Resource Management in IEEE 802.11 Based Wireless Networks

Ming Li
Department of Computer Science, California State University, Fresno
Fresno, CA 93740, USA
mingli@csufresno.edu

Roberto Riggio, Francesco De Pellegrini and Imrich Chlamtac
Create-Net, Via Solteri 38, 38100 Trento, Italy
{roberto.riggio, francesco.depellegrini, chlamtac}@create-net.it

1. Introduction

Next generation wireless networks will require significant flexibility on deployment so that the objective of providing seamless service to users anywhere at any time can be achieved. Having been successfully used as the ubiquitous last-mile technology in the present days, IEEE 802.11 has become the de-facto standard for providing wireless access in applications such as campus wireless LANs, wireless hotspots in hotels and airports, and military battlefields. The data rates guaranteed by the standard reaches as high as 54Mbps and 11Mbps for 802.11a and 802.11b, and current research is investigating enhanced PHY layers and data aggregation techniques in order to enhance the capacity of the WLANs channels. With this availability of broadband capability in IEEE 802.11 networks, it is a natural demand that appropriate Quality of Service (QoS) strategies are devised so that wireless users can access multimedia data such as voice, video, and 3D animation and experience satisfactory entertainment.

Depending on the actual applications and types of networks, the issues of QoS can be quite different. For instance, in single-hop wireless LANs, it is desirable to provide sufficient quality of service for real-time flows not only within each individual wireless LAN, but also among different wireless LANs when users roam in the network. Furthermore, it is not unusual that mobile users access/exchange information in wired networks. In this case, end-to-end QoS is a must in order to better support multimedia applications such as voice over IP and online video. On the other hand, in self-organizing multi-hop mobile ad hoc networks (MANETs), it is imperative that sufficient and reliable QoS is provided from the source node to the destination node, even when significant mobility and dramatic topology change is present. In addition, the concept of multi-hop ad hoc networks can be extended to wireless mesh networks where static wireless mesh routers are installed on the top of buildings to form what is called "mesh backbone". Existing wireless access networks can easily reach Internet via those wireless mesh routers. Since mesh routers can be equipped with multiple radios or multiple channels, the interference among different routers is reduced and thus the capacity of the mesh backbone is significantly increased. How to provide appropriate QoS by incorporating the multiple radio/channel capability is still a challenging issue.

Basically, there are two fundamental QoS issues in IEEE 802.11 based wireless networks. First, there is no QoS protocol in standard 802.11 MAC layer. Although IEEE 802.11e was proposed to prioritize the media access by stations of different priorities, how to improve the performance and provide QoS guarantee in 802.11e EDCF and HCF remains a challenge. Secondly, in multi-hop single channel wireless ad hoc networks, different links on the paths of the same flow may compete for the channel access and thus interfere with each other. In this case, it is difficult to provide high throughput. How to deal with this type of multi-hop interference and estimate the available bandwidth along a certain route is critically important for sufficient quality of service support.

Considerable works have been conducted on providing quality of service support to IEEE 802.11 wireless networks in recent years. Naturally, most approaches involve MAC layer protocol design with some approaches incorporate information at both MAC and network layers. From the network architecture aspects, these works can be classified to following categories:

- *Resource management in wireless LANs*: This work involves how to provide good QoS in the single hop wireless local area networks and can be further classified to further categories:
    - *DCF based protocols:* since the standard Distributed Coordination Function (DCF) only provides best-effort media access, it is important to incorporate certain distributed approaches to achieve desirable QoS guarantee and/or service differentiation in wireless LANs by appropriate modification of DCF.

On the other hand, improvement over the centralized Point Coordination Function (PCF) has also been proposed.

  - *Extension of 802.11e standard*: with the newly proposed Hybrid Coordination Function (HCF) MAC protocol in IEEE 802.11e standard, more efforts have been made on the improvement of 802.11e performance by addressing some of its shortcomings such as static parameter setting, starvation of low priority traffic, and limited QoS guarantee.

- *Resource management in multi-cell wireless LANs and heterogeneous wired/wireless networks*: this work involves how to provide good QoS when a mobile station roams among different wireless LANs. According to the heterogeneity of the network domains, we can classify it as follows:

  - *QoS handoff in multi-cell wireless LANs:* works in this direction usually check the resource availability before a station decides to move into another cell and gets associated with the corresponding access point (AP). Usually, certain bandwidth share is reserved for future traffic handoff.

  - *QoS guarantee in heterogeneous wired/wireless networks*: works in this direction usually address the issue of how to integrate the QoS strategies in wireless LANs such as Extended Distributed Coordination Function (EDCF) and other resource reservation protocols with QoS frameworks in wired Internet such as differentiated service (DiffServ) and Resource ReserVation Protocol (RSVP), thereby providing an end-to-end QoS guarantee for data access that spans over both wired and wireless network domains.

- *Resource management in wireless mobile ad hoc networks*: as mentioned earlier, multi-hop interference is a major issue for QoS in wireless ad hoc networks. Most routing protocols try to find routes that are better than shortest path. However, how to provide service differentiation or service guarantees in this type of network is very important for the support of multimedia applications. Protocols in this direction usually involves route available bandwidth estimation by incorporating interference and traffic information, devise of QoS aware routing protocols by piggybacking QoS information along with route discovery and route collection, and design of flow reservation frameworks.

- *Resource management in wireless mesh networks*: recently, the concept of wireless mesh network (WMN) has been clearly differentiated from the traditional ad-hoc network concept, becoming a strategical solution due to the drop of the cost for wireless mesh routers and the availability of novel softwares that are able to drive such devices. In practice, wireless mesh networks are configured as the very flexible extension of customary Ethernet LAN networks, and are particularly suitable in order to cover metro areas in a multi-hop fashion: the quest for QoS, then, come naturally from the need of uniformity with the wired trunks. Recent literature has focused on the design of wireless backhauls in such multi-hop scenario, and on related resources optimization problems; also, the problem of the channel allocation has to be tackled in the wireless mesh network domain. Furthermore, quite novel scenarios, which resemble similar issues arising in the peer-to-peer networks, are emerging in the field of community wireless mesh networks, where network design is performed on a best effort manner, but, the final resource utilization must be regulated through cooperation-enforcing mechanisms.

In this chapter, we focus on the resource management in IEEE 802.11 based wireless networks: due to the broadness of the topic of QoS and wireless technologies themselves, it is impossible to cover all aspects of QoS issues in many other wireless technologies such as cellular networks, UMTS, Bluetooth, and Wi-Max. Furthermore, with a less relevance to quality of service support, issues on capacity improvement such as using different routing protocols or topology control strategies will not be discussed in details. Instead, issues on resource management such as bandwidth provisioning, service differentiation, and admission control are given more attention. This chapter is organized as follows. Section 2 discusses representative QoS solutions based IEEE 802.11 DCF, EDCF, and HCA. Section 3 discusses the resource management schemes in multi-cell WLANs and heterogeneous wired/wireless networks. Section 4 discusses QoS framework, interference aware bandwidth estimation, and mobility aware QoS routing protocols in wireless mobile ad hoc networks. Section 6 describes issues of resources allocation and QoS in WMNs. Finally, Section 7 concludes the chapter with future trends.

