Admission Control Based on Available Bandwidth Estimation for Wireless Mesh Networks

Qiang Shen, *Student Member, IEEE*, Xuming Fang, *Member, IEEE*,
Pan Li, *Student Member, IEEE*, and Yuguang Fang, *Fellow, IEEE*

***Abstract*—Admission control plays a very important role in guaranteeing the quality-of-service (QoS) in wireless mesh networks (WMNs). In this paper, based on the channel and hidden terminal information, we first estimate the available bandwidth in the operation of medium access control (MAC). We then design an admission control algorithm (ACA) at the MAC layer to address the QoS issue for real-time and nonreal-time traffic. For real-time traffic, all nodes on a route make an admission control decision based on the estimated available bandwidth. For the nonreal-time traffic, a rate adaptation algorithm is proposed to adjust the sending rates of sources to prevent the network from entering the congestion state. Finally, through extensive simulations, we demonstrate the effectiveness of our algorithm.**

***Index Terms*—Admission control, medium access control (MAC), quality of service (QoS), wireless mesh networks (WMNs).**

I. Introduction

WIRELESS mesh networks (WMNs) have become a critical part of the future Internet. They can be widely deployed for many applications, such as campus networking, community networking, and so on, due to their fast configuration and low cost. However, the way to provide the proper QoS for multimedia traffic is an important design issue that has not been well addressed in the existing literature.

Recently, QoS provisioning for wireless local area networks (WLANs) or scheduled multihop ad hoc networks has been investigated. As we know, it is difficult to implement centralized scheduled WMNs through off-the-shelf 802.11 devices or even with some simple modifications. Furthermore, the point coordination function mode is just an optional access method [1] for such networks. Therefore, in this paper, we focus on the distributed coordination function mode instead.

QoS provisioning can be addressed in various protocol layers, such as the medium-access control (MAC) layer, the network layer, the transport layer, and so on. The IEEE 802.11e is one of the classical random MAC protocols in WLANs and WMNs, which addresses the QoS for different traffic types according to different priorities at the MAC layer. At the network layer, researchers attempt to choose the best route to accommodate new flows to address the QoS issue. For instance, the selected route should have the largest available bandwidth or minimum delay. At the transport layer, all nodes along a path must cooperate with each other to provide the proper QoS. The cooperation includes policing to decide whether the path has enough available bandwidth to accommodate new flows without affecting the QoS of other existing flows.

In addition, the aim of the QoS guarantee is to provide users with guaranteed QoS in terms of bandwidth, delay, and delay jitter. Generally speaking, bandwidth is the essential measurement of the wireless resource. Therefore, how to estimate the available bandwidth is a crucial problem that needs to be addressed. In the current literature, researchers focus on the bandwidth estimation at the network layer with the help of probing packets. Due to the dynamic characteristics of resource allocation, it is difficult to obtain the available bandwidth at the MAC layer. Fortunately, with the knowledge of the channel busyness ratio that has been proposed recently [10], [11], it is possible for a node to estimate its used bandwidth in WMNs.

Zhai *et al.* [10], [11] developed the admission and rate control scheme at the MAC layer for wireless LANs without considering the hidden terminals. In this paper, by considering the impact of hidden terminals, we develop an approach to estimating the available bandwidth for the first time in the multihop WMNs. Then, based on the estimation of the available bandwidth in the MAC layer, we propose an admission control algorithm (ACA) to support the QoS in WMNs. The ACA treats various traffic types in different ways. It provides the real-time flows with admission decisions at the MAC layer and the nonreal-time flows with sending rates. Since wireless resource is scarce, the ACA adjusts all sending rates of ongoing nonreal-time flows to maintain the QoS for the real-time flows.

The rest of this paper is organized as follows. Section II presents the related work on admission control in wireless multihop networks. Section III discusses the admission control in WMNs. Then, Section IV provides an approach to estimating the available bandwidth, and Section V describes the proposed

Manuscript received January 6, 2008; revised July 1, 2008. First published October 3, 2008; current version published May 11, 2009. This work was supported in part by the Ph.D. Innovation Fund of the Southwest Jiaotong University, by the National Science Foundation of China under Grant 60772085, by the Open Research Fund of the National Key Laboratory of Integrated Services Networks, Xidian University, under Grant ISN8-01, and by the 111 Project under Grant B08038. The review of this paper was coordinated by Prof. Y.-C. Tseng.

Q. Shen is with the Provincial Key Lab of Information Coding and Transmission, Southwest Jiaotong University, Chengdu 610031, China, and also with the National Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: q.shen@ufl.edu).

X. Fang is with the Provincial Key Lab of Information Coding and Transmission, Southwest Jiaotong University, Chengdu 610031, China (e-mail: xmfang@home.swjtu.edu.cn).

P. Li is with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: lipanleo@ece.ufl.edu).

Y. Fang is with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA, and also with the National Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: fang@ece.ufl.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2008.2006680

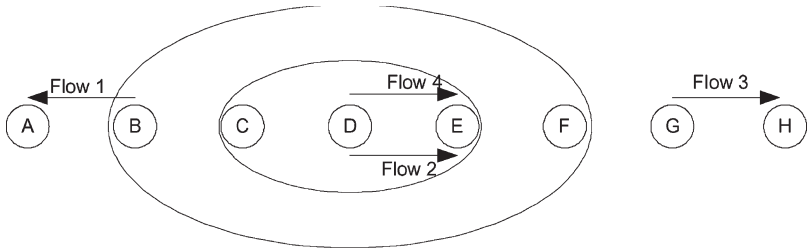


Fig. 1. Simple topology.

ACA in detail. By using simulations, Section VI shows the effectiveness of the proposed ACA. Finally, Section VII concludes this paper.

II. Related Work

Recently, a few schemes based on admission control [2]–[9] have been proposed to address the QoS issue at the network layer in wireless ad hoc or mesh networks. In INSIGNIA [3], in-band signaling allows it to quickly restore the flow state when a topology change occurs. In SWAN [4], the admission control mechanism collects the bandwidth information by a node listening to all transmissions in its transmission range. Unfortunately, the probing method introduces a lot of overhead and may not obtain an accurate value if any probe is lost. Meanwhile, Sun *et al.* [5] design an admission control mechanism considering the load at each node and the predicted delay to measure the network utilization. Luo *et al.* [6] propose an admission control to solve the QoS in a multirate environment, and it takes parallel transmissions into consideration.

The contention-aware admission control protocol (CACP) [7] and the perceptive admission control (PAC) [8] are protocols that enable a better QoS guarantee by limiting flows in the networks. However, the CACP has high overhead since a packet transmission using high power significantly affects the ongoing transmission. For the PAC, the extension of the sensing range decreases the spatial reuse, and it will result in some incorrect rejection decisions. Meanwhile, the CACP and the PAC assume that the available bandwidth has a fixed linear relationship to the idle channel time. Wei *et al.* [9] propose a call admission control method for WMNs, and their scheme is based on the interference capacity in a chain topology.

Zhai *et al.* [10]–[14] propose the original definition of the channel busyness ratio in WLANs, and based on the channel busyness ratio, they propose a call admission and rate control scheme in the MAC layer for Voice-over-Internet Protocol and the best effort traffic, respectively. In addition, Cheng *et al.* [15] adopt the channel busyness ratio concept in their framework to guarantee the accuracy of their analysis. However, the impact of the hidden terminals has not been taken into consideration in admission control schemes in all previous works.

