This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the ICC 2008 proceedings.

QOAR: Adaptive QoS Scheme in Multi-rate Wireless LANs

Ming Li[1], Yang Xiao[2], Hua Zhu[3], Imrich Chlamtac[4], B. Prabhakaran[5]

[1] Department of Computer Science, California State University, Fresno, CA 93740, USA (mingli@csufresno.edu)

[2] Department of Computer Science, The University of Alabama, Tuscaloosa, AL 35487-0290, USA (yangxiao@ieee.org)

[3] San Diego Research Center, San Diego, CA 92121 USA (hua.zhu@sdrcinc.net)

[4] Create-Net, Via Solteri 38, Trento 38100, Italy (chlamtac@create-net.org)

[5] Department of Computer Science, The University of Texas at Dallas, Richardson, TX 75080, USA (praba@utdallas.edu)

***Abstract*—With the availability of multiple rates in IEEE 802.11a/b/g wireless LANs, it is desirable to improve the network capacity and temporal fairness by sending multiple consecutive frames over high rate links, as proposed in opportunistic auto rate (OAR [1]). However, the basic OAR does not provide quality of service (QoS) guarantee and thus is not sufficient in supporting real time voice/video traffic. In this paper, we further enhance the OAR protocol with a set of QoS mechanisms. The proposed QOAR protocol consists of two protocols: (i) a traffic-differentiating flow weight adaptation protocol (FWA) that dynamically tunes both contention window and concatenation number per channel access; (ii) an admission control protocol (AC) that guarantees the bandwidth/delay requirements of multimedia services. Extensive simulation studies show that QOAR enables QoS for real-time traffic yet maximizes performance of best effort traffic.**

Keywords-multi-rate, wireless LAN, multimedia, QoS

I. INTRODUCTION

The IEEE 802.11 WLAN has been accepted as a complementary technology to high-speed IEEE 802.3 (Ethernet) for portable and mobile devices with its increasing data transmission rates and relatively low price. The IEEE 802.11, 802.11b, and 802.11a/g specifications provide up to 2, 11, and 54 Mbps data rates [3], respectively. With this broadband capability, it is desirable to support multimedia applications over IEEE 802.11 wireless LANs.

However, in standard CSMA/CA MAC protocol, each contending station can only transmit a single frame per successful channel access, regardless of link rate and frame size. This basic mechanism incurs high overhead, especially when link rate is high and frame size is small. To achieve better performance, opportunistic auto rate (OAR) [1] was proposed to improve the temporal fairness under various transmission rates such that the capacity of the basic Distributed Coordination Function (DCF) can be significantly improved by sending multiple frames appropriate to the corresponding link rate per each channel access, which is also known as Frame Concatenation [2]. With frame concatenation, a sender can send more data with the same overhead of backoff timer, RTS, CTS, and ACK frames, which significantly increases the network capacity.

Nevertheless, OAR only considers the total network capacity and does not guarantee QoS. For example, with OAR, a sender transmitting data flow packets over a high rate link will concatenate three to five back-to-back frames while another sender transmitting voice flow packets over a low rate link may only transmit one frame for each channel access. In this situation, the real-time voice flow may receive lower throughput than the best-effort data traffic, which is a violation of the principle of QoS. Therefore, how to improve OAR such that the dual objectives of high network capacity and desirable QoS can be achieved is a challenging issue.

Basically, service differentiation and admission control are two major approaches to support QoS in 802.11 networks. However, the lack of effective adaptation in existing works, such as EDCF, often limits the overall system performance. Furthermore, a comprehensive QoS solution that leverages and integrates various individual performance enhancement techniques may deliver greater performance improvement. For this purpose, we propose a solution, namely QOAR, which consists of flow weight adaptation (FWA) and admission control (AC). In FWA, senders exchange their performance and update flow weight information so that lower priority flows such as video and data reduce their flow weights adaptively to accommodate higher priority flows such as voice. Based on the updated flow weight, a sender may decide to increase its initial contention window ($CW_{\min}$) to reduce the channel access frequency and/or concatenation number (CN) to reduce the data rate, respectively. With this adaptation, data flows may still receive high throughput if real-time voice/video traffic is not present or of moderate load. On the other hand, more real-time flows can be admitted with this adaptation, thus achieving desirable QoS. In AC, performance of real-time traffic are constantly monitored and maintained. Extensive simulation results show that QOAR outperforms static differentiation based on either $CW_{\min}$ (e.g. EDCF) or CN (e.g. OAR) as well as simple admission control without adaptation.

