Dynamic Admission Control and QoS for 802.16 Wireless MAN

Haitang Wang, Wei Li and Dharma P. Agrawal
OBR Center for Distributed and Mobile Computing, ECECS
University of Cincinnati
Cincinnati, OH 45221-0030
{wangh7, liw0, dpa}@ececs.uc.edu

Abstract

IEEE 802.16 wireless MAN is expected to revolutionize of the broadband wireless access technology. The grant/request mechanism in IEEE 802.16 MAC povides different Quality of Service (QoS) for different service flows. In the paper, we discuss the QoS issue in IEEE 802.16 wireless MANs and propose an admission control scheme for services defined in the 802.16 specifications. Our proposed scheme provides the highest priority for UGS flows and maximizes the bandwidth utilization by bandwidth borrowing and degradation. We also develop an analytical model to evaluate the system performance. Some numerical results are included to provide a better understanding of our scheme.

1. Introduction

Wireless Metropolitan Area Networks (MANs) is being developed to replace the wireline infrastructure network with more efficient deployment and lower maintenance cost. IEEE Standard 802.16-2004 [1] defines the WirelessMAN^{TM} air interface specification for wireless MAN. As defined in IEEE 802.16, a wireless MAN provides network access to buildings through exterior antennas communicating with central radio based stations (BSs) [2].

The IEEE 802.16 Wireless MAN was initially proposed as a fixed broadband access system with multiple subscriber stations (SSs) link to a common BS. Within a SS, a large number of end users with different broadband access requirements can be present. The broadband access requirements can be classified into four types according to the scheduling service in IEEE 802.16: Unsolicited Grant Service (UGS), Real-time Polling Service (rtPS), Non-Real-Time Polling Service (nrtPS) and Best Effort (BE). Providing guaranteed QoS for four different types of multimedia service flows in IEEE 802.16 wireless MAN is a real challenging problem. IEEE 802.16 MAC, defined as connection-oriented, is designed to support different QoS for different services. Once a new service flow arrives at the SS of a Wireless MAN, the SS attempts to set up an end to end uplink connection with BS. Once the BS receives the request, it makes decision on whether to admit the connection or not, and how much bandwidth should be set aside to the connection for its entire transmission duration. The process we described above is called admission control process in IEEE 802.16. After the connection is set up, the end user starts transmitting data. Since data usually is generated in burst, when the connection has data to transmit, the SS uses the opportunities got from polling process to send bandwidth request to BS. After receiving the bandwidth request, BS makes a decision whether to grant or not grant the bandwidth and how much should be granted to the request. The decision making process on this step is called granting control in IEEE 802.16 wireless MAN. The decisions of both admission control and granting control are based on the current bandwidth utilization of system and the QoS requirements of the each type of connections. The difference of both controls is that the admission control considers the long term (i.e., the life time of the connection) performance and the bandwidth usage of the system while the granting control only concerns with the instantaneous effect when a bandwidth request is received. Intuitively, after the

This work has been supported by Ohio Board of Regents Doctoral Enhancement Funds and National Science Foundation under Grant No. CCR-0113361.

0-7803-8856-9/05/$20.00 ©2005 IEEE

optimal admission control and granting control process the demands of all the users in the system are satisfied. However, both important processes are not addressed at all in the standard of IEEE 802.16, and are our research goal.

In recent years, numerous multimedia admission control schemes and performance evaluations have been introduced [3] [4] [5] [6] [7] in the current existing wireless networks, like GPRS, Wireless LAN and CDMA. Most of existing works focus only on the two types of service in the current wireless networks: real-time service and non real-time service. As we know, in the wireless and mobile network, handoff calls should be given higher priority than new calls and has been totally ignored in IEEE 802.16. Though 802.16 working group is also looking at the mobile Wireless Broad Access (IEEE 802.16e), the most important concern for providing admission control and granting control in IEEE 802.16 Wireless MAN is how to maximize the bandwidth utilization and guarantee QoS of different service flows. To the best of our knowledge, there is no specific admission and granting control scheme and performance evaluation proposed for IEEE 802.16 Wireless MAN. In this paper, we firstly propose an admission control scheme and then introduce an analytical model to evaluate the performance of different multimedia service flows in IEEE 802.16 Wireless MAN. Trying to find good granting control schemes for IEEE 802.16 is one of the future works.