III. Motivation

The purpose of an admission control scheme is to determine whether the available resource in a network can accommodate new flows without affecting the QoS of existing flows. This section will discuss the motivation that is related to predicting the available bandwidth, locating the hot spot, providing Diff-Service, and supporting mobility.

A. Predicting the Available Bandwidth

Prediction of the available bandwidth is a critical issue when we design our admission control scheme. Due to the nature of the wireless channel, the available bandwidth is not only determined by communications in its sensing range but also affected by communications outside its sensing range. The CACP [7] uses high power to obtain the admission information of two-hop neighbors, and the PAC [8] extends the sensing range to guarantee that there is enough resource for a new flow. Without introducing any overhead, we attempt estimating the available bandwidth according to the information that is collected by a node itself during the MAC operation.

For example, in Fig. 1, we assume that there are flows 1, 2, and 3, and each flow consumes one third of the maximum bandwidth in their sensing range when there is no hidden terminal. All nodes can only communicate with their nearest neighbors, and the sensing range is twice as much as the transmission range. When flow 4 arrives, D needs to estimate its available bandwidth. G is in E's sensing range, but not in D's sensing range. Hence, G is D's hidden terminal when it communicates with E. The available bandwidth of D is significantly reduced due to the existence of flow 3. If we do not consider this decrease, flow 4 will be accepted, which would affect the QoS of existing flows. Hence, the correct estimation of the available bandwidth is very important for the admission control scheme to make a right decision on the QoS provisioning.

Referring to [10], if we can find the accurate available bandwidth when the bandwidth varies with the different channel busyness ratio in WMNs, we can use our admission control scheme to prevent a network from entering a congestion state.

B. Locating the Admission Control

In contrast to the case in WLANs, executing an admission control only in the gateways is not enough in maintaining the QoS in potentially multihop WMNs. Instead, the admission control needs to be implemented in all hot spots, particularly at the MAC layer.

For example, in Fig. 2, there are two flows. Flow 1 is from C to gateway E, and flow 2 is from A to gateway E. We observe that the wireless resource around C is shared by six active links. Therefore, C is the hot spot in this topology. Hence, the gateway is not always the hot spot, and thus, each node in a topology needs to implement an admission control scheme.

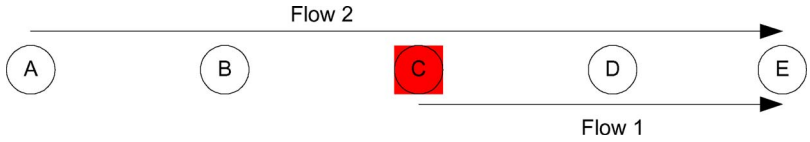


Fig. 2. Location of hot spots in a linear topology.

C. Differentiating Various Traffic

WMNs are designed to serve different traffic types, for which the admission control rules should be designed accordingly. Real-time traffic prefers the QoS in terms of bandwidth, delay, and delay jitter. In contrast, nonreal-time traffic usually demands reliability. Considering those characteristics, when the resource is not enough to accommodate a new flow, the new flow will be rejected if it is a real-time flow. However, if it is a nonreal-time flow, it can be accepted because our scheme can adjust the sending rates of all nonreal-time flows to accommodate this new flow.

D. Supporting Mobility

Mobility is another important issue that needs to be addressed in WMNs. Fortunately, the bandwidth that is allocated to the nonreal-time traffic is flexible. Hence, we can accommodate handoff flows by adjusting the sending rates of existing nonreal-time flows.

IV. Estimation of the Available Bandwidth

Since it has been shown that the channel busyness ratio provides an efficient way for estimating the resource in WLANs [10], [11], here, we attempt to study this issue in multihop WMNs.

A. Channel Busyness Ratio in WLANs

As described in [10], a slot could be an empty one, one with a successful transmission, or one with a collision. Let p_i, p_s, and p_c be the probabilities that the observed slot is one of those three types, respectively. Also, let $T_{\rm suc}$ and $T_{\rm col}$ be the average time periods that are associated with one successful transmission and one collision, respectively. In the case that a request-to-send/clear-to-send (RTS/CTS) [16] mechanism is used, we have

$$\begin{aligned} T_{\rm suc} &= T_{\rm rts} + T_{\rm cts} + T_{\rm data} + T_{\rm ack} + 3T_{\rm sifs} + T_{\rm difs} \\ T_{\rm col} &= T_{\rm rts} + T_{\rm cts\_timeout} + T_{\rm difs} = T_{\rm rts} + T_{\rm eifs}. \end{aligned}$$

In WLANs, the successful transmission probability of RTS frames (with RTS/CTS) is the same as the successful transmission probability of packets (RTS/CTS/DATA/ACK) because there is no hidden terminal. Hence, we can obtain

$$\begin{aligned} p_i &= (1-p_t)^n \\ p_s &= np_t(1-p_t)^{n-1} \\ p_c &= 1 - p_i - p_s. \end{aligned}$$

Based on previous equations, the definitions of channel idleness ratio R_i, channel busyness ratio R_b, and channel utilization R_s are as follows:

$$\begin{aligned} R_i &= \frac{p_i\sigma}{p_i\sigma + p_sT_{\rm suc} + p_cT_{\rm col}} \\ R_b &= 1 - R_i \\ R_s &= \frac{p_sT_{\rm suc}}{p_i\sigma + p_sT_{\rm suc} + p_cT_{\rm col}} \end{aligned}$$

where σ is the length of a time slot. In WLANs, the observed node can exactly distinguish the successful transmission, collision, and idle. However, in WMNs, there will be transmissions inside and outside a node's transmission range. In this case, we need to use a different approach to estimating the available bandwidth.

B. New Problems in Multihop WMNs

The derivation of the channel busyness ratio in [10] is based on the assumption that the successful transmission probability of RTS frames is the same as the successful transmission probability of packets. This assumption is valid in WLANs because all nodes are in the access point's transmission range. Whether this assumption is still suitable for WMNs is discussed here.

There are several good reasons why the successful transmission probabilities of RTS frames and packets are no longer equal in multihop networks. First, the blocking problem of the CTS transmission is obvious due to the existence of hidden terminals. In Fig. 1, G is D's hidden terminal when D is transmitting to E, while D has just finished the transmission of an RTS request. Without knowing this transmission, G initiates a new transmission, and this transmission may ruin the transmission of CTS from E to D. Consequently, there is no doubt that the successful transmission probability of RTS frames is not equal to that of packets. Second, collisions of DATA frames also need to be considered. For example, after the CTS transmission of E, G will set the network allocation vector with the length of $T_{\rm eifs}$. If the duration of the DATA frame is longer than this time, it is possible for G to initiate a new transmission when the DATA frame is still transmitting. In this case, there will be a collision at E between the DATA frame from D and the RTS frame from G. Since the duration of the ACK transmission is shorter than $T_{\rm eifs}$, there will be no collision. Finally, the influence of the retransmissions of RTS is also obvious. The retransmissions of RTS will affect the values of p_s, $T_{\rm suc}$, and $T_{\rm col}$. Consequently, the successful transmission probability of packets is affected by the successful transmission probabilities of RTS, CTS, and DATA frames. If we do not consider those impacts, the available bandwidth will be overestimated.