This paper is organized as follows. Section II discusses related works in multi-rate protocols and QoS. Section III presents the problem model and analysis. Section IV improves OAR and proposes QOAR, a flow weight based adaptation and admission control scheme in multi-rate wireless LANs. Section V evaluates the proposed schemes and discusses the simulations results and Section VI concludes this paper.

II. RELATED WORKS

Multi-rate MAC Protocols: As an enhancement of the distributed coordination function (DCF) [3], Auto Rate Fallback (ARF) protocol [4] was the first commercial MAC implementation with multi-rate capability enabled. With ARF, stations attempt to use higher transmission rates after consecutive successful transmissions, which usually indicate high channel quality, and revert to lower rates after failures. Then, Holland and Vaidya proposed Receiver Based Auto Rate (RBAR) [5] where receivers measure the channel quality using the signal strength of the request-to-send (RTS) message. Based on this signal strength information, receivers choose the most appropriate transmission rate and notify the sender through the clear-to-send (CTS) message. Since the RTS is sent shortly before data transmission, the channel condition estimation is quite accurate.

Frame Concatenation: The IEEE 802.11 standard includes a "frame train" model to handle frame segmentation for large data frames. Xiao [2] proposed a frame concatenation mechanism (CM) of non-segmentation frames for throughput improvement in IEEE 802.11 wireless LANs. In this scheme, each station can send a train of "back-to-back" small frames to significantly reduce the overhead of the legacy DCF protocol and thus achieve much higher network capacity. Zhai and Fang [7] proposed a distributed adaptive packet concatenation (APC) scheme to combine several short frames into a super frame based on a coherent time defined in OAR to eliminate unnecessary protocol overheads. Both schemes and OAR show that CM is a very effective technique for improving network capacity of IEEE 802.11 networks with varying packet sizes and link rates.

Sadeghi and Kanodia [1] further extended the idea of RBAR and proposed Opportunistic Auto Rate (OAR). In OAR, each contending station potentially can transmit 1, 3, and 5 frames per successful channel access according to the dynamically-determined channel rate of 2, 5.5, and 11 Mbps, respectively. With this improvement, OAR can achieve temporal fairness and explore the potential of high data rates, thus achieving significant throughput improvements. Wang and Zhai [6] investigated the problem of multi-user diversity in which a node concurrently communicates with several neighbors and proposed a similar frame concatenation scheme as OAR to further improve network performance. However, neither OAR nor OSMA provides service differentiation or QoS guarantee for multimedia flows.

Adaptive QoS Protocols: Xiao and Li [10][11][12] proposed several adaptive QoS strategies in IEEE 802.11e wireless LANs. The core idea is to adaptively tune the initial contention windows and arbitrary inter-frame space (AIFS) based on the experienced network condition such as number of transmission failures and collision rates. Li and Prabhakaran [13] proposed a MAC layer priority re-allocation scheme where data flows and new real-time flows reduce their priorities to accommodate existing high priority real-time flows. Since contention windows are differentiated by flow priorities in 802.11e wireless LANs, the idea is similar to increasing contention windows and other parameters.