2. Background and Motivation

Compared with a wired infrastructure, wireless LAN (WLAN) has unique advantages, such as broadband bandwidth capability and low deployment cost. Thanks to the technology provided by IEEE 802.11, the wireless LAN market is experiencing explosive growth in hot spots such as hotels, hospitals, and campuses, to mention just a few. With

WLANs being deployed in an unlimited way as access points, wireless users can access real-time and Internet services virtually anytime, anywhere, while enjoying the flexibility of mobility and guaranteed connectivity. IEEE 802.11 is the lead standard for wireless LAN. It adopts the standard 802 LLC (Logical Link Control) protocol but provides optimized PHY (Physical Layer) and MAC (Medium Access Control) sub-layers for wireless communications. 802.11 specify two physical layers: DSSS (Direct Sequence Spread Spectrum) and FHSS (Frequency Hopping Spread Spectrum). In this section, major MAC protocols in IEEE 802.11 standard are introduced.

2.1. Distributed Coordinator Function (DCF) and Point Coordinator Function (PCF)

Distributed Coordinator Function (DCF) is the basic medium access mechanism in IEEE 802.11 [81]. DCF is contention-based and it uses Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) algorithm to coordinate the access to the wireless channel. To resolve the hidden terminal problem, a Request-to-Send/Clear-to-Send (RTS/CTS) handshaking procedure is required to detect the transmission collision. Before a station (STA) sends out a data frame, it first senses the channel. If the channel is idle for at least a DCF inter-frame space (DIFS), the frame is transmitted. Otherwise, a backoff time slot is chosen randomly in the interval [0, *CW*), where CW is the contention window. The contention window is incremented exponentially with the increase of the number of attempts to retransmit the frame. During the backoff period, the backoff timer is decremented in terms of slot time as long as the channel is determined to be idle. When the backoff timer reaches zero, the data frame is sent out. If collision occurs, a new backoff time slot will be chosen and the backoff procedure starts over until some time limit is exceeded. Following the RTS/CTS, the sender STA waits for short inter-frame space (SIFS) and then sends out the data packets (DATA). Upon the receipt of the DATA, an ACK is sent from the receiver to acknowledge the success of the data transmission. After the successful RTS/CTS/DATA/ACK four way handshaking, the contention window is reset to CW_{min}. If the data packet is larger than the MAC threshold *Frag_threshold*, multiple fragments are transmitted and immediately acknowledged separately. Furthermore, a Network Allocation Vector (NAV) is set in RTS, CTS, and DATA. The value of NAV is the amount of the channel time the ongoing data transmission *will* need for completion. When another STA hears any of these frames, it defers its channel access only after a time period equivalent to NAV set in the frame, thus reserving the channel for the existing data transmission. Figure 1 illustrates the CSMA/CA channel access procedure with RTS/CTS and fragmentation. DCF suffers from collision seriously under high loads, and it does not provide any traffic differentiation and quality of service.

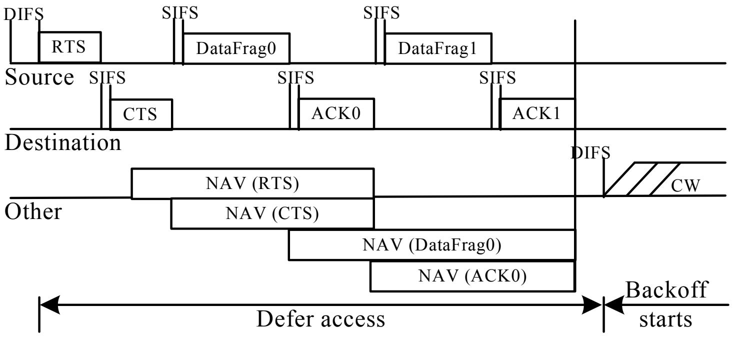

Figure 1. CSMA/CA-RTS/CTS with fragmentation access scheme

Point Coordinator Function (PCF) is an optional mechanism for IEEE 802.11. PCF coexists with DCF by providing a Contention Free Period (CFP), during which the Point Coordinator (PC) polls high priority stations and allocates time slots for them to transmit data frames. A STA is not allowed to transmit data packet without the permission from the PC. PCF Inter-frame Space (PIFS) is defined to make sure that low priority STAs do not interfere PCF operation. Also, DCF is supported in this case to prevent low priority stations from being starved. PCF is designed to offer QoS for real-time applications. But it is a centralized approach and suffers from location-dependent errors.

2.2. Hybrid Coordination Function (HCF)

Unlike IEEE 802.11, IEEE 802.11e standard [82] proposed a single coordination function called Hybrid Coordination Function (HCF), which combines two mechanisms: contention based enhanced distributed channel access (EDCA) and non-contention based HCF controlled channel access (HCCA). Basically, EDCA and HCCA are extensions of DCF and PCF, respectively and EDCA is also refereed as Enhanced Distributed Coordinator Function (EDCF).

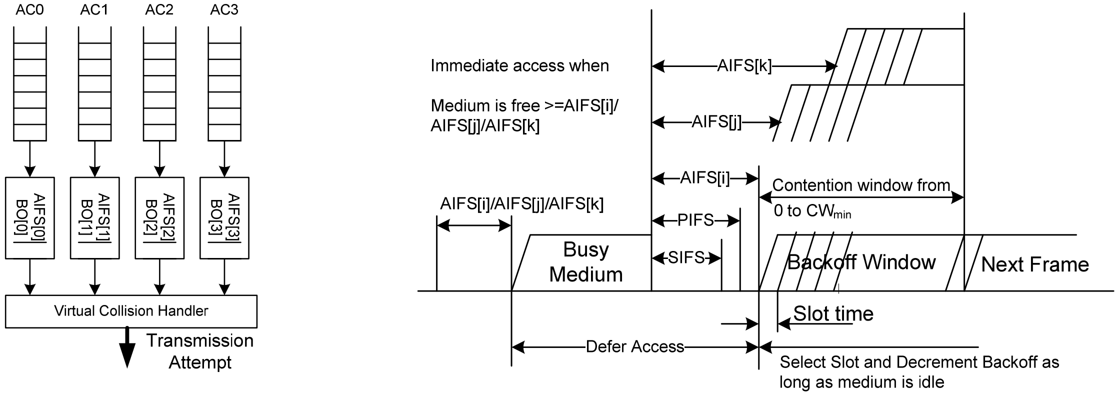

Figure 2. Channel access of EDCA [74]. (a) Virtual Collision handling; (b) Timing diagram

In EDCA, differentiated DCF access is provided to the wireless medium for prioritized access categories (ACs). As shown in Table 1, an EDCA station can implement at most 8 prioritized output queues, mapping to four different AC. Values for CW_{min}, CW_{max}, and arbitrary inter-frame space (*AIFS*) are set on per-AC basis. For AC i, the initial minimum contention window, maximum contention window, and arbitrary inter-frame space are $CW_{min}[i]$, $CW_{max}[i]$, and $AIFS[i]$, respectively. To make sufficient service differentiation, for AC i and j with $0 \leq i < j \leq 3$, it follows that $CW_{min}[i] \geq CW_{min}[j]$, $CW_{max}[i] \geq CW_{max}[j]$, and $AIFS[i] \geq AIFS[j]$, and at least one being "strict greater". Each station sends packets with its preferred priority, which is then mapped to a corresponding AC. By specifying smaller CW_{min} and *AIFS* values for higher priority queues, delay and throughput of high priority flows can be ensured while minimal service is offered to lowest priority queues. Virtual collisions (Figure 2a) between competing queues within a station are resolved by granting the transmission opportunity to the highest-priority queue involved in the collision. A virtual collision happens when the backoff intervals of two or more queues within one station counts to zero at the same time. Table 1 shows the time diagram of EDCF with three ACs: $i > j > k$. Furthermore, a new concept, transmission opportunity (TXOP) is introduced in 802.11e to represent a time period when a station wins the channel access for data transmission.