The remainder of this paper is organized as follows: Section 2 gives introduction of IEEE 802.16 Wireless MAN. Section 3 introduces our proposed scheme for IEEE 802.16 wireless MAN. Section 4 derives an analytical model to evaluate the QoS parameters for different multimedia service flows. Section 5 shows numerical results of our analytical model. Section 6 concludes the paper.

2. IEEE 802.16 Wireless MAN

Any IEEE 802.16 wireless MAN will include at least one BS and some SSs, the central BS handling multiple independent SSs simultaneously and regulates all the communications in the network. Data on the downlink to SSs is multiplexed in TDM fashion with individual SS allocated time slots serially. Uplink is multiple access shared media from SSs to BS in TDMA (IEEE 802.16) or FDMA (IEEE 802.16a).

In order to support the QoS for different services by scheduling the uplink access opportunity, four different scheduling services corresponding to uplink scheduler policy are defined in the standard: UGS, rtPS, nrtPS, and BE.

- UGS

UGS is designed to support real-time service flows that generate fixed-size data packets on periodic basis, such as T1/E1 and Voice over IP with silence suppression [1]. The BS allocates fix size grants to the UGS at periodic intervals without any explicit request from the SS which eliminates the overhead and the latency of bandwidth requests so as to meet the real time requirement of UGS service. The BS can adaptively allocate additional capacity to the SS when backlog in the transmission queue of SS is detected.

- rtPS

The rtPS is designed to support real-time service flows that generate variable size data packets on a periodic basis, such as moving pictures experts group (MPEG) streaming video [2]. The BS provides periodic dedicated request opportunities for SS to meet flow's real-time demands. In order for the service to work appropriately, the SS is allowed to use only unicast request issued by BS for connection and is prohibited from using any other contention request opportunities.

- nrtPS

The nrtPS is designed to support delay-tolerant data streams and consists of variable-sized data packets which require a minimum data rate, such as FTP [2]. The nrtPS is almost the same as the rtPS except that connections may utilize random access transmit opportunities for sending bandwidth request.

- BS

The BS service is designed to support data streams for which no minimum transmission rate is required and therefore maybe handled on a space-available basis [2], such as HTTP. The SS is allowed to use contention request opportunities as well as unicast request opportunities for BE service flow. The interval of unicast request opportunities should be longer than the nrtPS and the availability of dedicated opportunities is subject to the network load.

3. Admission control based on the scheduling services characteristics

In this section we describe our proposed admission control and bandwidth reservation scheme at the network layer. As we stated in the first section, before the transmission of a connection, BS should decide whether to accept or reject the user's request for connection. The decision is usually made based on long term bandwidth requirements of the connections and the current network state. The long term bandwidth requirement here means the estimation of bandwidth requirement during the whole transmission, which is different from the actually granted bandwidth for every service data unit (SDU) transmission. Similar to most admission controls, BS sets aside certain amount of bandwidth for the service flow, which supports efficient granting control in the MAC layer. The admission control performs a tradeoff between accepting a request for connection that may result in a QoS degradation of already admitted connections and rejecting a request for connection in order to support QoS of ongoing connections at a certain level.