III. PROBLEM MODELING AND ANALYSIS

Although OAR improves channel access efficiency, it does not provide QoS support for real-time flows. For example, if a sender is transmitting video packets over a low rate link, e.g., 2Mbps, it can only send one frame every time it grabs the channel. However, another sender transmitting data packets over a high rate link, e.g., 11Mbps, can send as many as five consecutive frames per channel access. Therefore, even though video flows have higher priority than data flows, its perceived throughput will be much less than the corresponding data flows, which is undesirable. Therefore, the issue of QoS in multi-rate wireless LANs can be stated as follows: *given a specific network topology, how to (i) guarantee the bandwidth and delay requirements of real-time flows under various load of data flows and link rate conditions; (ii) maintain low protocol overhead and maximize network capacity.*

Let us assume that there are M active stations in the network and the perceived throughput and averaged packet delay at station i ($1 \le i \le M$) are TP_i and d_i, respectively. Also, let BW_i and D_i be the throughput and delay requirements of the flow transmitted at station i ($1 \le i \le M$), respectively. Then, the mathematical model of the above problem is

$$\max \sum_{i=1}^{M} TP_i \tag{1}$$

subject to

$$TP_i \ge \alpha \cdot BW_i \qquad \text{for } i \text{ from 1 to } M$$

$$\Pr(d_i \le D_i) \ge \beta \qquad \text{for } i \text{ from 1 to } M$$

Where α and β are the general thresholds for throughput and delay requirement respectively. Further, given a specific time period of T, per flow throughput can be expressed as

$$TP_i == \frac{K_i CN_i L_i}{T} = \frac{K_i CN_i L_i}{\sum_{i=1}^{M} K_i \cdot (CW_i + \frac{RTS+CTS}{2} + CN_i \cdot (\frac{L_i + ACK}{r}))} \tag{2}$$

Where RTS, CTS, and ACK are the additional overhead for each channel access. K_i and CW_i are the number of successful channel access and average contention window of station i, respectively. L_i is the payload of station i and r is the actual channel rate. Then, the total throughput can be calculated as

$$\sum_{i=1}^{M} TP_i = \frac{\sum_{i=1}^{M} K_i \cdot CN_i L_i}{\sum_{i=1}^{M} K_i \cdot (CW_i + RTS + CTS + CN_i \cdot (\frac{L_i + ACK}{r}))} \tag{3}$$

It is obvious that given M, rate r and payload size L_i, the network performance depends on CW_i and CN_i only. The smaller the contention window CW_i is, the higher channel access frequency K_i is for a station, provided that the flow arrival rate is the same or channel is under saturation status, thus leading to high throughput. Then, with the fixed CW_i, the higher CN_i is, the less overhead per frame transmission, thus leading higher protocol efficiency and channel utilization.

Therefore, it is preferred for all stations to maintain lowest *CW* and highest *CN* to maximize their own throughput.

However, with more number of contending stations, it is impossible to maintain the desired throughput due to excessive channel contention and collision. In this case, the total throughput will reach the maximum value and may decrease slightly. Due to this collision, per flow throughput will also decrease significantly. In addition, the average delay of each flow increases with more number of competing stations. In summary, two mechanisms must be devised to meet the dual objectives of high network capacity and desirable QoS:

- *Parameter adaptation*: low priority stations increase their *CW* and/or decrease *CN* to accommodate the performance of high priority stations, under severe channel conditions. Initially, the smallest *CW* and highest *CN* are maintained for best performance. Then, with network condition change, these parameters are tuned accordingly.
- *Admission control*: new real-time flows are rejected to protect existing flows when channel is saturated and no adaptation can be made to admit new flows.

IV. Flow weight adaptation and admission control

To enable *CW* and *CN* adaptation, we introduce the concept of flow weight. The *flow weight* φ ($0<\varphi\leq 1$) for a flow is tuned as follows:

- *Voice flows*: $\varphi=1$, i.e., voice senders do not accommodate other traffic.
- *Video flows*: $\varphi=\max(\varphi-0.1, \varphi_{min,video})$ if at least one of the voice flows require accommodation without affecting video performance, and $\varphi=\min(\varphi+0.1, 1)$ if none of the voice flows require accommodation.
- *Data flows:* $\varphi=\max(\varphi/2, \varphi_{min,data})$ if at least one of the voice or video flows require accommodation and $\varphi=\min(\varphi+0.1, 1)$ if none of the voice and video flows require accommodation.