Table 1. Priority to Access Category Mapping

Priority	Access Category (AC)	Traffic type
7	3	Voice
6	3	Voice
5	2	Video
4	2	Video
3	1	Video Probe
2	0	Best Effort
1	0	Best Effort
0	0	Best Effort

Similar to PCF, HCCA provides a central control based channel access through polling. A QoS aware hybrid coordinator (HC) is defined at the QoS access point (QAP) and uses PIFS to gain control of the channel and allocate TXOP to QoS stations (QSTAs). An interesting feature of HCCA is that it can poll stations not only during contention-free period, but also during contention period. In order to appropriately allocate TXOP, HC requests a QSTA to send a QoS reservation request with its traffic specification (TSPEC) parameters and then determine the corresponding TXOP. The major parameters [24][82] included in TSPEC are:

- *Mean data rate* (ρ): the average bit rate for packet transmission, in bits per second;
- *Delay bound* (D): the maximum delay allowed to transport a packet across the wireless interface (including queuing delay), in milliseconds;
- *Maximum service interval* (SI_{max}): the maximum time allowed between neighbor TXOPs allocated to the same station, in microseconds;

- *Nominal MSDU size* (L): the nominal size of a packet, in octets;
- *Minimum PHY rate (R)*: the minimum physical bit rate assumed by the scheduler for calculating transmission time, in bits per second.

The service differentiation in EDCA is helpful in providing better-than-best effort quality of service for multimedia data traffic under low to medium traffic load condition. However, it does not perform well under high traffic load condition. In this case, admission control and bandwidth reservation becomes a must in order to guarantee QoS of existing traffics. Otherwise, the extremely large saturation delay may lead to the failure of supporting multimedia applications. On the other hand, HCCA provides traffic protection and QoS guarantees, but it is a centralized approach. Thus, there are two major research directions on the topic of QoS in wireless LANs: (i) Modification or extension of DCF/EDCA for distributed QoS support; (ii) Improvements of HCCA for better QoS support. Basically, the approaches in the former category dominate in the literature.

However, unlike wired networks, in IEEE 802.11 wireless networks, a station has no knowledge of the network resource availability and thus cannot make accurate decision on whether or not to admit a new flow. In addition, with the contention-based CSMA/CA channel access mechanism, bandwidth provisioning is almost impossible, leading to only soft QoS guarantee. Due to these two major difficulties, admission control and bandwidth reservation in IEEE 802.11 wireless network is quite difficult.

3. Resource Management in Wireless LANs

Based on the ways the resource management issue is approached and addressed, we can categorize existing schemes to three general categories:

- ***DCF based schemes*** [6][8][15][23][36][61][69][78]***:*** These schemes focus on how to provide sufficient quality of service without change of the legacy DCF/PCF scheduling algorithms.
- ***EDCF based schemes*** [7][41][43][58][72][74][73]***:*** These schemes improve the performance of real-time flows in addition to the service differentiation and QoS strategies offered by the legacy EDCF/HCF scheduling algorithms.

3.1. DCF based QoS MAC schemes

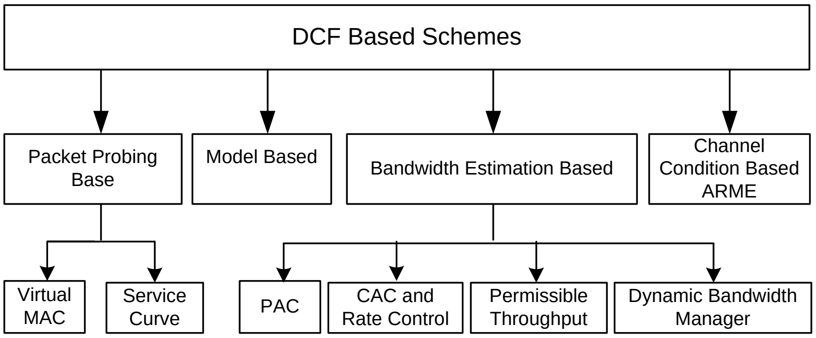

Figure 3. Taxonomy of DCF based QoS MAC schemes

Figure 3 shows the taxonomy of DCF based QoS MAC schemes. The common features of these schemes is to devise additional techniques to estimate the network resources such as bandwidth and delay and then enforce QoS strategy such as admission control and flow reservation to ensure the performance of real-time voice/video traffic. According to the techniques applied, these approaches can be further classified to following categories: *packet probing based approaches*, *calculation based approaches*, and *bandwidth estimation based approach*.

3.1.1. Calculation Based Approaches

Performance modeling of IEEE 802.11 DCF has been extensively investigated [9][70][71][79]. These models are based on the saturation status where all stations always have packets to send in the queue. However, admission control should be enforced before saturation status is reached to avoid performance degradation of real-time traffic. Ergen and Varaiya [23] proposed a Markov model for DCF that incorporates carrier sense, non-saturated traffic and SNR, for both basic and RTS/CTS access mechanisms. Then, based on the performance model, a system throughput can be calculated by solving a set of non-linear equations, given the assumption that all the flows are admitted.

Although significant contribution is made towards the performance modeling, especially on analysis of the non-saturation cases, using this approach for admission control has its own limitations. First, when the number of flows increases and traffic changes dynamically, the overhead of calculating system throughput is non-negligible. Second, the performance of a real network is far from ideal and the performance model does not adhere to the real conditions: In general, it is very difficult to obtain an accurate throughput with performance analysis. Due to these limitations, calculation based approaches are usually not widely used for practical QoS protocol design. Instead, measurement based approaches have the advantage of being simple, dynamic, and practical.

3.1.2. Channel Condition Based Approaches

Banchs and Pérez [6] proposed Assured Rate MAC Extension (ARME) to provide assured service to multimedia applications. The core idea of ARME is to use a token bucket and a queue to change the contention window (CW) and regulate the flow transmission. The token bucket is filled at the desired transmission rate and the queue expresses the willingness of a station to transmit packets. If queue is empty, i.e., no packet to fill and the current CW satisfies the sending needs of the user, CW is increased slightly. Otherwise, CW is decreased so that a station has more frequent channel access to increase its transmission rate. On the other hand, if the length of the packets in the token bucket is smaller than a certain limit, then the CW can be increased slightly. However, it is possible that all nodes try to decrease their CWs to meet their bandwidth requirements. In this case, the channel overloading occurs due to significant channel contention. To avoid this scenario, the average number of collision per packet transmission is measured. If the collision number is greater than a threshold c, then the CW is increased over the specified value by IEEE 802.11 DCF.

3.1.3. Packet Probing Based Approaches

Barry and Campell [8] proposed a Virtual MAC (VMAC) algorithm that passively monitors the channel by using virtual MAC frames and estimates local service level. From the observation that the wireless channel usually becomes delay limited before getting throughput limited, authors propose to estimate the MAC level delay to sense the channel condition as the basis for admission control. The idea of VMAC is that in order to accurately know how good the channel condition is, a virtually created packet is sent from the higher layer so that it will go through the channel contention and backoff procedure exactly like a real data packet, except that the virtual packet is not sent out through the air interface. Thus, VMAC does not impose traffic overhead to the network and can be executed continuously in parallel to the real applications without triggering network congestion.