UGS flow, like E1/T1 and VoIP, is the most common way used by people for daily communication. The non UGS flows like rtPS, nrtPS and BE flows, are used to support stream video, ftp, or HTTP applications. Most of applications are used for entertainment and the actual rate varies during transmission. From the viewpoint of end user, blocking the new UGS flow causes more serious problem than blocking the new non-UGS flow. We therefore give UGS connection higher priority over non UGS connection so that every request for UGS connection is admitted and its required bandwidth for transmission is guaranteed. In other words, the request for UGS connection is accepted without restriction if bandwidth is available, while the request of non-UGS is only accepted when the total used bandwidth is not greater than the predetermined value. We assume that the scenario considered in this paper is homogenous and thus, we can examine a single SS in isolation. Suppose the total bandwidth for a SS is B, the predetermined value is B-U, where U is the bandwidth exclusively reserved for UGS.

Assume the bandwidth required by a UGS connection and rtPS connection to be b_{UGS} and b_{rtPS} respectively. The required bandwidth must be satisfied in order to meet QoS requirements of UGS and rtPS connections.

Due to the property of nrtPS flow, the required bandwidth of an nrtPS flow may vary within the range of $[b_{nrtPS}^{\min}, b_{nrtPS}^{\max}]$, where $b_{nrtPS}^{\max}$ and $b_{nrtPS}^{\min}$ are the maximum and minimum bandwidth required for the nrtPS flow, respectively. If sufficient bandwidth is available (i.e., fewer connections), each nrtPS flow can be transmited at a higher rate. As the number of connections increases, the existing nrtPS flows can give up some bandwidth to new connections in order to have more UGS, nrtPS or rtPS connections in the system. We call this as a degradation model. The degradation is performed stepwise and δ is the amount of degraded bandwidth for every degradation step. All the nrtPS connections in the system are allowed to maintain the same degradation level. Let l_{nrtps}^{n} be the current degradation level. Thus, the current reserved bandwidth for each nrtPS connection is $b_{nrtPS}^{\max} - l_{nrtps}^{n}\delta$ which satisfies $b_{nrtPS}^{\max} - l_{nrtps}^{n}\delta \geq b_{nrtPS}^{\min}$. The maximum degradation step is $l_{nrtps}^{\max} = (b_{nrtPS}^{\max} - b_{nrtPS}^{\min})/\delta$

In what follows, we describe our admission control scheme in more details.

- When a request for UGS connection arrives at the BS, if the bandwidth currently set aside for all ongoing connections plus b_{UGS} is less than or equal to B, the request is accepted, then BS sets aside b_{UGS} bandwidth for this connection during its lifetime. Otherwise, the request for the UGS connection is rejected.
- When a request for rtPS connection arrives at the BS, if total bandwidth set aside for all ongoing connections plus b_{rtPS} is less than or equal to B-U, the connection is set up and BS sets aside b_{rtPS} bandwidth for the connection. Otherwise, BS degrades the bandwidth set aside for all ongoing nrtPS connections until all bandwidth set aside for all ongoing connections plus b_{rtps} is not greater than B-U. If the currently set aside bandwidth plus b_{rtps} is still greater than B-U and the maximum degradation step $l_{nrtps}^{\max}$ has been reached, the request for rtPS connection is rejected.

- When a request for nrtPS connection arrives at the BS, if the total bandwidth already set aside for all ongoing connections plus $b_{nrtPS}^{\max} - l_{nrtPS}^{n}\delta$ is less than or equal to B-U, the connection is set up. Otherwise, BS degrades the bandwidth set aside for nrtPS ongoing connections until the current total bandwidth for all ongoing connections plus the bandwidth for the new nrtPS connection is not greater than B-U. In both cases, the bandwidth set aside for the new connection is $b_{nrtPS}^{\max} - l_{nrtps}^{n'}\delta$, which is the same for all nrtPS connections in the system. $l_{nrtps}^{n'}$ ($l_{nrtPS}^{n} \le l_{nrtps}^{n'} \le l_{nrtPS}^{\max}$) is the updated degradation level of all nrtPS connections after admitting new connetion. Otherwise, the request for nrtPS connection is rejected.
- When the request for BE connection arrives at the BS, the request is always admitted, but BS will not set aside any bandwidth for such a connection. In 802.16 MAC layer, the BS connections get the transmission opportunities only when other service connections do not transmit. Generally, BS connections do have long idle period (think time) and data in each transmission is relatively small, especially in the uplink direction. Therefore, QoS of BS can be easily satisfied. In the following discussion, we only consider the other three services.