It should be noted that video and data flows follow different adaptation patterns. The reason is that video flows only accommodate voice flows by reducing their flow weights when their QoS requirements can still be guaranteed. Thus, it is better to decrease the flow weight slowly. However, for data flows, it is not necessary to maintain QoS, so we can quickly decrease their flow weights so that voice and video flows avoid significant throughput loss. For this reason, we should set a much lower threshold for data flows than video flows when reducing the flow weights. Study of the effect of $\varphi_{min,video}$ and $\varphi_{min,data}$ is omitted due to limited space. Of course, when all real-time flows receive satisfactory throughput and delay, video and data flows can gradually increase their flow weight back until the maximum value of 1 is reached.

Based on the adapted flow weights, we can then tune *CW* and *CN* as follows. Let CW_{min} and CW_{max} be the minimum and maximum contention windows, we have that

$$CW=\min(CW_{max}, \lfloor \mathrm{CWmin}/\varphi \rfloor) \tag{4}$$

$$CN=\lceil CN_{max}\cdot\varphi\cdot CW / CW_{min} \rceil \tag{5}$$

It is clear that the initial contention window is non-decreasing according to the flow weight. If traffic load is not very high and φ is close to 1, *CW* will be slightly increased and *CN* will be maintained as CN_{max} (set by OAR) for maximum possible protocol efficiency. However, if traffic load is even higher, *CW* will reach CW_{max} and *CN* has to be also reduced to further reduce the bandwidth share of low priority flows. It should be noted that both *CW* and *CN* are determined per packet with the update of flow weights (every second in our implementation).

The idea of admission control is described as follows. Initially, every station continuously listens to the channel and thus can overhear RTS/CTS/DATA/ACK of their neighbors and obtain information in the frame headers. In this scheme, three fields, *traffic_type*, *flow_weight*, and *is_satisfied*, are included in RTS and CTS frame headers. The flag *is_satisfied* indicates whether or not the throughput and delay requirements is met according to equation (1). Based on these information, a station makes decision on whether or not it should start a new voice/video traffic as follows:

- If data traffic exist and the highest overheard data flow weight equals $\varphi_{min,data}$, reject the new voice/video flow since there is no way to accommodate new flows. Otherwise, the flow is admitted.
- If no data traffic exists and at least one voice flow is not satisfied, reject any new voice or video flows.
- If no data traffic exists and at least one video flow is not satisfied, reject any new video flows. However, we can still admit new voice flows since voice flows usually require much less bandwidth than video flows.

Of course, it is difficult to know in advance if the newly accepted voice/video flow can be supported or not. We can adopt a simple method to resolve it. After the admission, the source station of the flow monitors RTS/CTS of other stations for a short period. If at least one of the flows in the same category is not satisfied, the station drops the newly admitted flow to avoid performance effect on existing flows.

To set the *is_satisfied* flag, each station currently sending voice/video flows measures its perceived throughput by periodically counting the number of bytes successfully transmitted. Also, the percentage of data packets with delay higher than the delay bound is updated periodically (two seconds in our implementation). Then, before sending RTS/CTS, *is_satisfied* flag is set to 1 if the throughput/delay requirements are met according to equation (1). For simplicity, we refer the proposed protocol QoS OAR (QOAR).

An alternative approach of enforcing admission control is to calculate available bandwidth based on the effective link capacity and channel busy ratio. However, with differentiated contention windows and concatenation numbers, it is quite difficult to know exactly how much bandwidth is achievable for flows of different categories. Thus, we believe that measuring the throughput and delay is reasonable and effective for supporting soft QoS.

V. PERFORMANCE EVALUATION

The performance of several schemes are evaluated and compared through extensive simulations in *ns*-2 network simulator. The proposed QOAR protocol is implemented by modifying the OAR software [9] provided by Rice Network Group. In all simulations, transmission range is set to 250 meters and the link rates are 11, 5.5, and 2 Mbps for nodes within transmission range of 80, 120, and 250 meters, respectively. 50 nodes are randomly distributed in an area with size of 120×120 m^2.