VMAC works as follows. First, virtual packet is time-stamped at MAC layer and put into a virtual buffer. Then, if "virtual collision" is detected, i.e., whenever any other mobile station chooses the same slot for transmission, it will wait for a delay equal to RTS timer and then enters a backoff process. In this way, VMAC can measure MAC packet delay, loss, and collision, and can provide a very reasonable basis for admission control. Then, a Virtual Source (VS) algorithm is proposed to measure the application layer delay. Whenever a new flow is initiated, its delay requirement is compared with the measured delay. If delay requirement is satisfied, the flow is admitted, otherwise, it is dropped. With this technique, VMAC and VS work well for delay constrained voice traffics.

Instead of using virtual packet, Valaee and Li [69] proposed to use a sequence of small size probing packets to measure the probing delay and establish the service curve, which is then used as the criteria for call admission control. The probing is a greedy manner, i.e., a new probing packet is generated once an existing probing packet is transmitted. Thus, the total delay of ith probing packet, δ_i, is approximately τ_i-$\tau_{i-1,}$ where τ_i is the time instance at which a probing packet is delivered to the destination. On the other hand, δ_i includes a waiting time in the queue w_i and the actual transmission time b that includes the duration of multiple MAC frames. Thus w_i=δ_i-b= δ_i-T_{DIFS}-3T_{SIFS}-T_{RTS}-T_{CTS}-T_{ACK}-T_{DATA}. Thus, a sequence of total waiting time is used to obtain the service curve $S_{\text{€}}(t)$, which is defined as a percentile of the delay elements. Finally, call admission control (CAC) algorithm can accept a flow if the induced service curve will stay above the universal service curve and reject it otherwise.

3.1.4. Bandwidth Estimation Based Approaches

Probing packet usually involves algorithmic overhead [8] or traffic overhead [69]. Furthermore, the accuracy of the admission control decision largely depends on the packet sizes and frequency of the probing, especially when the network traffic is dynamic. Thus, many proposed schemes are designed to monitor the wireless channel, measure the channel utilization (or the busy ratio) and/or use existing traffic to directly estimate available bandwidth.

Chakeres and Belding-Royer [15] proposed PAC, a simple approach to estimate the channel utilization. In this approach, by monitoring the amount of time the channel is sensed busy, sending, or receiving, a node can measure not only transmissions that occur within its reception range (or transmission range), but also those within its carrier sensing range. Then, the network available bandwidth B_{avail}=(1-U)·B_{max}, where U is the estimated channel utilization and B_{max} is an approximation of channel maximum achievable bandwidth. For 2Mbps channel rate, B_{max} is set to 1.2Mbps. Furthermore, to prevent channel congestion and make sure network is running under saturation status, a small portion of the bandwidth B_{rsv} is reserved. Then, a new flow is only admitted if $R_f < B_{max}$-B_{rsv},where R_f is the flow rate requirement. With appropriate utilization estimation, satisfactory performance can be obtained for real-time multimedia applications. However, simply assuming a constant B_{max} may not be very desirable. With many factors such as network configuration, traffic characteristics, network topology, and mobility, the maximum achievable bandwidth varies significantly [9].

Zhai and Chen [78] proposed two algorithms to provide QoS in IEEE 802.11 wireless LANs. In their approach, where call admission control (CAC) is used to restrict number of real time flows and rate control scheme is to regulate the transmission rates of best effort flows. The core idea of the CAC is to measure the channel utilization and then calculate the available bandwidth. Let B_U be the channel utilization corresponding to the optimal operating point and R_b be the channel busyness ratio, then the available *normalized* throughput s_a=(B_U-R_b)·T_{data}/T_{suc}, where T_{data} is the time taken to transmit the data packet and T_{suc} is the average time period associated with one successful transmission (defined as T_{data}+T_{ACK}+T_{SIFS}+T_{DIFS} without RTS/CTS). However, this estimation is too optimistic. Basically, due to overhead and collision, it is impossible to achieve normalized total throughput as high as T_{data}/T_{suc}, even with B_U considered (set as 0.9 without RTS/CTS and 0.95 with RTS/CTS). Thus, this scheme and PAC do not provide an accurate and practical method to obtain the maximum achievable channel bandwidth.

Kazantzidis and Gerla [36] proposed a measurement based approach to estimate the available bandwidth. Instead of using virtual packets or probing packets, existing traffic is used to eliminate the extra overhead and algorithm complexity. In this approach, a *permissible throughput* is calculated as

$$r_{(source,destination)} = (1-u) \cdot \frac{S}{t_q + (t_s + t_{CA} + t_{overhead}) \cdot R + \sum_{r=1}^{R} B_r}$$

Where u is the link utilization, t_q is the MAC queuing time, t_s is the transmission time of S bits, t_{CA} is the collision avoidance phase time, $t_{overhead}$ is the total control overhead time such as *RTS*, *CTS*, *ACK*, etc, R is the number of necessary transmissions due to collision and loss, and B_r is the backoff time for rth transmission. As an approximation, a constant of 1200 μs is used for $t_{overhead}$ for simplicity. A window of 16 to 32 packets is used to smooth the estimation. However, compared to t_s, t_{CA} may not be negligible and is independent of S. Thus, this estimation scheme is affected by packet size.

Shah and Chen [61] proposed an admission control and dynamic bandwidth management scheme to provide fairness and a soft rate guarantee in single hop ad hoc networks. In this approach, a cross-layer architecture is proposed to deal with the bandwidth estimation, allocation, and adaptation of real-time flows. The architecture has the following three components: *Rate Adaptor* (*RA*), *Total Bandwidth Estimator* (*TBE*), and *Bandwidth Manager* (*BM*).

- *Rate Adaptor* (*RA*) converts a flow's bandwidth requirements into channel time percentage (CTP) requirements and communicates it to the bandwidth manager (BM). After RA obtains an allocated CTP for the flow, it will control its transmission rate according to the allocated CTP. Let the minimum bandwidth of a flow f be $B_{\min}(f)$ and the total bandwidth perception by flow f be $B_p(f)$, then the corresponding minimum CTP of f, $p_{\min}(f)$=$B_{\min}(f)$/$B_p(f)$..
- *Total Bandwidth Estimator* (*TBE*) takes advantage of the MAC layer information and makes the estimation on the total channel bandwidth. Similar to the measurement technique proposed by [36], the throughput of transmitting a packet is calculated as $TP = S/(t_r - t_s)$, where S is the size of the packet, t_s is the time-stamp that the packet is ready at the MAC layer, and t_r is the time-stamp that an ACK has been received, as illustrated in Figure 4. This time interval t_r–t_s includes the channel busy and contention time. The measurement is made for different neighboring nodes due to channel condition variation and only on active links. For inactive links, a default initial bandwidth can be set since there is no way to measure t_r–t_s over non-transmitting links. Since the packet size affects the accuracy of the bandwidth estimation, authors proposed to normalize the estimated TP value by incorporating packet size. In Figure 4, we can that T_d=S/BW_{ch} is the actual time for the channel to transmit the DATA packet, where BWch is the channel's bit-rate. Then, the transmission times of two packets should differ only in their times to transmit the DATA packets. Therefore, it is obvious that $(t_{r1}–t_{s1})–S_1/BW_{ch} = (t_{r1}–t_{s1})–S_2/BW_{ch} = S_2/TP_2–S_2/BW_{ch}$, where S_1 is the

actual data packet size, and S_2 is a pre-defined standard packet size. With this equation, normalized throughput $TP2$ can be calculated from a standard size packet.