C. Estimation of the Successful Transmission Probability of Packets

Instead of assuming that the successful transmission probability of RTS frames is the same as that of packets, we consider all the successful transmission probabilities of RTS, CTS, DATA, and ACK frames. Let $p_{\rm srts}$, $p_{\rm scts}$, $p_{\rm sdata}$, and $p_{\rm sack}$ be the successful transmission probabilities of RTS, CTS, DATA, and ACK frames, respectively, and let $p_{\rm crts}$, $p_{\rm ccts}$, $p_{\rm cdata}$, and $p_{\rm cack}$ be the collision probabilities of RTS, CTS, DATA, and ACK frames, respectively. Also, we assume that n is the total number of nodes in the observed node's sensing range, n_1 is the number of hidden terminals, and p_t is the average transmission probability of each node. According to the characteristics of the IEEE 802.11 MAC in multihop networks, we assume that $p_{\rm scts}=1$ and $p_{\rm sack}=1$, namely, $p_{\rm ccts}=0$ and $p_{\rm cack}=0$.

Let $p_{{\rm ss}_i}$ denote the successful transmission probability of an RTS transmission in slot i. Then

$$\begin{aligned} p_{\rm srts} &= \sum_{i=0}^{r-1}(1-p_{\rm srts1})^i p_{\rm srts1} \\ p_{\rm srts1} &= \prod_{i=1}^{\lceil T_{\rm rts}/T_{\rm slot}\rceil} p_{{\rm ss}_i} \\ p_{\rm crts} &= 1-p_{\rm srts} \end{aligned}$$

where r is the $ShortRetryLimit$, which is defined as the retransmission times of RTS in 802.11, and the value in the standard is 7. Meanwhile, $p_{\rm srts1}$ is the successful transmission probability of a single RTS transmission. In addition, $p_{{\rm ss}_i}$ is shown as

$$p_{{\rm ss}_i}=\begin{cases}(1-p_t)^{n+n_1-1}, & i=1\\ (1-p_t)^{n_1}, & 2\le i\le\lceil T_{\rm rts}/T_{\rm slot}\rceil.\end{cases}$$

Generally speaking, the sensing range will not be smaller than twice the transmission range, and, in this case, the successful transmission probability of DATA frames can be derived as follows:

$$\begin{aligned} p_{\rm sdata} &= \prod_{i=1}^{\lceil T_{\rm data}/T_{\rm slot}\rceil} p'_{{\rm ss}_i} \\ p_{\rm cdata} &= 1-p_{\rm sdata} \end{aligned}$$

where

$$p'_{{\rm ss}_i}=\begin{cases}1, & \text{if } i\le\lceil (T_{\rm eifs}-T_{\rm sifs})/T_{\rm slot}\rceil\\ (1-p_t)^{n_1}, & \text{otherwise.}\end{cases}$$

Whether a packet is transmitted successfully is determined by whether all RTS, CTS, DATA, and ACK frames are transmitted successfully. Hence, we can obtain the successful transmission probability of packets, which is denoted $p_{\rm spacket}$, as follows:

$$p_{\rm spacket}=p_{\rm srts}p_{\rm scts}p_{\rm sdata}p_{\rm sack}=p_{\rm srts}p_{\rm sdata}.$$

D. Sensing-Range Bandwidth

The channel busyness ratio in WMNs is contributed by two parts: one is the transmission in its transmission range, which is the same as that in WLANs, and the other part is the transmission in its sensing region, excluding its transmission range.

TABLE I
SYSTEM PARAMETERS IN IEEE 802.11

Data rate	2 Mbps
Base Rate	1 Mbps
Backoff Slot time	20 μs
SIFS	10 μs
DIFS	50 μs
Phy header	192 bits
MAC header	224 bits
Data packet	4096 bits + Phy header + MAC header
RTS	160 bits + Phy header
CTS/ACK	112 bits + Phy header

Considering the retransmissions of RTS and the backoff time, the duration of the RTS/CTS transmission should be

$$T_{\rm rts+cts}=\sum_{i=0}^{r-1}(1-p_{\rm srts1})^i p_{\rm srts1}(iT_{\rm col}+T_{\rm rts}+T_{\rm cts}+T_{\rm sifs}).$$

Then, the successful transmission duration T_s and the collision transmission duration T_c in WMNs are as follows:

$$\begin{aligned} T_s &= T_{\rm rts+cts}+2T_{\rm sifs}+T_{\rm data}+T_{\rm ack}+T_{\rm difs} \\ T_c &= (1-p_{\rm srts})rT_{\rm col}+p_{\rm srts}p_{\rm cdata}T_s. \end{aligned}$$

Besides the above modification, R_i, R_b, and R_s should be changed as follows:

$$\begin{aligned} R_i &= \frac{p_i\sigma}{p_i\sigma+p_sT_s+p_cT_c} \\ R_b &= 1-R_i \\ R_s &= \frac{p_sT_s}{p_i\sigma+p_sT_s+p_cT_c} \end{aligned}$$

where p_i, p_s, and p_c are defined as follows:

$$\begin{aligned} p_i &= (1-p_t)^n \\ p_s &= np_tp_{\rm spacket} \\ p_c &= 1-p_i-p_s. \end{aligned}$$

Finally, the normalized bandwidth in the observed node's sensing range, denoted by s, is expressed as follows:

$$s=R_s\times T_{\rm data}/T_s.$$

With all parameters shown in Table I, we can obtain the numerical results illustrated in Fig. 3. We observe that the optimal operational points are different when n_1/n changes. The case when $n_1=0$ corresponds to what was proposed in [10].

E. Estimation Without Neighbor Information

As we know, it is difficult to determine the number of nodes in the observed node's sensing range and the number of hidden terminals. Fortunately, according to the above discussion, we can obtain that $p_{\rm sdata}$ is a function of p_t and n_1; R_b is a function of p_t, n, and n_1. If we assume that n is a constant, and $p_{\rm sdata}$ and R_b can be obtained by monitoring the communications at

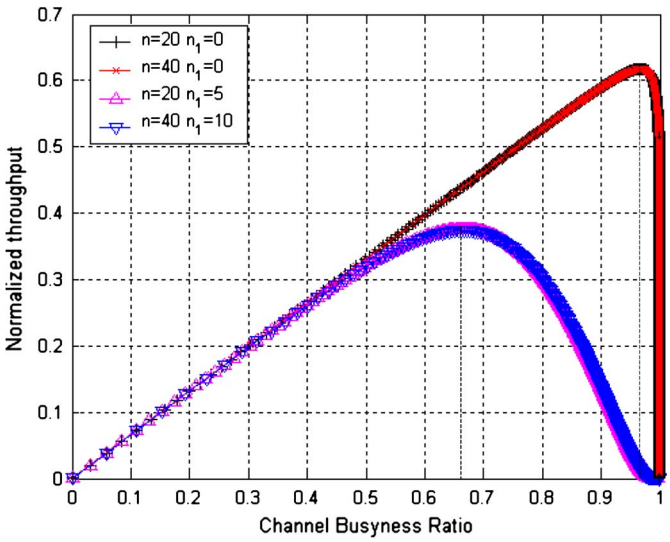


Fig. 3. Channel busyness ratio.

the MAC layer, with the information, p_t and n_1 can be calculated. As shown in Fig. 3, the normalized throughput is only related to n_1/n. Thus, we can find the maximum bandwidth, the used bandwidth, and the available bandwidth, which are denoted by $B_{\max}$, B_{use}, and B_a, respectively. For example, if $p_{\text{sdata}} = 1$ and $R_b = 0.75$, the curve we obtain is shown in Fig. 3 when $n = 20$ and $n_1 = 0$ (or $n = 40$ and $n_1 = 0$). At this time, $B_{\max} = 0.62$, $B_{\text{use}} = 0.495$, and $B_a = 0.125$.