4. Analytical Model

We compute the performance of the admission scheme based on the following assumptions:

The arrival process of new connection requests for UGS, rtPS and nrtPS is Poisson with rate λ_{UGS}, λ_{rtPS}, λ_{nrtPS} respectively. Let $\lambda_{total} = \lambda_{UGS} + \lambda_{rtPS} + \lambda_{nrtPS}$, $\lambda_{UGS} = \lambda_{total} \times \alpha_{UGS}$, $\lambda_{rtPS} = \lambda_{total} \times \alpha_{rtPS}$, and $\lambda_{nrtPS} = \lambda_{total} \times \alpha_{nrtPS}$.

The service time for UGS rtPS and nrtPS connections is exponentially distributed with mean $1/\mu_{UGS}$, $1/\mu_{rtPS}$, and $1/\mu_{nrtPS}$, respectively.

We model each SS with a four dimensional continuous Markov chain. The state of SS can be represented by (n_{UGS} , n_{rtPS} , n_{nrtPS} , l_{nrtPS}^{n}), where n_{UGS}, n_{rtPS}, n_{nrtPS} is the current number of connections of UGS, rtPS, and nrtPS in the system, respectively. l_{nrtPS}^{n} is the current degradation level of nrtPS. The stead probability of the state s = (n_{UGS} , n_{rtPS} , n_{nrtPS} , l_{nrtPS}^{n}) is $\pi_{(n_{UGS}, n_{rtPS}, n_{nrtPS}, l_{nrtPS}^{n})}$.

A state is a valid state if and only if:

$$n_{UGS}b_{UGS} + n_{rtPS}b_{rtPS} + n_{nrtPS}(b_{nrtPS}^{\max} - l_{nrtPS}^{n}\delta) \le B$$
$$\wedge\, n_{rtPS}b_{rtPS} + n_{nrtPS}(b_{nrtPS}^{\max} - l_{nrtPS}^{n}\delta) \le B - U$$
$$\wedge\, l_{nrtPS}^{n} \le l_{nrtPS}^{\max}$$

We define S is the state space for all possible states that

$$S = \{s = (n_{UGS}, n_{rtPS}, n_{nrtPS}, l_{nrtPS}^{n}) \mid$$
$$n_{UGS}b_{UGS} + n_{rtPS}b_{rtPS} + n_{nrtPS}(b_{nrtPS}^{\max} - l_{nrtPS}^{n}\delta) \le B$$
$$\wedge\, n_{rtPS}b_{rtPS} + n_{nrtPS}(b_{nrtPS}^{\max} - l_{nrtPS}^{n}\delta) \le B - U$$
$$\wedge\, l_{nrtPS}^{n} \le l_{nrtPS}^{\max}\}$$

For a given state s = (n_{UGS}, n_{rtPS}, n_{nrtPS}, l_{nrtPS}^{n}), state transition occurs when a new request is accepted or when an ongoing communication completes. Fig. 1 shows the general state transition diagram. Note that $l_{nrtPS}^{n'}$ represents the degradation level of nrtPS flows after the state transition and may have value different from l_{nrtPS}^{n}. Also, state transitions to some directions may not exist under certain conditions. Therefore, the state transition diagram can be different for different states.