- *Bandwidth Manager* (*BM*) performs admission control at two different times: (1) when a new flow is established; (2) when an existing flow is torn down. When a new flow f is introduced and requests its $p_{\min}(f)$ from RA to BM, BM checks if $1-\sum_{g\in F} p_{\min}(g) \ge p_{\min}(f)$. If this is true, the new flow is admitted. Otherwise, it is rejected with a reply to the RA at the sender. After the admission, BM must redistribute free channel time $1-\sum_{g\in F} p_{\min}(g)$ according to a min-max fair fashion to maximize the performance of existing flows. When a flow f terminates, the BM will eliminate the flow from the flow list, calculate the total channel time percentage, and redistribute the free channel time to existing flows through the RA.

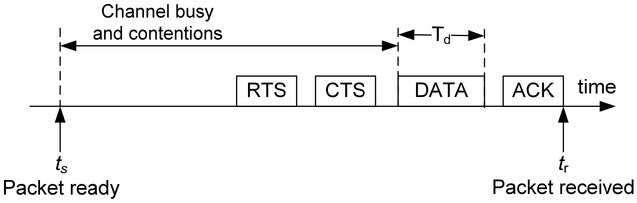

Figure 4. IEEE 802.11 unicast packet transmission sequence.

3.2. EDCA based call admission control schemes

With EDCA being proposed in IEEE 802.11e standard, many research works have concentrated on the resource management issues in term of how to dynamically improve the bandwidth sharing and delay guarantee by tuning the network configuration parameters. These schemes have the desirable features of good compatibility and thus are more practical. Figure 5 shows the taxonomy of EDCA based QoS MAC schemes. According to the techniques applied, these approaches can be further classified to following categories: *priority reallocation and admission control*, *CW and TXOP adjustment based approaches*, and *data and admission control approaches*.

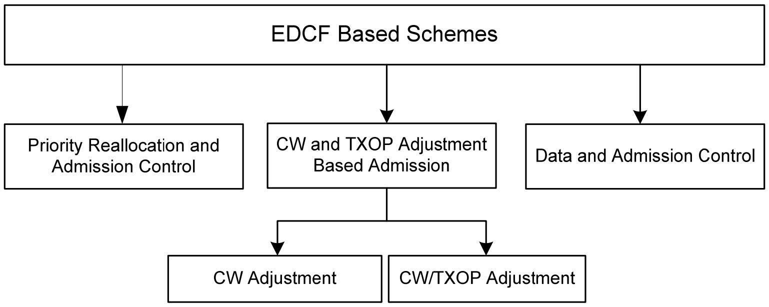
EDCF Based Schemes
Priority Reallocation and Admission Control
CW and TXOP Adjustment Based Admission
Data and Admission Control
CW Adjustment
CW/TXOP Adjustment

Figure 5. Taxonomy of EDCA based QoS MAC schemes

3.2.1. Priority Reallocation and Admission Control

Li and Prabhakaran [41] proposed a general admission control and flow reservation strategy for distributed MAC schedulers in IEEE 802.11 wireless LANs such as DCF, EDCF, and DFS [68]. On the one hand, the approach is general since the flow reservation and bandwidth estimation work for all existing schedulers. On the other hand, they proposed a priority re-allocation scheme to improve the capacity of EDCF in the infrastructure mode.

The total bandwidth estimation technique is the same as the one proposed in [36] and [61]. However, the flow reservation scheme ensures better QoS guarantee than the dynamic bandwidth management in [61]. To facilitate easy implementation and better compatibility, authors adopted a simple request/response pattern for flow reservation and admission control. No bandwidth re-negotiation is enforced because soft-QoS is offered. Since each wireless station does not have global information of the LAN, a Wireless Bandwidth Manager (WBM) is specified for admission control and flow reservation. Every high priority mobile stations, before data transmission, must send their QoS requirements to WBM, which will accept/reject the requests according to the availability of the bandwidth in the wireless LAN.

Many schemes [1][60] try to alleviate serious flow contention by dynamically changing the contention window. However, these approaches require modification of the standard. In fact, given a certain per priority contention window setting in EDCF, changing user priority also change contention window, only in a coarse scale. Thus, it is equivalently effective to assign a new priority appropriately for an incoming flow at the application layer. In this approach [41][42], flows are classified into two classes, *high priority* and *low priority*, which have priorities from 4 to 7, and 1 to 3, respectively. Correspondingly, flows are called *high priority flows* and *low priority flows*, respectively. Two flows are said to be of *the same class* (higher or lower) if they are in the same range specified. Also, *Flow_length* of a priority *p*, *Flow_length*(*p*), is defined as the total bandwidth demands of all the flows of priority *p* in the network. Then, when a new flow is initiated, its priority is reassigned according to two rules: (i) A flow is assigned a priority of the same class with the smallest *Flow_length*; (ii) Among all the priorities satisfying the above condition, the one closest to the original priority is chosen. In low or medium traffic load, assigning a lower priority does not decrease the received throughput of a flow because of enough idle slots in the channel. In high traffic load, our algorithm can effectively improve the overall throughput of the network by minimizing the collision rate. Another desirable feature of this algorithm is that it can easily punish misbehaving flows by assigning a very low priority to them and protect other flows. Figure 6 compares the overall received throughput and average normalized throughput, respectively. It can be seen that when all flows are of the same priority, the overall throughput is decreasing after the number of flows exceeds 10. The higher the priorities, the earlier the decrease occurs. This result is consistent since that under low and medium traffic load, re-allocating priorities does not decrease the overall throughput because of the low collision rate, and when traffic load becomes high, low priority is preferred for smaller collision rate. However, high overall throughput is always obtained with even priority distribution and the improvement is up to 44%.

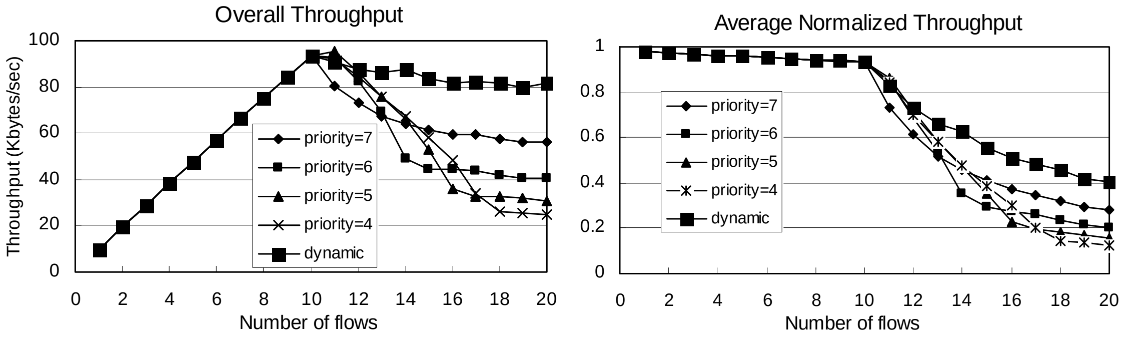

Figure 6. Comparison of dynamic priority reallocation scheme with different fixed user priorities.