With the consideration of the existence of hidden terminals, the curve changes dramatically. Hence, through this approach in estimating the available bandwidth, the problem described in Section III-A can be addressed more effectively.

V. Admission Control Algorithm

The ACA uses an admission control and rate adaptation scheme to tune the optimal network operating point at an unsaturated state. To address the problem described in Section III-C, the real-time and nonreal-time traffic is differently treated in the ACA.

A. Real-Time Traffic

To provide the minimum QoS requirement for the nonreal-time traffic, we need to set an upper bound, which is denoted by $B_{r_{\max}}$, for the bandwidth that is consumed by the real-time traffic. For example, we can set $B_{r_{\max}}$ to be 80% of the threshold bandwidth, which is denoted by B_{th}, and B_{th} is 85% of the maximum bandwidth. $B_{r_{\max}}$ and B_{th} are adaptable to the traffic configuration of a specific network. This way, we can guarantee that the nonreal-time traffic can occupy at least 20% of B_{th}. $(B_{\max} - B_{\text{th}})$ is reserved for the route discovery and other control messages. Furthermore, we assume that the gateway is always the source or the destination, so the proportion of various traffic can be properly handled by the gateway. Without this assumption, the gateway should perform the same as an intermediate node.

Bandwidth is the most important factor in determining whether a QoS requirement can be satisfied. Hence, three parameters, i.e., $(B_{\text{ave}_i}, B_{\text{peak}_i}, \text{Len}_i)$, are used to describe the basic resource requirement for flow i. B_{ave_i} is the average data rate of flow i, B_{peak_i} is the peak data rate, and Len_i is the average packet length in bits.

The total bandwidth that is occupied by all the admitted flows is recorded by the ACA when a flow joins or leaves a network, and the total bandwidth that is consumed by the real-time traffic is denoted by $(B_{\text{ave}_a}, B_{\text{peak}_a})$. Unfortunately, due to the route break and mobility, it is very expensive to refresh the information in time for intermediate nodes other than sources and destinations. For instance, once a link is broken, the source cannot notify the nodes between the broken link and the destination, which indicates that this flow's information cannot be used anymore. If that information is not clear, expired flows will still occupy the resource. Hence, we use $(B_{\text{ave}_a}, B_{\text{peak}_a})$ only at the gateway.

The estimation of bandwidth consumption for a single flow is another problem that needs to be addressed by an admission control scheme. We assume that the route to the destination is known before performing the admission control. When we use the *ad hoc* on-demand distance vector (AODV) as the routing protocol, we can obtain the previous hop (*prehop*), the next hop (*nexthop*), the number of hops m_1 to the source S, and the number of hops m_2 to the destination D. Since the sensing range is always between two and three times the transmission range, we can estimate the number of hops m in the observed node's sensing range in the following way:

$$\begin{aligned} &\text{if } (m_1 > 2), \quad h_1 = 2 \text{ else } h_1 = m_1 \\ &\text{if } (m_2 > 2), \quad h_2 = 2 \text{ else } h_2 = m_2 \\ &\qquad\qquad m = h_1 + h_2. \end{aligned}$$

Based on those information, the bandwidth that is consumed by flow i can be estimated as follows:

$$\Gamma(B_{\text{ave}_i}) = mB_{\text{ave}_i}.$$

In addition, B_{ave_a} is defined as the summation of $\Gamma(B_{\text{ave}_i})$ for each real-time flow in the observed node's buffer. In the same way, we can calculate B_{peak_a} corresponding to all B_{peak_i} for real-time flows.

After receiving a real-time connection request from the application layer, a node checks whether it has enough resource to establish this new flow. If so, it initiates an admission request to the destination to verify whether all other nodes on the path have enough resource to accommodate this flow.

For the ACA of all nodes except for the gateway, the admission decision must be based on the information collected by themselves. Those nodes cannot use B_{ave_a} and B_{peak_a} because they may be outdated. Instead, they use the ratio of various traffic by monitoring the channel busyness ratio during a short period. Denote R_{real} as the contribution from the real-time traffic to the channel busyness ratio. Then, we must maintain

$$R_{\text{real}} B_{\text{use}} + \Gamma(B_{\text{ave}_{\text{new}}}) \le B_{r_{\max}} \quad (1)$$

$$R_{\text{real}} B_{\text{use}} + \Gamma(B_{\text{peak}_{\text{new}}}) \le B_{\text{th}}. \quad (2)$$

If both (1) and (2) are satisfied, the ACA will locally accept this flow and then forward this request to the next hop.

Otherwise, it immediately initiates an admission reply with a rejection decision without considering the other nodes along this path. If this admission request arrives at the destination, the destination will issue an admission reply with the final decision. After receiving the admission reply, the source is notified with the admission result. If this flow is accepted, the source will send packets that are stored in its buffer for the flow to the destination. By implementing this scheme, the problem discussed in Section III-B can be addressed.

For the ACA at the gateway, after receiving the admission request, if the following constraints are satisfied, this application will be admitted locally:

$$B_{\mathrm{ave}_a} + \Gamma(B_{\mathrm{ave}_{\mathrm{new}}}) \leq B_{r_{\max}}$$

$$B_{\mathrm{peak}_a} + \Gamma(B_{\mathrm{ave}_{\mathrm{new}}}) \leq B_{\mathrm{th}}.$$

After a flow finishes the transmission, the source sends a termination request to the destination to release the resource that is allocated to this flow. Apparently, (1) and (2) depend on the accuracy of R_{real}. A larger estimation of R_{real} leads to a smaller ratio of the real-time traffic. Since it is difficult to accurately estimate R_{real}, we assume that the traffic ratio that a node monitors in its communication range is the same as that in its sensing range. For the purpose of differentiating real-time and nonreal-time packets, one reserved bit in the MAC header is used.

The observed channel busyness ratio is contributed by three parts: one from the real-time traffic with a decodable MAC header R_{b_1}, one from the nonreal-time traffic with a decodable MAC header R_{b_2}, and another one from an undecodable MAC header R_{b_3} due to various reasons such as collisions or transmissions in its sensing range, but not in its transmission range. The approximation of R_{real} is calculated as follows:

$$R_{\mathrm{real}} \approx R_{b_1} \times \left(1 + \frac{R_{b_3}}{R_{b_1} + R_{b_2}}\right) = \frac{R_{b_1} \times R_b}{R_{b_1} + R_{b_2}}$$

where we assume that R_{b_3} is composed of the real-time and nonreal-time traffic according to the traffic ratio of R_{b_1}/R_{b_2}.