Each state transition diagram has a corresponding state transition balance equation. As an example,

if $n_{UGS} > 0 \wedge n_{rtPS} > 0 \wedge n_{nrtPS} > 0 \wedge l_{nrtPS}^{n} \le l_{nrtPS}^{\max}$

$$\wedge\, n_{UGS}b_{UGS} + n_{rtPS}b_{rtPS} + n_{nrtPS}(b_{nrtPS}^{\max} - l_{nrtPS}^{n}\delta) + b_{UGS} \le B$$
$$\wedge\, n_{UGS}b_{UGS} + n_{rtPS}b_{rtPS} + n_{nrtPS}(b_{nrtPS}^{\max} - l_{nrtPS}^{n}\delta) + b_{rtPS} \le B - U$$
$$\wedge\, n_{UGS}b_{UGS} + n_{rtPS}b_{rtPS} + n_{nrtPS}(b_{nrtPS}^{\max} - l_{nrtPS}^{n}\delta) + (b_{nrtPS}^{\max} - l_{nrtPS}^{n}\delta) \le B - U,$$

the state balance equation is:

$$(\lambda_{UGS} + \lambda_{rtPS} + \lambda_{nrtPS} + n_{UGS}\mu_{UGS} + n_{rtPS}\mu_{rtPS} + n_{nrtPS}\mu_{nrtPS})\pi_{(n_{UGS}, n_{rtPS}, n_{nrtPS}, l_{nrtPS}^{n})}$$
$$= \lambda_{UGS}\pi_{(n_{UGS}-1, n_{rtPS}, n_{nrtPS}, l_{nrtPS}^{n})} + \lambda_{rtPS}\pi_{(n_{UGS}, n_{rtPS}-1, n_{nrtPS}, l_{nrtPS}^{n})} + \lambda_{nrtPS}\pi_{(n_{UGS}, n_{rtPS}, n_{nrtPS}-1, l_{nrtPS}^{n})}$$
$$+ (n_{UGS}+1)\mu_{UGS}\pi_{(n_{UGS}+1, n_{rtPS}, n_{nrtPS}, l_{nrtPS}^{n})} + (n_{rtPS}+1)\mu_{rtPS}\pi_{(n_{UGS}, n_{rtPS}+1, n_{nrtPS}, l_{nrtPS}^{n})}$$
$$+ (n_{nrtPS}+1)\mu_{nrtPS}\pi_{(n_{UGS}, n_{rtPS}, n_{nrtPS}+1, l_{nrtPS}^{n})}$$

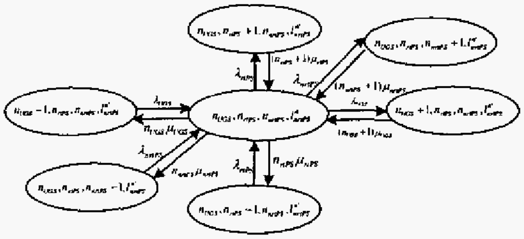

Figure 1. Transition diagram for the state $s = (n_{UGS}, n_{rtPS}, n_{nrtPS}, l^{n}_{nrtPS})$

The number of equations equals the number of states. After we get all global balance equations, we combine them with the normalized condition that sum of all steady state probabilities is 1:

$$\sum_{s\in S}\pi(s) = 1 \cdot$$

By solving all the linear equations using the Markov model [8], we can get all the steady state probabilities $\pi(s)$ and the QoS parameters of performance evaluation can be determined by those state probabilities.

a. New connection blocking probability (CBP)

The new *connection blocking probability* (CBP) is the probability of rejecting a new request for connection. As an example, the CBP of UGS is the probability of blocking a new arriving UGS request.

$$BP_{UGS} = \sum_{s\in S'}\pi(s),$$

where $S' = \{s = (n_{UGS}, n_{rtPS}, n_{nrtPS}, l^{n}_{nrtPS}) \mid (n_{UGS}+1)b_{UGS} + n_{rtPS}b_{rtPS} + n_{nrtPS}b^{\min}_{nrtPS} > B\}$.