The above even distribution affects the fairness of existing flows. Thus, in [43], the authors proposed to only re-allocate priorities of new flows to avoid severe channel contention with existing high priority flows. In this approach, the mean MAC level delay experienced by a sender for flows of different priorities, i.e., per priority MAC delay, is measured. Let D_i^m be the average MAC delay experienced by station *i* on transmitting the *m*th priority flow packets. Correspondingly, let *D_Max*[*m*] be the maximum MAC delay that can be tolerated for *m*th priority flows. Then, a new flow of priority *m* is droped if D_i^m >*D_Max*[*m*]. Otherwise, if D_i^m >Th_{pr}·*D_Max*[*m*], the flow is assigned a lower priority, where Th_{pr} is a threshold range from 0.6 to 0.8. Thus, under light traffic load, no priority reallocation is necessary since the mean MAC delay is quite low. However, under high network traffic, existing flows can be well protected with less contention from new flows. In addition, authors also proposed to enforce admission control to further guarantee the performance of admitted flows. Furthermore, a dynamic flow dropping scheme is devised to address the issue of *false admission*, i.e., multiple flows are admitted simultaneously even though the network cannot commit the service to all of them. With dynamic flow dropping, some flows will be appropriately dropped when the MAC delay is larger than a certain threshold after the admission.

3.2.2. CW and TXOP Adjustment Based Approaches

Similar to [6], Banchs and Perez-Costa [7] proposed a scheme to adjust CW and enforce admission control in IEEE 802.11e WLANs. In their scheme, a Wireless LAN Bandwidth Broker (WBB) is defined at the Access Point. Whenever a new station *n*+1 with throughput requirements R_{n+1} arrives, the WBB first re-computes the optimal contention windows of stations from 1 to *n*+1. Then, the expected throughput of the all stations is calculated. Then, station (*n*+1) is accepted only if for all stations, $r_i \geq R_i$, where r_i is the throughput of station *i*.

Pong and Moors [58] proposed a call admission control strategy for QoS of flows in IEEE 802.11e single-hop networks. In this approach, the authors adopt the analytical model proposed by Bianchi [9] to calculate the achievable bandwidth. The probabilities in the equation can be obtained by measuring the collision rate. Then, the admission control is enforced at the Access Point (AP). Every time a new flow request arrives, the collision rate is initialized to the average collision rate of a flow with similar achievable throughput as the desired bandwidth of the new flow. Then, the AP iteratively reduces CW or increase TXOP and re-estimate achievable throughput. If the requested bandwidth is smaller than the achievable throughput after adjustment, the flow is accepted with the new CW and/or TXOP distributed to all other stations. Otherwise, the flow is rejected. The desirable feature of this approach is that it adaptively tunes the parameters to achieve the maximum admission ratio.

3.2.3. Data and Admission Control Approaches

Xiao and Li [72][73][74] proposed both centralized and distributed algorithms for data control and admission control in IEEE 802.11e EDCF. In centralized algorithms, the Access Point (AP) collects all the information such as successful and failing data transmissions from each station and makes a global decision, while in distributed algorithms, each station estimates its own information locally by monitoring the channel. Basically, the ideas for the centralized and distributed algorithms are quite similar. In this section, we focus on the centralized version [74]. Readers are encouraged to study the original papers for more details on the distributed version. Two major components are proposed:

- *Data Control:* The parameters such as CW_{min}, CW_{max} and *AIFS* of best-effort data flows are adjusted dynamically to protect the performance of existing voice and video traffic. Basically, these parameters should be increased if there are two many data transmissions, and vice versa.
- *Admission Control:* The *TXOP* budget and transmission time of all Access Categories are recorded as the basis of admitting/rejecting a new flow and assigning maximum transmission time for existing voice/video flows at next beacon interval.

For global data control, the key idea is how to identify the "event" of more or less data traffic. The idea in [74] is to use two variables: successful transmission time (STT) per AC and failed transmission time (FTT). These two variables are measured and estimated periodically during every beacon interval *t*. Then, the AP checks if *FTT* increases significantly. If that is the case, it will try to increase CW_{min}[0], CW_{max}[0], and *AIFS*[0] if $FTT(t)/(FTT(t)+sum(STT[i](t), i \in [1,3]))$ is greater than a small pre-defined threshold. The reason for this is that higher *FTT* usually indicates heavier data traffic and it is necessary to limit the channel access frequency of data flows by increasing these parameters. On the other hand, if *FTT* decreases significantly and at the same time *STT* increases significantly for at least one AC, then CW_{min}[0], CW_{max}[0], and *AIFS*[0] are decreased to improve the performance of data traffic.

For admission control, the idea is to maintain a transmission budget for voice/video traffic. If the budget for an AC is depleted, new stations will be rejected, while existing stations cannot increase their transmission time per beacon interval. Thus, the admission control protects admitted voice/video streams. To enforce the admission control, five variables are maintained: *TxUsed*[i] (the amount of time occupied by transmission of AC *i* stations), *TxCounter*[i] (time of successful transmissions of AC *i* stations), *TxLimit*[i] (maximum transmission time of AC *i* stations), *TxRemainder*[i] (transmission time of AC *i* stations that can be carrier over to the next beacon interval), and *TxMemory*[i] (the amount of resource that AC *i* of a station utilizes at a beacon interval). These variables are updated at each beacon interval. The TXOPBudget of a an AC *i* station is calculated as

$$TXOPBudget[i] = \max(ATL[i] - TxTime[i] \times SurplusFactor[i], 0)$$

Where *ATL*[i] is the maximum amount of time that may be used for transmissions of AC *i* per beacon interval. If *TXOPBudget*[*i*] becomes zero, *TxMemory*[i] and *TxRemainder*[i] will be set to zero and new stations cannot start transmission with AC *i*. Otherwise, *TxMemory*[i] will be set to an initial value between 0 and *TXOPBudget*[*i*]/ *SurplusFactor*[*i*]. This time will be used for stations of AC *i* in the next beacon interval. In this case, new stations can possibly transmit flows of AC *i*. With both data control and admission control, two objectives can be achieved: (i) existing voice/video flows are not affected by the transmission of new flows of the same AC; (ii) existing voice/video flows are not affected by heavy data transmissions. Thus, sufficient protection for real-time traffic is provided in the proposed scheme.

In contrast to EDCA, no much research has been done on possible extensions of the HCCA scheme. One reason for this is that as a centrally controlled channel access scheme, HCCA can naturally enforce admission control. However, with HC being a central point, HCCA does not posses many advantages of the distributed EDCA mechanism and thus lacks its popularity in practice.