Three different traffic types are involved in all simulations. Each voice stream is a CBR flow with rate of 64Kbps, generated by a constant inter-arrival time of 25ms and a fixed payload size of 200 bytes. Each video stream is traffic-shaped CBR flow with rate of 320Kbps, generated by a constant inter-arrival time of 20ms and a fixed payload size of 800 bytes. We also introduce background data traffic, which is a UDP flow with payload size of 1000 bytes and frame arrival interval of 5 ms. 25 flows are initiated every two seconds with random source and destination pairs to gradually increase the traffic load in the network. The ratio of number of voice, video, and data flows is 2:2:1. All voice and video flows last for 50 seconds and data flows lasts until the simulation finishes. The total simulation time is 150 seconds.

In the simulation, $\varphi_{min,video}$ and $\varphi_{min,data}$ are set to 0.5 and 0.02, respectively. α and β thresholds are set to 0.95 and 0.9, respectively. Also, the delay bound is 150ms for both voice and video flows. We compare OAR, the proposed flow weight adaptation (FWA) protocol, and the integration of OAR and EDCF (referred as EDCF for simplicity). For the FWA, CW_{max} is set to 31, 127, and 255, respectively. For EDCF, CW_{max} is set to 31, 63, and 127, respectively. CW_{min} is set to 31 for all protocols. The CW_{max} of FWA is set to be higher than EDCF to allow more flexibility on adaptation.

A. *Effect of flow weight adaptation*

Table I. THROUGHPUT COMPARISON OF OAR AND FWA

Throughput	Voice (Kbps)	Video (Kbps)	Data (Kbps)	Overall (Mbps)
OAR	54.93	252.02	657.21	3.90
FWA	60.24	275.97	626.87	3.82

Table I compares the averaged per flow and total throughput obtained by OAR and FWA, respectively. It can be seen that with flow weight adaptation, a slight decrease of total throughput occurs due to the fact that some low priority stations increase their initial contention windows. However, around 10% increase is achieved for throughput of voice and video flows, on average. Obviously, this improvement will be more significant with heavier traffic of voice and video flows.

Figure 1 depicts the flow weight adaptation of the first and last data flow. It can be seen that when the traffic load increases, data flow weight quickly decreases until the minimum value is reached. After most voice and video flows finish around 80 seconds, all existing high priority flows perceive satisfactory performance and all data traffic begin to increase their flow weight linearly until reaching maximum value of 1. Also, we can see that when the flow weight changes, the received throughput changes accordingly. Especially, with flow weight becoming 1, the throughput of data flows can be as high as 1.2 Mbps on average.

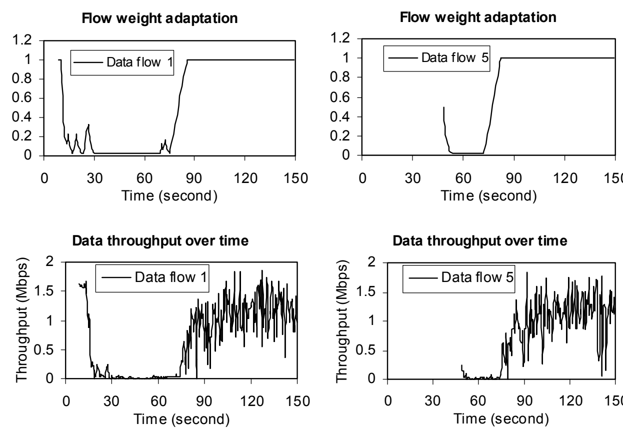

Figure 1. Performance of OAR+FWA under dynamic traffic characeteristics. Top: data flow weight over time; Right: data throughput over time.