B. Nonreal-Time Traffic

A rate adaptation scheme is designed for a nonreal-time flow to adjust its sending rate according to the network status. When there is a nonreal-time connection request at the beginning, the ACA is used to determine a suitable initial sending rate for the new flow.

Obviously, an initial sending rate should be determined first for a nonreal-time flow. If B_{use} is larger than B_{th} at any node on the path, namely, the node works on a saturated status, we can set the initial sending rate $B_{\mathrm{nr}_{i,j}}$, for flow i at node j, with a default value, for example, one packet per second. Otherwise, all nodes except the gateway will use the following equation to determine their local initial sending rates:

$$B_{\mathrm{nr}_{i,j}} = \begin{cases} \Gamma^{-1}(B_{\mathrm{th}} - B_{\mathrm{use}}), & if\ \Gamma(B_{\mathrm{ave}_i}) > B_{\mathrm{th}} - B_{\mathrm{use}} \\ B_{\mathrm{ave}_i}, & \text{otherwise.} \end{cases}$$

For the gateway, how to estimate its local initial sending rate is different from other nodes. The bandwidth that is consumed by the real-time traffic is $(B_{\mathrm{ave}_a}, B_{\mathrm{peak}_a})$ as mentioned before. Hence, the maximum bandwidth that the nonreal-time traffic can occupy at the gateway, which is denoted by $B_{\mathrm{nr}_{\max}}$, is described as follows:

$$B_{\mathrm{nr}_{\max}} = \begin{cases} B_{\mathrm{th}} - B_{\mathrm{peak}_a}, & \text{if } B_{\mathrm{peak}_a} < 0.8B_{\mathrm{th}} \\ B_{\mathrm{th}} - B_{r_{\max}}, & \text{otherwise.} \end{cases}$$

If the bandwidth that is consumed by the nonreal-time traffic, denoted by $B_{\mathrm{nr}_{\mathrm{con}}}$, is smaller than $B_{\mathrm{nr}_{\max}}$, all the requested bandwidth will be allocated. $B_{\mathrm{nr}_{\mathrm{con}}}$ is defined as the sum of $\Gamma(B_{\mathrm{ave}_i})$ corresponding to all nonreal-time flows in the gateway's buffer. Otherwise, we can calculate the initial sending rate, which is denoted by B_{nr_i}, for nonreal-time flow i at the gateway according to the following rule:

$$B_{\mathrm{nr}_{i,g}} = \frac{B_{\mathrm{nr}_{\max}}}{B_{\mathrm{nr}_{\mathrm{con}}}} B_{\mathrm{ave}_i}.$$

After all acceptance decisions are made, if the sending rates of other nonreal-time flows need to be adjusted, the destination will send an adjustment notification to each source with a new sending rate. This way, we can maintain rough fairness among all nonreal-time flows. The initial rate should be the minimum sending rate among local initial rates of nodes on the path from the source to the destination, i.e.,

$$B_{\mathrm{nr}_i} = \min(B_{\mathrm{nr}_{i,j}}, B_{\mathrm{nr}_{i,g}}).$$

Once the new nonreal-time flow is established, related nonreal-time flows will adjust their sending rates again if the wireless resource is tight. Namely, the initial sending rate cannot guarantee that a network operates at an unsaturated status all the time. For example, if a node discovers that B_{use} is larger than B_{th}, this node will send an adjustment notification to all known sources of nonreal-time flows.

We introduce a novel and simple way to adjust the sending rate of a nonreal-time flow as follows:

$$B_{\mathrm{nr}_{\mathrm{new}}} = \frac{R_{\mathrm{th}} - R_{\mathrm{real}}}{R_b - R_{\mathrm{real}}} B_{\mathrm{nr}_{\mathrm{old}}} \tag{3}$$

where R_b is the channel busyness ratio corresponding to B_{th}. Unlike the derivation of R_{real} in Section V-A, we use a different way to estimate it here. First of all, we give a lower bound and an upper bound of R_{real} as follows:

$$R_{b_1} \leq R_{\mathrm{real}} \leq R_{b_1} + R_{b_3}.$$

Then, to enforce a conservatively increasing and aggressively decreasing rule, we set R_{real} here as follows:

$$R_{\mathrm{real}} = \begin{cases} R_{b_1}, & \text{if } R_b \leq R_{\mathrm{th}} \\ R_{b_1} + R_{b_3}, & \text{otherwise.} \end{cases}$$

For a nonreal-time flow, all nodes on the path are qualified to send the rate adjustment to the source to decrease the sending rate. However, only the destination is entitled to send the rate adjustment to increase it. To avoid the frequent change of the

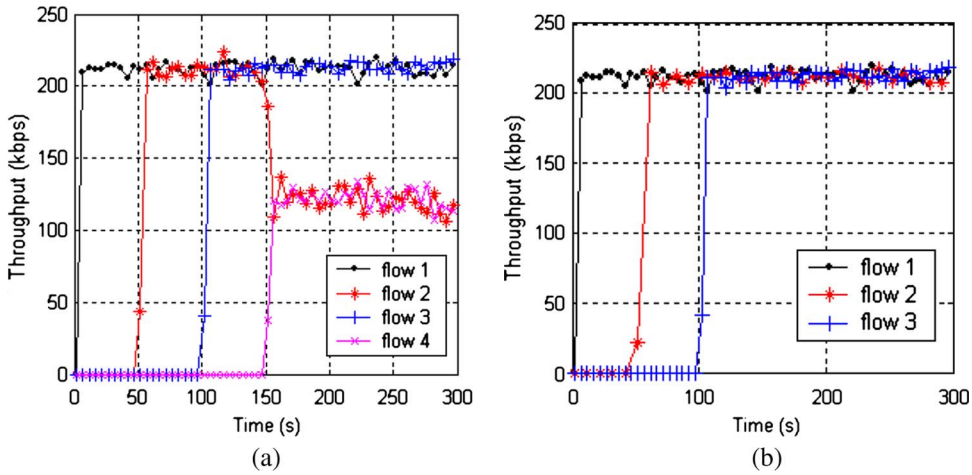


Fig. 4. Throughput for the topology in Fig. 1.

sending rate due to changes of the channel busyness ratio, this rate adaptation is made periodically, and all the parameters in (3) should be the average values over a predetermined period.

C. Mobility Support

Mobility support is another basic service that our scheme is designed to offer. When one source moves from one access point to another, it is possible that the available bandwidth is not enough to accommodate handoff flows. Hence, our resource allocation scheme should be flexible enough to accommodate handoff flows.

Compared with the mobility management in cellular systems, it is expensive to reserve excessive resource for handoff flows in resource-limited WMNs. Due to the adjustable resource allocation scheme for the nonreal-time traffic and the reserved bandwidth $(B_m - B_{\rm th})$, it is practical to provide mobility service by the ACA.

After receiving packets of handoff flows, the new access point forwards those packets to the next hop as the route indicates. After a short period, the node will notify related sources of nonreal-time flows to adjust their sending rates to accommodate those handoff flows.