4. Resource management in multi-cell wireless LANs and Heterogeneous Wired/Wireless Networks

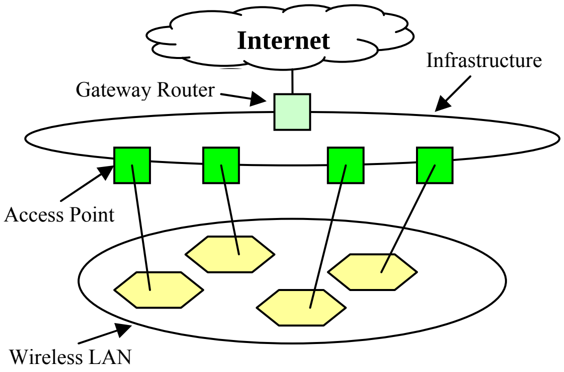

Figure 7. Architecture of Wireless Internet

IEEE 802.11 wireless LANs are being successfully used as the last-mile technology in the present-day pervasive computing environments. In many instances, wireless/mobile users need to access/exchange information stored in some servers located in wired networks. The architecture of wireless Internet is depicted in Figure 7 (revised from MIRAI architecture [30]). Multiple basic service sets (BSSs) are inter-connected by a distribution system (DS) and form an extended service set (ESS). An ESS is connected to the Internet via a gateway router. In each BSS, all mobile hosts (MHs) are within the broadcast region of their associated AP and can only access the infrastructure through the AP. All MHs can roam among different BSSs of the same or different ESSs.

In such a wired-cum-wireless network, mobile/wireless users usually access/exchange information stored somewhere on the Internet via their associated APs and the gateway routers. For multimedia applications such as audio/video streaming, end-to-end QoS guarantee is highly desirable in order to ensure user satisfaction. Basically, two requirements should be satisfied:

- Seamless roaming is supported when users move among different BSSs. With QoS handoff, ongoing multimedia traffics will not suffer from severe performance degradation due to insufficient network resource in the new cell.
- End-to-end QoS guarantee is provided through wired/wireless signaling and accurate wireless LAN admission control.

Basically, QoS handoff in multi-cell wireless LANs and end-to-end QoS signaling in wired/wireless networks are complimentary and can be combined to maximize wireless users' experience [44].

4.1. QoS handoff in multi-cell wireless LANs

IEEE 802.11 standard [81] specifies a procedure called *hand-off* where mobile hosts (MHs) may move between BSSs and transfer its associated AP from one cell (home cell) to another cell (target cell). In IEEE 802.11b, 11 different channels are available for use and channel 1, 6, and 11 are non-overlapping. Usually, handoff occurs when a MH moves across the boundary of two or more wireless APs and detects weak signal reception with its current associated AP or it experiences significant QoS deterioration. To facilitate the handoff, each AP broadcasts the beacon signal periodically and a MH scans the beacon signal and sees if there exists another AP with stronger beacon signal. Then, the MH sends authentication and reassociation request to the AP with strongest beacon signal. Upon receipt of the request, the new AP makes necessary security checking and sends the reassociation response if it accepts the handoff. The reassociation response includes information such as supported bit rates and station ID. Meanwhile, the new AP sends an Inter-access-point-protocol (IAPP) message to the old AP to inform the completion of the handoff. Figure 8 illustrates the handoff procedure in 802.11 WLANs. From a mobile user's perspective, it is desirable to have very low handoff latency. Basically, handoff latency is composed by *probe delay*, *authentication delay*, and *re-association delay*. Experiment results in [46] have shown that among these delays, probing delay consists of the biggest part (over 90%) of the overall handoff latency. Thus, existing research works mainly focus on techniques to reduce the probing delay.

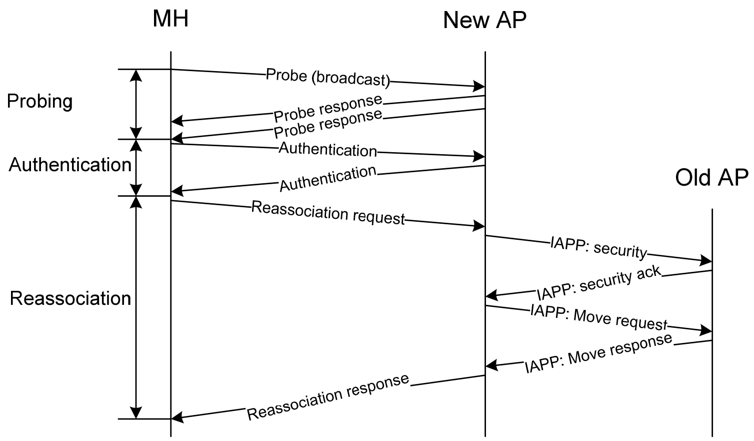

Figure 8. Illustration of the handoff procedure in IEEE 802.11 wireless LANs

4.1.1. Minimization of Handoff Delay

Shin and Forte [66] proposed to two techniques to reduce the probing delay:

- *Selective scanning algorithm* that improves the scanning procedure. In this algorithm, a full channel scanning is made initially. When a station scans APs, a channel mask is built. Also, channel masks of channels 1, 6, and 11 are also set. After the stations associates with the best AP, the channel being used is removed from the channel mask since the likelihood of an adjacent AP on the same channel is very small. Then in the next handoff, if no APs are discovered with the current channel mask, the channel mask is inverted and a new scan is performed. If still no APs are discovered, a full scan on all channels is performed until some APs are found. Experiments show that selective scanning reduces the handoff latency by 30-60%, compared to the original handoff scheme in 802.11 DCF standard.
- *Caching technique* that minimizes the number of times the previous scanning procedure is needed. In this technique, an AP cache consisting of a table (with MAC address as the key) is created. With the selective scanning algorithm, multiple APs may be found and the best two APs with strongest signal strength are entered in the cache with the old AP as the key. Then, at the time of handoff, a station first checks the cache to see if there is a hit. If successful, the station sends a message to associate with the new AP. Otherwise, a selective scanning is performed. If the station fails to connect with the first entry, the second entry is tried. If both attempts fail, the selective scanning is performed to update the cache. Experiments show that caching can reduce the probing delay to very small values (5-20 ms).

4.1.2. Bandwidth Aware Handoff

However, the AP with strongest signal strength is not necessarily the best one for handoff. If everyone tries to move to the new cell with high bit rate, the AP is usually handling heavy network traffic. In this case, it is highly possible that a MH may move from a less crowded cell to a more crowded cell and its QoS requirements are no longer satisfied, leading to serious performance degradation.

To address this issue, Lo and Lin [46] proposed traffic load aware handoff scheme. In this scheme, if the weighted average traffic load in the old AP does exceed a threshold, the signal strength is considered. If the signal strength in the new AP is significantly higher than the old AP, the handoff is initiated. Otherwise, the handoff request to the candidate new AP is denied since in this case the handoff does not provide any benefit. On the other hand, if the traffic load in the old AP is higher than a threshold, both the load and signal strength need to be considered. The handoff will be initiated only when two conditions are satisfied: (i) the signal strength of the new AP is higher than a threshold; (ii) traffic load in the new AP is not significantly higher than the old one. Thus, by enforcing more strict criteria, the problem of overloading due to handoff can be alleviated. However, since handoff is still allowed when traffic load is heavy, performance degradation after handoff still occurs.

To prevent QoS degradation when a node moves to a new cell, Li and Zhu [44] proposed to enforce admission control during the handoff. First, a fixed amount of channel available bandwidth (15-20%) is reserved for handoff to minimize handoff blocking probability. The actual proportion depends on the frequency of node mobility and traffic load. Second, admission control is enforced every time a node moves into a new cell with ongoing traffic. The available bandwidth estimation is similar to [36][61]. If there is enough bandwidth to support the flow, the handoff is initiated. Otherwise, another candidate AP is selected for handoff attempt.

