddots & \\
\mathbf{0} & & \frac{-\rho_K^2 e^{-\rho_K x_K}}{(1 - e^{-\rho_K})}
\end{bmatrix}
\end{aligned}
$$

Obviously, if $\rho > 0$, $1 - e^{-\rho} > 0$. The matrix M is said to be negative-definite. Thus, there exist the locate largest extreme values for the utility of BS, U_{BS}. In the following, we show how to find the optimum solution. First, set the derivative $dF = 0$, which yields the system of equations:

$$
\begin{aligned}
\frac{\partial F}{\partial x_{M+1}} &= \frac{\rho_{M+1} e^{-\rho_{M+1} x_{M+1}}}{(1 - e^{-\rho_{M+1}})} - \lambda B_{M+1} = 0 \qquad &(15\text{-}M+1) \\
\frac{\partial F}{\partial x_{M+2}} &= \frac{\rho_{M+2} e^{-\rho_{M+2} x_{M+2}}}{(1 - e^{-\rho_{M+2}})} - \lambda B_{M+2} = 0 \qquad &(15\text{-}M+2) \\
\vdots \quad & \vdots \quad \vdots & \\
\frac{\partial F}{\partial x_K} &= \frac{\rho_K e^{-\rho_K x_K}}{(1 - e^{-\rho_K})} - \lambda B_K = 0 \qquad &(15\text{-}K) \\
\frac{\partial F}{\partial \lambda} &= B_{total} - \sum_{i=1}^{M} B_i - \sum_{i=1}^{K} x_i B_i = 0 \qquad &(15\text{-}K+1)
\end{aligned}
$$

According to Eqn.(15), we have

$$
\begin{aligned}
\frac{\partial F}{\partial x_i} &= \frac{\rho_i e^{-\rho_i x_i}}{(1 - e^{-\rho_i})} - \lambda B_i = 0 \\
\Rightarrow \quad x_i &= \frac{1}{\rho_i} \ln \left(\frac{\rho_i}{\lambda B_i (1 - e^{-\rho_i})} \right)
\end{aligned}
$$

We substitute the x_i into Eqn.(15-K+1), and then

$$
\begin{aligned}
B_{total} - \sum_{i=1}^{M} B_i &= \sum_{i=M+1}^{K} x_i B_i \\
&= \sum_{i=M+1}^{K} \frac{B_i}{\rho_i} \ln \left(\frac{\rho_i}{\lambda B_i (1 - e^{-\rho_i})} \right)
\end{aligned}
$$

Based on the equation above , we can find the solution of λ .

$$\lambda = \exp\left(\frac{\sum_{i=M+1}^{K} \frac{B_i}{\rho} \ln\left[\frac{\rho}{B_i(1-e^{-\rho})} \right]}{\sum_{i=M+1}^{K} \frac{B_i}{\rho}} + \frac{\sum_{i=1}^{M} B_i - B_{total}}{\sum_{i=M+1}^{K} \frac{B_i}{\rho}} \right)$$

and then we have

$$\begin{aligned} x_i = {} & \frac{1}{\rho} \ln \frac{\rho}{B_i(1-e^{-\rho})} \\ & - \frac{1}{\rho} \left(\frac{\sum_{j=M+1}^{K} \frac{B_j}{\rho} \ln\left[\frac{\rho}{B_j(1-e^{-\rho})} \right]}{\sum_{j=M+1}^{K} \frac{B_j}{\rho}} \right) \\ & - \frac{1}{\rho} \left(\frac{\sum_{j=1}^{M} B_j - B_{total}}{\sum_{j=M+1}^{K} \frac{B_j}{\rho}} \right) \end{aligned} \tag{15}$$

Finally, the solution of optimum bandwidth allocation is shown.

$$b^* = [b^*_{M+1}, \cdots, b^*_K]$$

$$\begin{aligned} b^*_i = x_i B_i = {} & \frac{B_i}{\rho} \ln \frac{\rho}{B_i(1-e^{-\rho})} \\ & - \frac{B_i}{\rho} \left(\frac{\sum_{j=M+1}^{K} \frac{B_j}{\rho} \ln\left[\frac{\rho}{B_j(1-e^{-\rho})} \right]}{\sum_{j=M+1}^{K} \frac{B_j}{\rho}} \right) \\ & - \frac{B_i}{\rho} \left(\frac{\sum_{j=1}^{M} B_j - B_{total}}{\sum_{j=M+1}^{K} \frac{B_j}{\rho}} \right) ; i \in [M+1, \cdots, K] \end{aligned} \tag{16}$$

Therefore, we just obtain only one solution of b^* to maximize the QoS level of the BS according to the derivations above.