VI. Performance Evaluation

In this section, we demonstrate by simulations using NS-2 that the ACA manages flows well to provide good QoS for all admitted flows. We use the IEEE 802.11 MAC as the MAC protocol. In this case, we can show how effective the ACA can be to address the QoS issue. In addition, we use the AODV as the basic routing protocol, and the admission control is incorporated with the routing protocol. The parameters are set in Table I, and others are the default settings in NS-2.

A. Effectiveness of the Bandwidth Estimation

The topology shown in Fig. 1 is used to show the effectiveness of the bandwidth estimation. We only use the real-time traffic in this scenario and assume that each of the flow consumes 200 kb/s, namely, 50 packets/s when the packet size is 512 B/packet. Meanwhile, flows from flow 1 to 4 are added to the network every 50 s.

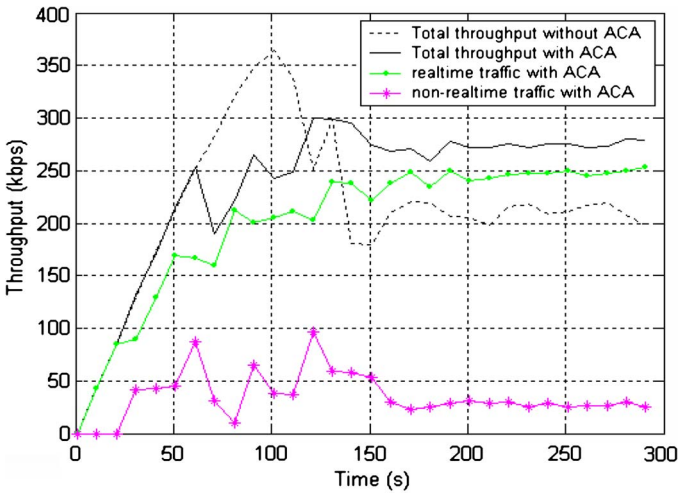


Fig. 5. End-to-end throughput in grid networks.

If we use the CACP [7], when flow 4 starts, the available bandwidth of node D is 24.7% of the channel capability, namely, 482 kb/s; it is reasonable for the network to accommodate flow 4. Fig. 4(a) shows the performance when all those four flows are accepted. Apparently, the QoS of flow 4 is very low, and the throughput for flow 2 is affected significantly. Namely, the CACP cannot provide good QoS in this situation.

In contrast, the ACA detects the existence of hidden terminals, and the maximum bandwidth that a node can achieve is much lower than that without considering the existence of hidden terminals. For example, take node D without considering the existence of hidden terminals. B_m is 62% of the basic data rate, and $B_{r_{\max}}$ is 823 kb/s at this time. By monitoring the communication of node D, the successful transmission probability of data frames is 0.68, and $B_{r_{\max}}$ is changed to 703 kb/s. We should notice that this does not mean that flow 4 can be accepted because (1) is not satisfied at this time. The reason is that retransmissions cost a lot of wireless resource, leading to a significant decrease in the available bandwidth. Hence, flow 4 is rejected, and the QoS of other flows are guaranteed. The performance with the ACA is shown in Fig. 4(b).

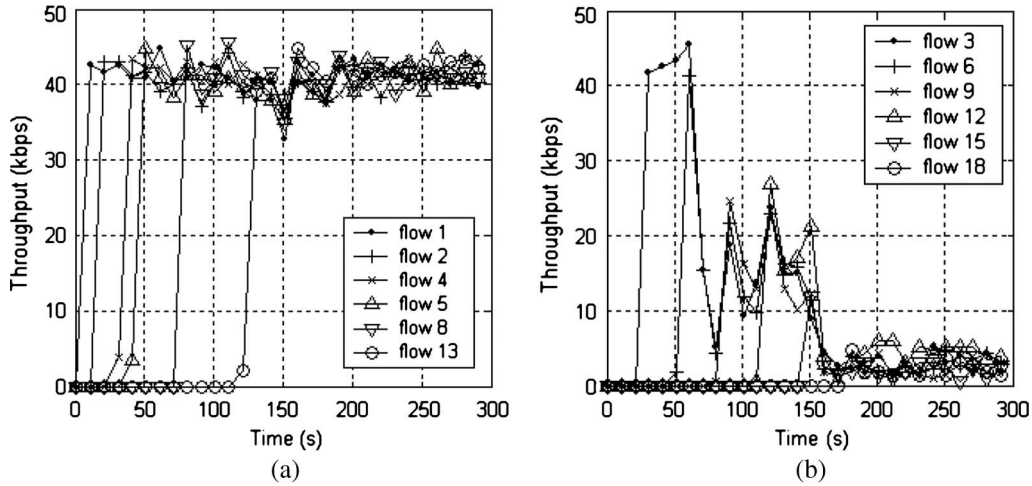


Fig. 6. Throughput in grid networks with the ACA.

B. *Effectiveness in Grid Networks*

Here, we use a grid topology to demonstrate the effectiveness of the ACA. We use a grid topology with 7×7 nodes, and nodes can only communicate with their closest neighbors. There will be a new nonreal-time flow every two real-time flows, and the sources are chosen randomly. Meanwhile, each flow consumes 40 kb/s and will be added to the network every 10 s. In addition, the total number of flows is 20.

Fig. 5 shows the comparison of the end-to-end throughput in the grid network with and without the ACA. As illustrated in Fig. 5, the throughput without the ACA aggressively increases first; however, as the number of flows increases, it significantly decreases after 100 s. In contrast, the throughput with the ACA can be kept in a stable value after it reaches a certain threshold. The curves for the real-time and nonreal-time traffic with the ACA show that the bandwidth that is consumed by the nonreal-time traffic when the number of real-time flows is small is much larger than that when the number of real-time flows is large. However, the nonreal-time traffic can still consume a certain proportion of the total bandwidth when the number of real-time flows is large.

In addition, since the original protocol prefers short-hop flows, the protocol without the ACA can achieve a high end-to-end throughput at first due to the acceptance of short-hop flows. Later, the throughput decreases due to the congestion. In contrast, with the ACA, whether a flow is accepted is only determined by whether there is enough bandwidth compared with the requirement of the new flow and its traffic type. Hence, it seems that the available bandwidth is underestimated. It is not the case.

The throughput for the real-time and nonreal-time traffic in the grid network is shown in Fig. 6(a) and (b). A stable throughput for each real-time flow is guaranteed once it is established. Note that flow 7 is rejected; however, flow 8 is accepted because the location of flow 8 is different from that of flow 7. For the same reason, flow 13 is also accepted. At the cost of better QoS for the real-time traffic, sending rates for nonreal-time flows are adjusted from time to time, and the delay of the nonreal-time traffic is larger than that of the real-time traffic.

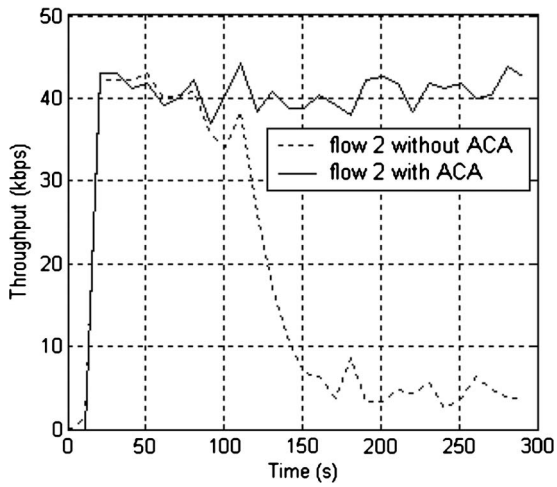


Fig. 7. Throughput for flow 2 in grid networks with and without the ACA.