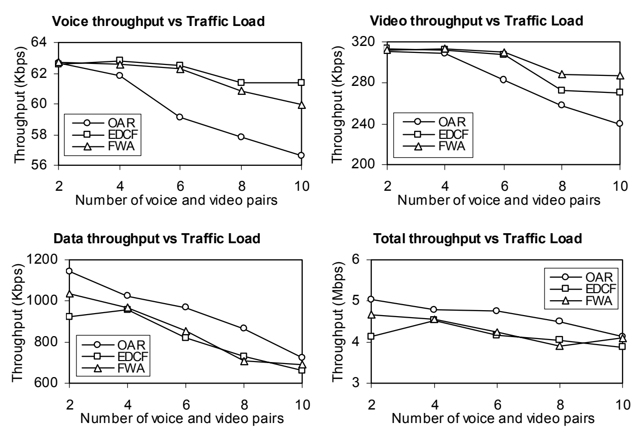

Figure 2. Throughput comparison of OAR, EDCF, and FWA under various traffic load of real-time flows.

Then, we compare the performance of OAR, FWA, and EDCF for the performance under various real-time traffic loads. In this simulation, we still maintain five data flows as described before but vary the pair of voice and video flows from 2 to 10. It can be seen from Figure 2 that OAR achieves highest total throughput since it tries to concatenate the maximum number of frames. However, with higher traffic, both voice and video flows receive much lower throughput with OAR than EDCF and FWA, which suppress data traffic by reducing the CW_{max} and/or *CN*. On the other hand, FWA outperforms EDCF under low traffic load where data traffic is not suppressed due to satisfactory performance received by voice and video flows. EDCF, however, receives low

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the ICC 2008 proceedings.

throughput with its static parameter setting. Furthermore, with adaptation, FWA achieves higher per-flow throughput for videos under high traffic load.

B. *Performance of admission control*

Finally, we evaluate the performance of admission control, i.e., FWA vs. FWA+AC (i.e., QOAR) v.s. OAR+AC. It can be seen from Table II that with QOAR, per flow throughput of voice and video traffic increases significantly to 98% and 97% of the corresponding flow rates of 64 Kbps and 320 Kbps, respectively, indicating satisfaction of QoS support. Also, compared with simple admission control (Table III), named OAR+AC, QOAR admits 3 more video flows (equivalent to 0.96 Mbps rates totally) because of the adaptation. The reason is that no consideration is given to data flow weight at the time of admission control in OAR+AC, while QOAR suppresses data traffic and allocates more bandwidth to real-time voice and video flows. From Figure 3, we can see that QOAR provides very stable throughput for both voice and video flows while OAR+AC does not due to the effect of significant data traffic.

Table II. THROUGHPUT COMPARISON OF ADMISSION CONTROL

Throughput	**Voice (Kbps)**	**Video (Kbps)**	**Data (Kbps)**	**Overall (Mbps)**
FWA	60.24	275.97	626.87	3.82
OAR+AC	61.76	307.63	998.77	4.55
QOAR	62.71	310.76	825.04	4.05

Table III. ADMISSION CONTROL RATIO

Admission Ratio	**Voice**	**Video**
OAR+AC	50%	20%
QOAR	50%	50%

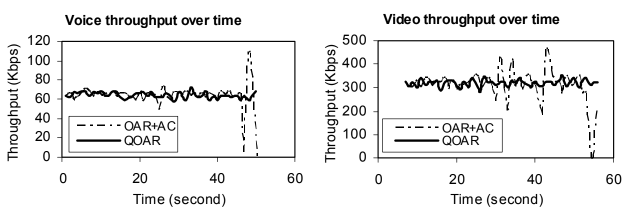

Figure 3. Throughput of admitted voice and video flows over time

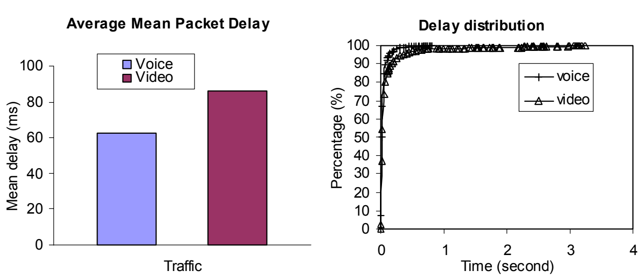

Figure 4. Packet delay with QOAR. Left: averaged mean delays; Right: delay distribution of first voice and video flows.