5. PROPOSED CAC SCHEME

We proposed a simple CAC scheme, which is based on the optimum solution of the bandwidth allocation in Section 4. In the CAC scheme, we introduce two new features. One is that the allocated bandwidth of a connection will be larger than its minimum bandwidth limit, $b^*_i \geq B_i^{min}$. The minimum limit of the connection depends on its application requirement. if it is a UGS connection such as voice applications, its minimum bandwidth is set to its required bandwidth. If it is an rtPS connection, its minimum bandwidth is less than its required bandwidth. If the value b^*_K is less than the minimum requirement of bandwidth B_K^{min}, we reject the connection. The reason is that the quality of service of the connection will be poor, even if it is granted to enter the BS. The other is that the priority order is introduced for ongoing connections and new connections. The ongoing connections have higher priority to share the resource than new connections. That is, the ongoing connections will be not dropped in our CAC scheme.

Since the two features, this CAC mechanism causes the process of bandwidth allocation to be iterated. For example, if the allocated bandwidth amount of an ongoing connection, calculated by Eqn.(17), is lower than a bottom limit B_i^{min}, CAC scheme will not drop the connection but continue providing service for the connection. At that time, what the CAC scheduler does is to cancel the bandwidth allocation, exclude the ongoing connection from the bandwidth allocation process, and then restart a new bandwidth allocation procedure. The iterated process does not stop until the allocated bandwidth of the reminder connections are greater than their minimum bandwidth requirement. If there are nothing in the set of reminder connections in the end, the new connection will be blocked.

The detail of the scheme is shown in Algorithm I. After a new incoming connection sends its request of entry to a BS, CAC process starts. If the BS has the available bandwidth enough for the its demand at that time, the new connection K will be given the access right and its required resource is allocated to it. Otherwise, the optimum value of bandwidth b^* for the connections. Under a lack of resource, if a connection is accepted, two conditions should be satisfied. One is that the value b_K^* is greater than the threshold B_K^{min} for the new connection. The other is that allocated bandwidth b_i^* is greater than B_i^{min} for the ongoing connections. After the allocated bandwidth b^* is ready, we adjust allocated bandwidth of other $K-1$ ongoing connections using their optimum values. And then we allocate the b_K^* bandwidth to the new connection to the end. Thus, we can keep the QoS level of the BS as high as possible all the time. The step 7, 8 ,9 mean that we exclude the specific ongoing connection from the process of bandwidth allocation, whose optimal value of allocate bandwidth is less than its lowest limit. The iterated bandwidth allocation will not stop until the reminder connections satisfy the conditions, $b_i^* \geq B_i^{min}$.

Algorithm 1: Proposed CAC scheme

1 Waiting for a new connection arrival;
2 the Kth connection (a new incoming connection) sends its entry request given that there are $K-1$ connections served by a BS now;
3 **if** $B_K <= B_{ava}$ **then**
4   Accept the connection request, and allocate B_K to the new connection;
5 **else**
6   Calculate the optimum bandwidth value $\mathbf{b}^*$ according to Eqn.(18) ;
7   **if** $b_i^* < B_i^{\min}$ **then**
8   The ith connection is excluded for the bandwidth allocation this time, and $B_{total} = B_{total} - b_i^{old}$, $K = K-1$;
9   **go to 6**;
10   **end**
11   **if** $b_K < B_K^{min}$ **then**
12   Block the new connection request;
13   **go to 1**;
14   **else**
15   Accept the new connection request;
16   **for** $i = 1$ **to** $K-1$ **do**
17   The connections excluded by CAC scheduler keep their original resources.;
18   Adjust the bandwidth of connections based on the optimum bandwidth value b_i^* ;
19   **end**
20   Allocate the bandwidth b_K^* to the new connection;
21   **end**
22 **end**
23 **go to 1**;

6. SIMULATIONS

We conduct computer simulation to test the performance of the proposed scheme through a simulation model, shown in Fig.3. To evaluate the overall performance of the CAC scheme, the simulation scripts are run multiple times to obtain statistically meaningful results.