The throughput curves of the real-time and nonreal-time traffic show that our rate adaptation scheme is efficient in adjusting the sending rates of nonreal-time flows to make full use of the wireless resource when the bandwidth that is consumed by the real-time traffic is smaller than $B_{r_{\max}}$.

In Fig. 7, the throughput without the ACA for flow 2 decreases when some nodes on the path work at a saturated state, for example, at 110 s. Due to the effectiveness of the ACA, each node still works on an unsaturated status. In other words, the ACA prevents the throughput from a dramatic decrease.

The delay for flow 2 in each case is shown in Fig. 8(a). As a result of competing with many later flows, the delay becomes very large after 80 s in the scenario without the ACA. It means that when nodes work close to the saturated status, the wireless resource becomes very tight; hence, the delay becomes very large. However, with the ACA, the delay can be kept in a certain range. Note that, when we use the ACA, the delay for the first few packets is very large due to the delay of the connection request and reply. In Fig. 8(b), the delay jitter with the ACA is also much better than that without the ACA.

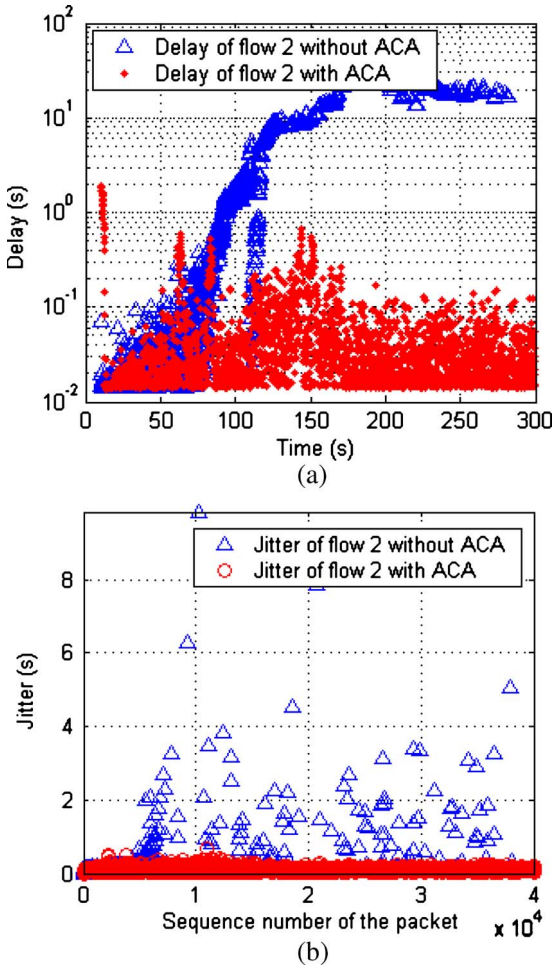


Fig. 8. Delay and delay jitter for flow 2 in grid networks with and without the ACA.

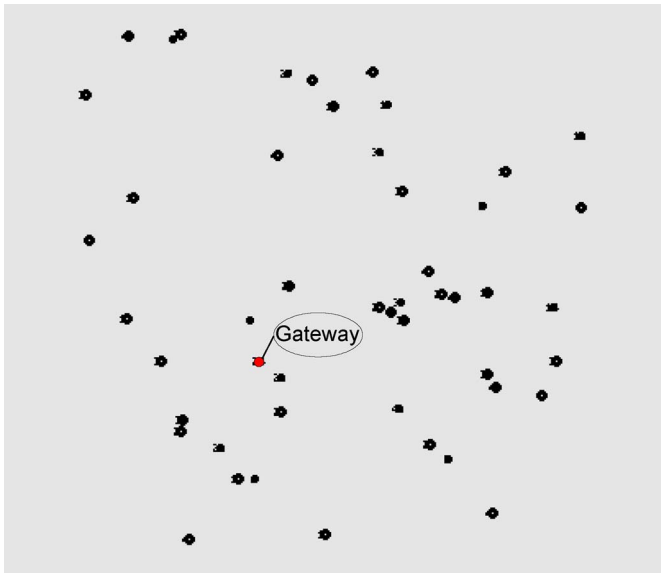


Fig. 9. Random topology.

C. Effectiveness in a Random Topology

The simulation topology is shown in Figs. 9 and 10 is the comparison of the throughput in the random topology with and without the ACA. Compared with the results shown in Fig. 5, both the throughput with the ACA in grid and random topologies can be kept in a stable value after it reaches a certain threshold. At the same time, the real-time and nonreal-time traffic can be guaranteed to be allocated with a certain proportion of the total bandwidth.

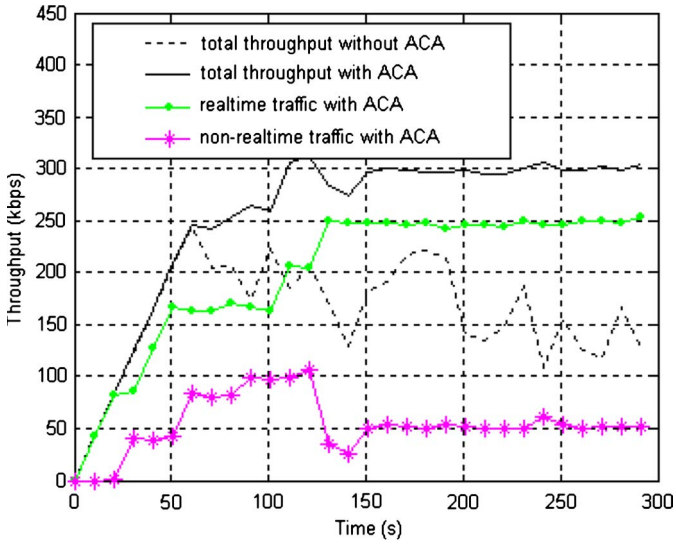


Fig. 10. End-to-end throughput in a random topology.

VII. Conclusion

Providing the QoS over multihop WMNs is always a significant challenge. The support level of the QoS really depends on the knowledge of the network resource and the traffic situation. Due to the existence of hidden terminals, the estimation of the available bandwidth is very difficult. With the extension of the derivation of the channel busyness ratio for WLANs, for the first time, we have developed a novel approach to estimating the available bandwidth by considering the impact of hidden terminals in the multihop WMNs. Based on this bandwidth estimation, we have designed a new ACA to effectively address the QoS issue for various traffic. Extensive simulation study shows that our proposed scheme indeed provides good QoS support while efficiently utilizing the residual resource for the best-effort data traffic.