$$BP_{nrtPS} = \sum_{s\in S''}\pi(s),$$

where $S'' = \{s = (n_{UGS}, n_{rtPS}, n_{nrtPS}, l^{n}_{nrtPS}) \mid n_{UGS}b_{UGS} + n_{rtPS}b_{rtPS} + (n_{nrtPS}+1)b^{\min}_{nrtPS} > B-U\}$

b. Bandwidth utilization

The *bandwidth utilization* is the average ratio of the used bandwidth to the total bandwidth. Formally, it can be calculated as:

$$BU = \sum_{s\in S}(n_{UGS}b_{UGS} + n_{rtPS}b_{rtPS} + n_{nrtPS}(b^{\max}_{nrtPS} - l^{n}_{nrtPS}\delta))\pi(s)/B$$

c. Transmission capacity

The transmission capacity is the traffic volume transmitted within a time interval T by the system. In our proposed scheme, system can dynamically changes the transmission rate of nrtPS to accommodate more connections into the system. In the degradation mode, the blocking probability of nrtPS is decreased, as well as the blocking probability of other services. That means more connections can be admitted into the system in the degradation mode. The transmission capacity of nrtPS is an expected value and can be calculated as:

$$TC_{nrtPS} = \sum_{s\in S} n_{nrtPS}(b^{\max}_{nrtPS} - l^{n}_{nrtPS}\delta)\pi(s)T,$$

where T is time interval.

5. Numerical Results

In this part we present our numerical results for the QoS parameters discussed in the last part. The system parameters used in our theoretical model are shown in Table 1.

Table 1: System parameters

B	1280 kbps
δ	32kbps
b_{UGS}	32kbps
b_{rtPS}	128 kbps
$b^{\max}_{nrtPS}$	128kbps
$b^{\min}_{nrtPS}$	32kbps

Figure 2 and Figure 3 compare the blocking probabilities of UGS and nrtPS with different value of U. We see that the reservation method is useful for decreasing CBP of UGS. In Figure 3, the CBP of nrtPS is increased by the reservation method. As we know, UGS connections are more important than nrtPS connections, the reservation method could be said to be an efficient method to give higher priority to some services to improve their performance.

Figure 4 compares the bandwidth utilization when nrtPS works in the degradation mode and in the constant rate mode. When working in the constant rate mode, the nrtPS flows can only transmit at the maximum rate (i.e., 128kbps). Figure 4 illustrates the degradation mode has better bandwidth utilization than the constant rate mode. Moreover, the transmission capacity of nrtPS in the degradation mode is also larger than that in the constant rate mode because more nrtPS can be accepted in the degradation mode, even though the average transmission rate of nrtPS in the degradation mode is smaller. The results are shown in Figure 5.