The simulation model is constructed according to the CAC model in Section 3.1, and includes the components listed below:

- Pseudo Connection Generators: It creates connections with the inter-arrival times, just like "customers" in queuing theory. The inter-arrival time T of a Poisson arrival process is an exponential random variable, The arrive rate of connections is presumed as 30 to 100 per hour according to different applications and services. Every connection has its minimum bandwidth requirement and its application characteristic parameters. The bandwidth requests of the connections are from 64Kbps to 2Mbps. A service time span that a connection stays in the system is assigned to every connection. The value of the service time is also drawn from exponential distribution with mean value from 600 to 1 500 seconds.
- Ongoing Connection Management: When a new connection is accepted by CAC, it will be placed into a dynamic queue of ongoing connections. The ongoing connection is removed by the Connection Queue Management when its service time is end.

- Call Admission Control Scheduler: It carries out a CAC algorithm including QoS level calculation, providing the access right to a connection, and allocating the total bandwidth (Bt o t al = 75M b p s) to connections.

Traffic Type	ρ	Required bandwidth (Kbps)	Minimum bandwidth (Kbps)	Arrival rate (/hour)	Service time (seconds)
UGS	N/A	64	64	100	600
BE	0.01	90	64	90	700
rtPS	4	128	64	80	800
rtPS	5	256	128	70	900
BE	0.01	384	192	60	1000
nrtPS	0.01	768	384	50	1100
nrtPS	0.01	1024	512	40	1200
rtPS	6	1512	756	40	1200
rtPS	7	1768	884	30	1500
rtPS	8	2048	1024	30	1500

TABLE 3: Connection Types

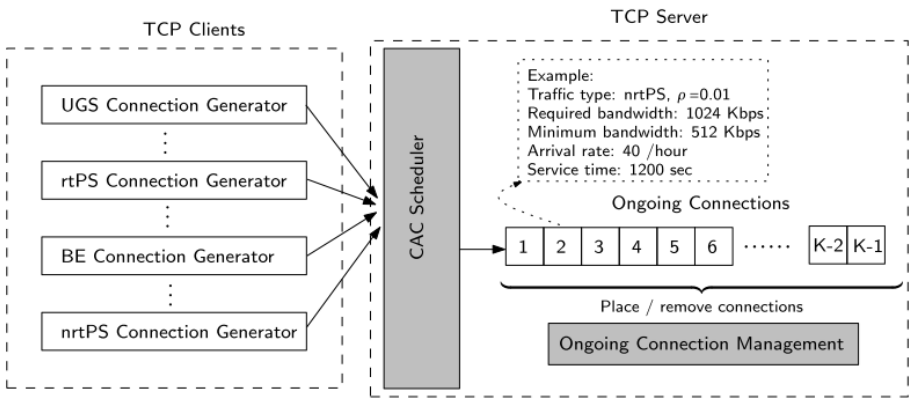

FIGURE 3: The Structure of simulation model

Here, we choose two other CAC methods as the simulation benchmarks. The first one is labeled as 'Baseline'. In the method, we do not take any extra CAC strategy. We just accept one new connection if available bandwidth is enough for it, or block it if not. The other is labeled as 'Elkadi' algorithm which is a typical resource reservation method. It makes bandwidth reservation for new connections, whose resources are taken from the ongoing connections [15].

	Baseline	Elkadi	Proposed
Block rate (%)	4.15	0	0.82
Bandwidth utilization(%)	86.70	83.57	88.42
Connection capacity	137.55	145.60	144.01
Avg. QoS level of connections	1.00	0.94	0.98
QoS level of BS	137.55	136.28	141.19

TABLE 4: Summary of Simulation Results

The simulation scripts are run multiple times to obtain statistically meaningful results to evaluate the overall performance of the CAC scheme. Table IV summarizes the simulation results for the three CAC methods. As can be seen, we investigated the performance through not only the traditional measurement parameters such as block rate, bandwidth utilization, and connection capacity, but also the proposed metric: average QoS level of connections and the overall QoS level of BS. For example, the proposed method is the best one in the bandwidth utilization and

QoS level of BS. Compared to the Baseline method, the proposed method decreases the block rate five times, which varies from 4.15% to 0.82%. The value 0.82% is small, approaching to the performance of Elkadi method. The Elkadi method does not block any new connections even under a heavy load traffic condition, because it makes a resource reservation enough for all the new connections. The average connection capacity of our method increases by about more than 5% connections than the baseline one. In addition, its value 144.01 is close to 145.60, the one of the Eldadi method. Therefore, we think the proposed method can achieve comprehensive performance better than others can.