Figure 4 shows the averaged mean packet delays and delay distribution of voice and video flows, respectively. It can be seen that on average, with various flow weight determination schemes, voice and video flows experience packet delay of 30-90ms and 30-140ms on average, respectively. From Figure 4, we can also see that for voice and video flows, the percentages of packets experiencing delay lower than 150ms are 100% and 89% respectively. For soft QoS, this is quite satisfactory.

VI. CONCLUSTION

In this paper, OAR protocol has been enhanced to support quality of services of multimedia applications under various network conditions. The proposed QOAR protocol can adapt dynamically according to the experienced throughput and delay. Under low traffic load, data traffic take advantage of the high capacity offered by frame concatenation in OAR, while under heavy traffic load, video and data traffic may increase their contention windows and/or decrease the frame concatenation number to accommodate high priority flows. In addition, an admission control scheme based on the flow weight adaptation algorithm is proposed to provide guaranteed quality of service for voice/video flows. In the future, we will incorporate frame size in QOAR for a more comprehensive QoS solution in multi-rate wireless LANs.

REFERENCES

[1] B. Sadeghi, *et. al*, ``Opportunistic Media Access for Multirate Ad Hoc Networks," in Proc. of ACM MobiCom'02.

[2] Y. Xiao, "IEEE 802.11 Performance Enhancement via Concatenation and Piggyback Mechanisms," *IEEE Transactions on Wireless Communications*, Vol. 4, No. 5, Sep. 2005, pp. 2182- 2192.

[3] IEEE, "IEEE std 802.11 – wireless LAN medium access control (MAC) and physical layer (PHY) specification", 1997.

[4] A. Kamerman and L. Monteban, "WaveLAN II: A high-performance wireless LAN for the unlicensed band, " in *Bell Labs Technical Journal*, pages 118-133, Summer 1997.

[5] G. Holland, *et. al*, "A rate-adaptive MAC protocol for multi-hop wireless networks". In Proceedings of ACM MOBICOM'01.

[6] J. Wang, et.al, "OMAR: Utilizing Multiuser Diversity in Wireless Ad Hoc Networks," *IEEE Transactions on Mobile Computing*, vol. 5, no. 12, pp. 1764-1779, Dec. 2006.

[7] H. Zhai and Y. Fang, "A Distributed Adaptive Packet Concatenation Scheme for Sensor and Ad Hoc Networks," in Proc. of IEEE MilCom'05.

[8] E. Kim and Y.-J. Suh, "ATXOP: An Adaptive TXOP Based on the Data Rate to Guarantee Fairness for IEEE 802.11e Wireless LANs," in Proc. of IEEE VTC 2004-Fall.

[9] OAR implementation, Rice network group. Available at: http://www-ece.rice.edu/networks/software/OAR/OAR.html.

[10] Y. Xiao, F. H. Li, and S. Choi, "Two-Level Protection and Guarantee for Multimedia Traffic in IEEE 802.11e Distributed WLANs", Wireless Networks, accepted in February 2007.

[11] Y. Xiao, F. H. Li, " Local Data Control and Admission Control for QoS Support in Wireless Ad Hoc Networks", IEEE Transactions on Vehicular Technology (TVT), Vol. 53, No.5, pages: 1558-1572, September 2004.

[12] Y. Xiao, F. H. Li, "Voice and Video Transmissions with Global Data Parameter Control for the IEEE 802.11e Enhanced Distributed Channel Access" IEEE Transactions on Parallel and Distributed Systems (TPDS) , Vol. 15, No. 11, pages: 1041-1053, November 2004.

[13] Ming Li, B. Prabhakaran, "MAC Layer Admission Control and Priority Re-allocation for Handling QoS Guarantees in Non-cooperative Wireless LANs", ACM/Springer Mobile Networks and Applications (MONET), pages 947 – 959, Vol. 10, Issue 6, December 2005.