References

[1] *IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, ISO/IEC 8802C11: 1999(E), 1999.
[2] S. Valaee and B. Li, "Distributed call admission control for ad hoc networks," in *Proc. VTC*, Vancouver, BC, Canada, Sep. 2002, pp. 1244–1248.
[3] S. B. Lee, G. S. Ahn, X. Zhang, and A. Campbell, "INSIGNIA: An IP-based quality of service framework for mobile ad hoc networks," *J. Parallel Distrib. Comput.—Special Issue Wireless Mobile Comput. Commun.*, vol. 60, no. 4, pp. 374–406, Apr. 2000.
[4] G. Ahn, A. Campbell, A. Veres, and L. Sun, "SWAN: Service differentiation in stateless wireless ad hoc networks," in *Proc. INFOCOM*, New York, 2002.
[5] Y. Sun, E. M. Belding-Royer, X. Gao, and J. Kempf, "A priority-based distributed call admission protocol for multi-hop wireless ad hoc networks," Univ. California, Santa Barbara, Santa Barbara, CA, UCSB Tech. Rep. 2004-20, Aug. 2004.

[6] L. Luo, M. Gruteser, H. Liu, K. Huang, and S. Chen, "A QoS routing and admission control scheme for 802.11 ad hoc networks," in *Proc. DIWANS*, Los Angeles, CA, 2006, pp. 19–28.
[7] Y. Yang and R. Kravets, "Contention-aware admission control for ad hoc networks," *IEEE Trans. Mobile Comput.*, vol. 4, no. 4, pp. 363–377, Jul./Aug. 2005.
[8] I. D. Chakeres, E. M. Belding-Royer, and J. P. Macker, "Perceptive admission control for wireless network quality of service," *Ad Hoc Netw.*, vol. 5, no. 7, pp. 1129–1148, Sep. 2007.
[9] H. Wei, K. Kim, A. Kashyap, and S. Ganguly, "On admission of VoIP calls over wireless mesh network," in *Proc. ICC*, Istanbul, Turkey, Jun. 2006, pp. 1990–1995.
[10] H. Zhai, X. Chen, and Y. Fang, "A call admission and rate control scheme for multimedia support over IEEE 802.11 wireless LANs," *ACM Wireless Netw.*, vol. 12, no. 4, pp. 451–463, Jul. 2006.
[11] H. Zhai, J. Wang, and Y. Fang, "Providing statistical QoS guarantee for voice over IP in the IEEE 802.11 wireless LANs," *IEEE Wireless Commun.*, vol. 13, no. 1, pp. 36–43, Feb. 2006.
[12] H. Zhai, X. Chen, and Y. Fang, "Improving transport layer performance in multihop ad hoc networks by exploiting MAC layer information," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1692–1701, May 2007.
[13] H. Zhai, J. Wang, and Y. Fang, "DUCHA: A new dual-channel MAC protocol for multihop ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 5, no. 11, pp. 3224–3233, Nov. 2006.
[14] H. Zhai and Y. Fang, "Distributed flow control and medium access in multihop ad hoc networks," *IEEE Trans. Mobile Comput.*, vol. 5, no. 11, pp. 1503–1514, Nov. 2006.
[15] Y. Cheng, X. Ling, W. Song, L. Cai, W. Zhuang, and X. Shen, "A cross-layer approach for WLAN voice capacity planning," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 4, pp. 678–688, May 2007.
[16] V. Bharghavan, A. Demers, S. Shenker, and L. Zhang, "MACAW: A media access protocol for wireless LANs," in *Proc. ACM SIGCOMM*, 1994, pp. 212–225.

Qiang Shen (S'07) received the B.S. degree in computer science in 2004 from Southwest Jiaotong University, Chengdu, China, where he is currently working toward the Ph.D. degree with the Department of Communication Engineering.

From April 2007 to April 2008, he was a Visiting Student with the Department of Electrical and Computer Engineering, University of Florida, Gainesville. He is also currently with the National Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, China. His research interests include cross-layer design, admission control, medium-access control, routing algorithms, and testbed development in wireless multihop networks.

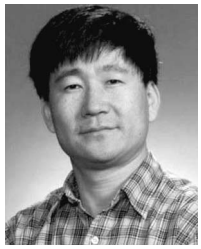
Xuming Fang (M'00) received the B.E. degree in electrical engineering in 1984, the M.E. degree in computer engineering in 1989, and the Ph.D. degree in communication engineering in 1999 from Southwest Jiaotong University, Chengdu, China.

He was a Faculty Member with the Department of Electrical Engineering, Tongji University, Shanghai, China, in September 1984. He then joined the School of Computer and Communications Engineering (currently the School of Information Science and Technology), Southwest Jiaotong University, Chengdu, where he has been a Professor since February 2001 and the Chair of the Department of Communication Engineering. He held visiting positions with the Technical University at Berlin, Berlin, Germany, in 1998 and 1999, and with the University of Texas at Dallas, Richardson, in 2000 and 2001. He has, to his credit, over 100 high-quality research papers in journals and conference publications. He has authored or coauthored five books or textbooks. His research interests include wireless mesh networks, multihop relay networks, mobile ad hoc networks, scheduling for broadband wireless access, admission control, access control, power control, flow control, cognitive radio, multihop relay cooperation, etc.

Pan Li (S'06) received the B.E. degree in electrical engineering from Huazhong University of Science and Technology, Wuhan, China, in 2005. He is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Florida, Gainesville.

His research interests include capacity and connectivity analysis, medium-access control, routing algorithms, and cross-layer design in wireless networks.

Dr. Li is a Student Member of the Association for Computing Machinery.



Yuguang Fang (S'92–M'97–SM'99–F'08) received the Ph.D. degree in systems engineering from Case Western Reserve University, Cleveland, OH, in 1994 and the Ph.D. degree in electrical engineering from Boston University, Boston, MA, in 1997.

From July 1998 to May 2000, he was an Assistant Professor with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark. In May 2000, he joined the Department of Electrical and Computer Engineering, University of Florida, Gainesville, as an Assistant Professor and then got an early promotion to Associate Professor with tenure in August 2003 and to Full Professor in August 2005. He holds a 2006 to 2009 University of Florida Research Foundation Professorship and a Changjiang Scholar Chair Professorship with Xidian University, Xi'an, China, from 2008 to 2011. He has published over 250 papers in refereed professional journals and conferences.

Dr. Fang is a member of the Association for Computing Machinery. He was a recipient of the 2006 Best Paper Award from the IEEE International Conference on Network Protocols, the 2002 Office of Naval Research Young Investigator Award, the 2002 IEEE Technical Committee on Gigabit Networking Best Paper Award at the IEEE High-Speed Networks Symposium, the IEEE Global Telecommunications Conference (GLOBECOM) Award, and the 2001 National Science Foundation Faculty Early Career Award. He is also active in professional activities. He has served on several editorial boards of technical journals, including the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, *IEEE Wireless Communications Magazine*, and *ACM Wireless Networks*. He was the Editor for the IEEE TRANSACTIONS ON MOBILE COMPUTING and currently serves on its Steering Committee. He has been actively participating in professional conference organizations such as serving as the Steering Committee Cochair for QShine, the Technical Program Vice Chair for the 2005 IEEE Conference on Computer Communications (INFOCOM), the 2004 Technical Program Symposium Cochair for IEEE GLOBECOM, and a member of the Technical Program Committee for the IEEE INFOCOM (1998, 2000, and 2003–2009).