On the other hand, Fig. 4 - 8 shows the measurement curves of the three methods during the whole simulation time. There are six curves in each figure, where the horizontal lines with dot line describe the mean value of the corresponding measurement variable. In Fig.4, we can see that the block rate curve of the proposed method keeps zero during half of simulation time. The curve of the baseline method always fluctuates because it often blocks the new connections under the heavy load traffic. Fig.5 illustrates that the number of ongoing connections of the proposed method and Elkadi method are similar during the simulation, and their performance are better than the baseline method. The bandwidth utilization curve of the Elkadi method is lower than other two methods, illustrated in Fig.6. The reason is that the Elkadi method makes a bandwidth reservation to accommodate new incoming connections by means of cutting down the bandwidth of ongoing connections. On the other hand, its connection capacity benefits from the reservation methodology and its curve can be always keeping higher, shown in Fig. 5.

For the view of users, the curves of QoS level are more meaningful for individual connection. Our target is to satisfy both the ongoing connections and new connections. The curve of average QoS level of connections for the Elkadi method falls below the others, which suggests the dissatisfactions of ongoing connections. The baseline method can indeed keep the average QoS level of ongoing connections as the maximum value. However, it should be noticed that it denies many requests of new incoming connections frequently, which contributes to the highest block rate. This is, the satisfaction degree of new connections is low in this case. These two benchmark methods cannot deal with the tradeoff among the requirements of the new connections and ongoing connections. On contrast, the proposed method makes the average QoS level of ongoing connections higher than Elkadi method as well in the simulation. At the same time, the block rate of new connections is kept close to a tiny value. The result of the proposed method can satisfy both ongoing connections and new connections.