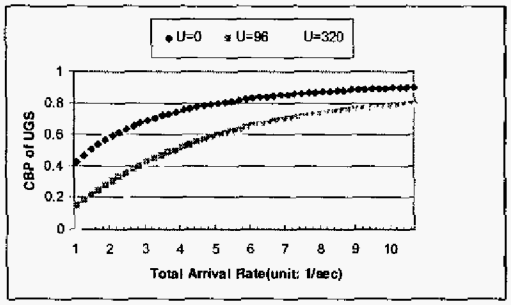

Figure 2. Comparison of CBP of UGS for different

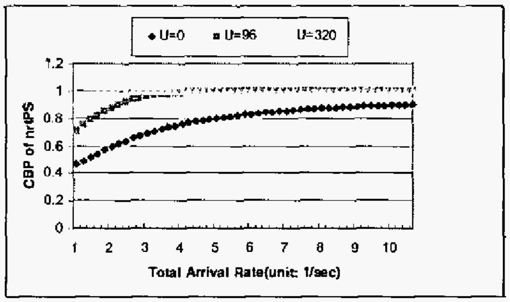

Figure 3. Comparison of CBP of nrtPS for different U

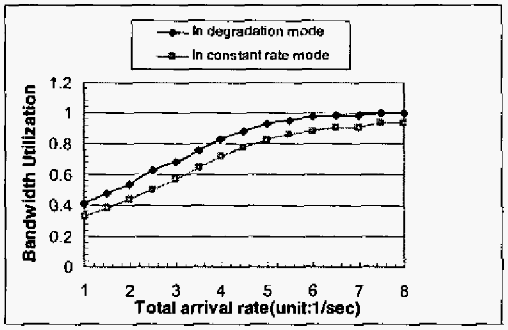

Figure 4. Comparison of Bandwidth Utilization

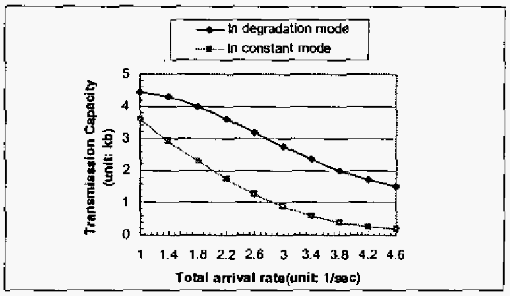

Figure 5. Comparison of Transmission Capacity

6. Conclusion

IEEE 802.16 wireless MANs supports service differentiation. In this paper, we discuss the QoS issue in IEEE 802.16 wireless MANs and propose an admission control scheme for services defined in the 802.16 specification. The proposed scheme gives the highest priority for UGS flows and maximizes the bandwidth utilization by bandwidth borrowing and degradation. We also develop an analytical model to evaluate system performance and some numerical results are provided.

We believe that the work presented in this paper is the first step of our research on QoS of IEEE 802.16 wireless MAN. In order to satisfy the QoS requirements of different services, the admission control must work together with an appropriate granting control at the MAC layer so that the actual bandwidth can be allocated when data unit is ready for transmission. We are planning to investigate these issues in our future work.

Reference

[1] IEEE Std 802.16-2004 (Revision of IEEE Std 802.16-2001): IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems.

[2] C. Eklund, R. B. Marks, K. L. Stanwood, and S. Wang, "IEEE Standard 802.16: A Technical Overview of the WirelessMAN™ Air Interface for Broadband Wireless Access," *IEEE Communications Magazine* , June 2002, Vol. 40, pp. 98-107.

[3] C. W. Leong, W. Zhuang, Y. Cheng, and L. Wang, "All Admission control for Integrated On/Off Voice and Best Effort Data Services in

Mobile Cellular Communications," *IEEE Transactions. on Communications,* May 2004, Vol.52, pp. 778-790.

[4] J. Hou, J. Yang, and S. Papavasiliou, "Integration of pricing with call admission control to meet QoS requirement in cellular networks," *IEEE Transactions. on Parallel and Distributed Systems,* Sept. 2002, Vol. 13, pp. 898-910.

[5] S. Shen, C-J. Chang, C.Y. Huang, and Q. Bi, "Intelligent Call Admission Control for Wideband CDMA Cellular Systems," *IEEE Trans. on Wireless Communications,* Sept. 2004, Vol. 3, pp. 1810-1821.

[6] R. Q. Hu, J. Babbitt, H. Abu-Amara, C. Rosenberg, and G. Lazarou, "Connectivity planning and call admission control in an on-board cross-connect based multimedia GEO satellite network," *IEEE International Conference on Communications,* May. 2003.

[7] J. Levendovszky, and A. Fancsali, "Real-time call admission control for packet-switched networking by cellular neural networks," *IEEE Trans. on Curcuits and Systems* I, June 2004, Vol. 51, pp. 1172-1183,.

[8] Q.Zeng, D. P. Agrawal, "Modeling and efficient handling of handoffs in integrated wireless mobile networks" *IEEE Trans. on Parallel and Distributed Systems*, Dec. 2002, Vol. 3, No. 12, pp. 1290 – 1302,