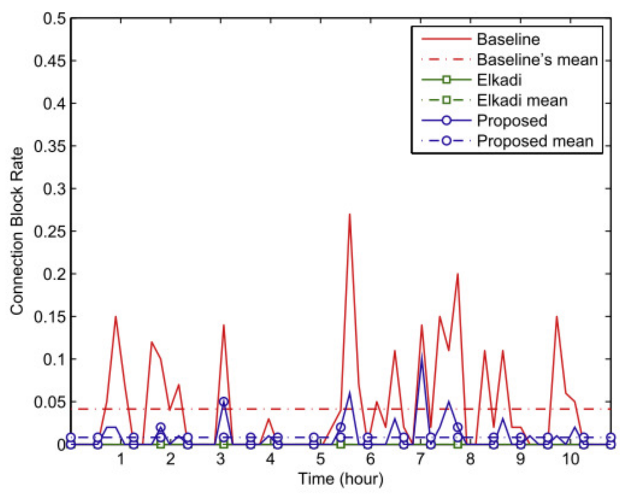

FIGURE 4: Comparison of block rate of connections

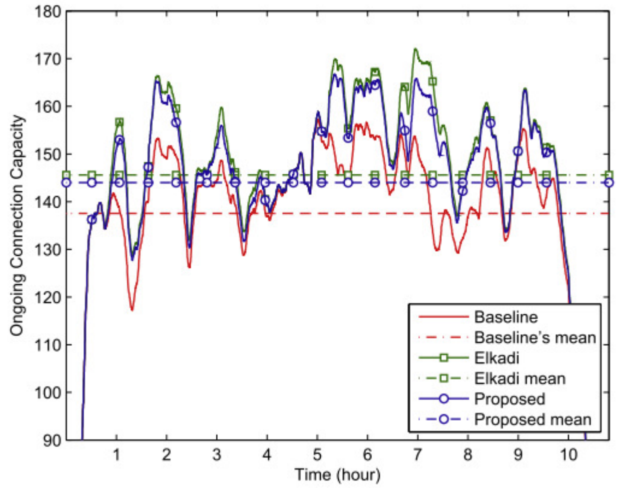

FIGURE 5: Comparison of capacity of connections

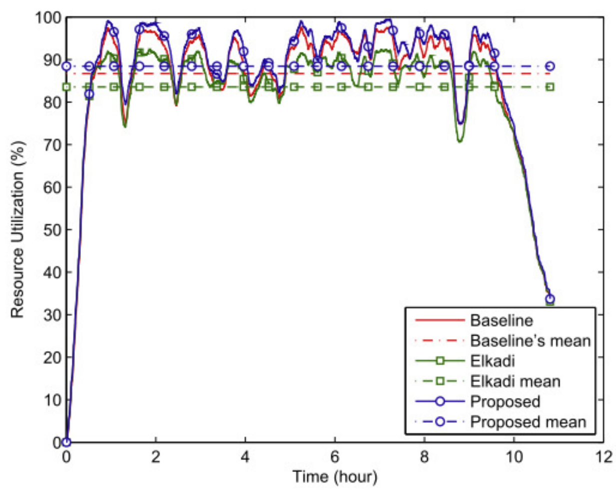

FIGURE 6: Comparison of bandwidth utilization

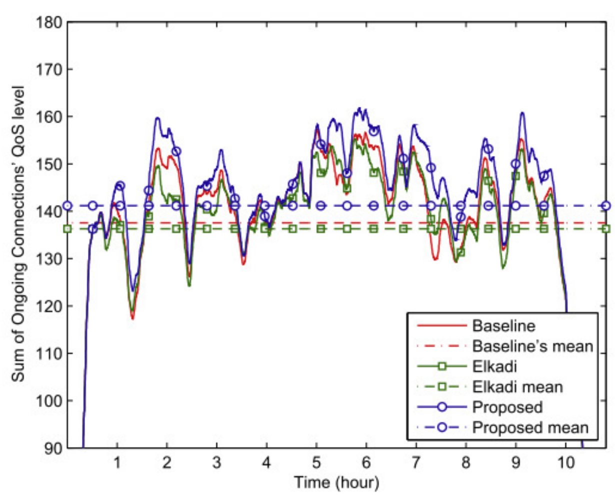

FIGURE 7: Comparison of QoS Level of a BS

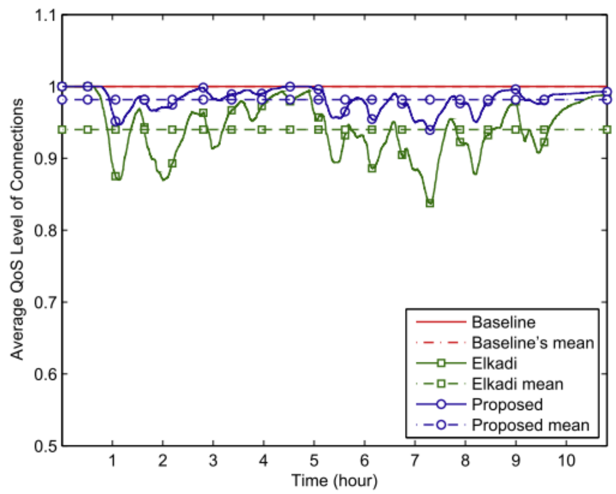

FIGURE 8: Comparison of average QoS level of connections

7. CONCLUSION

We introduce a metric of individual QoS level for multimedia data defined in WiMAX standard in the paper. Using the metric, a CAC scheduler can take account of the QoS level of the new connections and the ongoing connections, and then obtain the optimum solution of bandwidth allocation. The solution is provided to the CAC scheduler to determine whether a connection is accepted or not. The simulation results indicate that our proposed method can achieve better trade-off among the new connections, the ongoing connections and BS system performance.

REFERENCES

[1] H. G. Perros and K. M. Elsayed, "Call Admission Control Schemes: A Review," *IEEE Communications Magazine*, vol. 34, no. 11, pp. 82–91, 1996.

[2] K. Mase, "Toward scalable admission control for VoIP networks," *IEEE Communications Magazine*, vol. 42, no. 7, pp. 42–47, 2004.

[3] Cisco-System. (2001) Voip call admission control document. [Online]. Available: http://www.cisco.com/- univercd/~cc/td/doc/cisintwk/intsolns/voipsol/cac.htm

[4] Y. Fang and Y. Zhang, "Call admission control schemes and performance analysis in wireless mobile networks," *Vehicular Technology, IEEE Transactions on*, vol. 51, no. 2, pp. 371 –382, mar 2002.

[5] Y. Xiao, P. Chen, and Y. Wang, "Optimal admission control for multi-class of wireless adaptive multimedia services," *IEICE Transactions on Communications*, vol. E84-B, no. 4, pp. 795–804, april 2001.

[6] Y. Qian, R.-Y. Hu, and H.-H. Chen, "A call admission control framework for voice over wlans," *Wireless Communications*, IEEE, vol. 13, no. 1, pp. 44 – 50, feb. 2006.

[7] G. Djukanovic, N. Gospic, and M. Sunjevaric, "Cac negotiation algorithm proposal for higher qos in heterogeneous wireless networks," in Telecommunications in Modern Satellite, Cable and Broadcasting Services, 2007. TELSIKS 2007. 8th International Conference on, 26-28 2007, pp. 19 –22.

[8] W. Song and W. H. Zhuang, "Multi-service load sharing for resource management in the cellular/wlan integrated network," *IEEE Transactions on Wireless Communications*, vol. 8, no. 2, pp. 725–735, 2009.

[9] O. Yilmaz and I. R. Chen, "Utilizing call admission control for pricing optimization of multiple service classes in wireless cellular networks," *Computer Communications*, vol. 32, no. 2, pp. 317–323, 2009.

[10] X. P. Zhai, G. G. Bi, and P. P. Xu, "Call admission control with bandwidth reallocation for adaptive multimedia in high-rate short-range wireless networks," *Chinese Journal of Electronics*, vol. 14, no. 1, pp. 169–173, 2005.

[11] W. L. Ni, W. Li, and M. Alam, "Determination of optimal call admission control policy in wireless networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 2, pp. 1038–1044, 2009.

[12] H. Zhai, X. Chen, and Y. Fang, "A call admission and rate control scheme for multimedia support over ieee 802.11 wireless lans," *Wireless Networks*, vol. 12, no. 4, pp. 451–463, 2006.

[13] H. Chen, S. Kumar, and C. J. Kuo, "Dynamic call admission control and resource reservation with interference guard margin (igm) for cdma systems," in Wireless Communications and Networking, IEEE, vol. 3, no. C, 2003, pp. 1568–1572.

[14] Sunil and C. J. Kuo, "Qos-aware radio resource management scheme for cdma cellular networks based on dynamic interference guard margin (igm)," *Computer Networks*, vol. 46, no. 6, pp. 867–879, 2004.

[15] M. El-Kadi, S. Olariu, and H. Abdel-Wahab, "A rate-based borrowing scheme for qos provisioning in multimedia wireless networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 13, no. 2, pp. 156–166, 2002.

[16] K. Elsayed, A. Saad, and M. El-Hadidi, "Performance evaluation of resource reservation and call admission policies for deterministic services in pgps-based packet networks," *Elsevier Computer Communications*, vol. 25, pp. 1513–26, 2002.

[17] K. Tsagkaris and P. Demestichas, "Wimax network," IEEE Vehicular Technology Magazine, vol. 4, no. 2, 2009.

[18] J. G. Andrews, A. Ghosh, and R. Muhamed, Fundamentals of WiMAX. Prentice Hall, 2007.

[19] ITU-T Recommendation G.107. The e-model, a computational model for use in transmission planning, MC68175/D, ITU, December.

[20] K. N. N. Kitawaki and T. Yamada, "Objective quality assessment of wideband speech coding," *IEICE Transactions on Communications*, vol. E88-B(n3), 2005.

[21] ITU-T. (2008) R value calculator. [Online]. Available: http://www.itu.int/itu-t/studygroups/com12/emodelv1/calcul. php

[22] Z. He, J. Cai, and C. W. Chen, "Joint source channel rate-distortion analysis for adaptive mode selection and rate control in wireless video coding," *Circuits and Systems for Video Technology, IEEE Transactions* on, vol. 12, no. 6, pp. 511–523, Jun 2002.

[23] E.-H. Yang and X. Yu, "Rate distortion optimization for h.264 interframe coding: A general framework and algorithms," *Image Processing, IEEE Transactions on*, vol. 16, no. 7, pp. 1774 – 1784, july 2007.

[24] J. Liu, Z. Guo, and Y. Cho, "Bit allocation for joint spatial-quality scalability in h.264/svc," in Image Processing(ICIP), 2009 16th IEEE International Conference on, 7-10 2009, pp. 3737 –3740.

[25] J. Postel and J. Reynolds, "Rfc959, file transfer protocol (ftp)," 1985. [Online]. Available: http://tools.ietf.org/html/rfc959

[26] R. C. W. C. Fielding and Irvine, "Rfc2616 the hipertext transfer protocol," p. 176, 1999. [Online]. Available:http://www.ietf.org-/rfc/-rfc2616.txt