4.2. QoS guarantee in heterogeneous wired/wireless networks

Integrated Service (IntServ) [12] and Differentiated Service (DiffServ) [11] have been proposed to enhance the QoS support in broadband Internet access. In IntServ, resource reservation and admission control is enforced to provide guaranteed service to users demanding minimum bandwidth or maximum delay. RSVP/SBM [13][77] has been widely accepted as a reservation scheme for flow reservation in IEEE 802 style LANs to support Integrated Service (IntServ) [12]. In DiffServ, service levels are prioritized such that higher priority users gain more network bandwidth for data transmission. Therefore, a natural idea would be the integration of RSVP or DiffServ in wired infrastructure and a QoS schemes in wireless LAN to provide end-to-end QoS support in a cost-effective manner.

Park and Kim [54] proposed a QoS architecture between DiffServ and IEEE 802.11e [82]. The idea is that priorities in IEEE 802.1D/Q can be mapped to IEEE 802.11e to provide end-to-end service differentiation. In order to support QoS and Mobility for wireless Internet, García-Macías and Rousseau [23] presented a hierarchical QoS architecture to extend DiffServ to WLANs with flexible mobility management.

Moon and Aghvami [50] proposed a QoS mechanism in all-IP wireless access networks. In their scheme, re-routing of RSVP branch path toward the crossover router at every handoff event is made for reduction of resource reservation delays and signaling overheads. Also, advance reservation is made via a new BS for on-going flows to maintain QoS guarantee. Shankar and Choi [62] proposed a MAC-level QoS signaling for IEEE 802.11e WLAN and address its interaction with RSVP and SBM. However, no effort is made to handle the specific issue of admission control in wireless LANs. Also, mobility issue is not addressed.

Based on a proposed MAC layer flow reservation and admission control protocol in IEEE 802.11 WLAN, called WRESV, Li and Zhu [44] suggested to integrate RSVP and WRESV for the support of IntServ in heterogeneous wired-cum-wireless networks. In their approach, to establish end-to-end flow reservation, signaling messages need to be sent between wireless users and wired servers through wireless AP, gateway router, and intermediate IP routers. As QoS signaling schemes for the Internet and wireless LAN, RSVP and WRESV are integrated to realize end-to-end flow reservation. Since WRESV follows the same procedure of request/response as RSVP does, mapping of signaling messages can be made between the reservation messages of Path/Resv of RSVP, and REQUEST/RESPONSE of WRESV, respectively. Features of RSVP and the characteristics of wireless medium are carefully considered, e.g., multicast receivers in a wireless LAN can be handled by broadcasting at the AP. Message mappings at the AP are implemented by cross-layer interaction and user priorities are mapped to 802.11 MAC priorities with 802.1p. Since WRESV can work with most of the existing MAC schedulers such as DCF, EDCF, and DFS, this integration scheme is more general and leaves space for further enhancement. Advantages of this approach include: (i) Reservation based schemes can provide ensured QoS guarantee stronger than service differentiation, especially when network traffic is high. (ii) Per-flow signaling for end-to-end reservation becomes quite easy through integration of a RSVP-like flow reservation protocol in wireless LAN and RSVP. (iii) The message mapping and resource management are only needed at the access point. (iv) The overhead of the integration can be minimized by cross-layer interaction between MAC and upper layers. Furthermore, this integration scheme also considers support of both node mobility and QoS in the situation of handoff.

4.3. QoS and Mobility Management in Hybrid Wireless Networks

In addition to roaming and horizontal handoff among 802.11 WLANs, supporting QoS anytime, anywhere, and by any media requires seamless vertical handoffs between different wireless networks such as WLAN, MANET, Bluetooth, UMTS, and WCDMA. Many new architectures/schemes have been proposed recently for seamless integration of WLAN and various wireless network interfaces, which are discussed as follows.

- **Integration of WLAN and MANET:** Lamont and Wang [38] investigated the issue of maintaining session connectivity while mobiles continuously roam across multiple WLANs and MANETs. In the proposed network architecture, routing within MANETs is handled by the Optimized Link State Routing (OLSR) protocol and handoff between WLANs and MANETs are supported through automatic mode-detection and node-switching capabilities of the mobiles. To achieve efficient mobility management, functionalities of OLSR are extended to support Mobile IPv6.

- **Integration of WLAN and Bluetooth:** Conti and Dardari [19] proposed an integrated analytical model for evaluation of the interference between IEEE 802.11 and Bluetooth. The model takes both physical layer and MAC layer into account and can be easily implemented. The performance is evaluated by packet error probability in term of the relative distances between the two systems for different conditions.

- **Integration of WLAN and 3G Wireless Networks:** An architecture for Integrating UMTS and IEEE 802.11 WLANs was proposed by Jaseemuddin [31]. Since 802.11 is used primarily for high-speed best-effort service, a mobile node can maintain two connections in parallel, i.e., data connection through WLAN, and voice connection through UMTS. Park and Yoon [53] investigate vertical handoff between IEEE 802.11 WLANs and CDMA cellular networks. In their handoff strategy, traffic characteristics are considered in order to guarantee low handoff latency. Specifically, real-time traffic takes into account the handoff delay and best-effort traffic takes into account of throughput only. Finally, Buddhikot and Chandranmenon [14] suggest to combine the features of wide-coverage but low-rate 3G networks, and high-rate but small-coverage WLAN, to improve the QoS and flexibility of wireless data services. A loose integration approach is realized with an IOTA gateway and a new client software in order to support seamless mobility, QoS guarantees and multi-provider roaming agreements.

With the decreasing size of cells in next generation multimedia enabled wireless networks, the number of handoffs during a call's life time increases. Thus, for integration of WLAN and 3/4-G wireless networks, an essential element of seamless end-to-end QoS guarantee is the ensuring of low call dropping probability in the 3/4-G networks. Lou and Li [47] proposed an adaptive bandwidth allocation scheme, termed measurement-based pre-assignment technique, to prevent handoff failure in wireless cellular networks. With periodical measurement of traffic status within a local cell, the number of channels reserved for a handoff can be adjusted, thus eliminating the signaling overhead of status information exchange between involved cells.

5. Resource management in MANETs

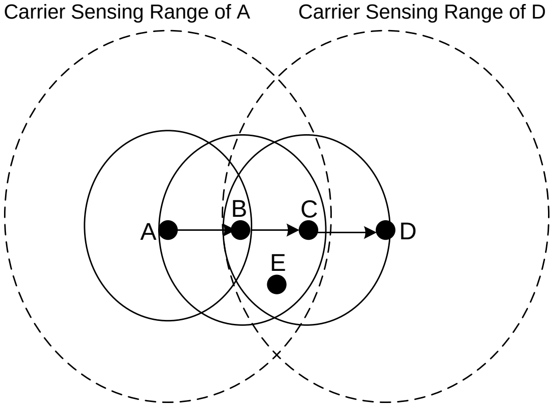

Figure 9. Illustration of channel interference

In IEEE 802.11 MAC, nodes cannot transmission and receive data simultaneously. Therefore, both the reception and transmission will take the bandwidth resource of nodes. For example, in Figure 9, let node A send packets to node D, through nodes B and C. At the time when A sends to B, B may also tries to forward packet to C, and C forwards packets to D. In this case, A will sense busy channel if B's broadcast arrives A first, i.e., B's transmission consumes available bandwidth at A. Meanwhile, since A is within the carrier sense range (CS range, RCS) of C, C actually also contends with A when the flow is introduced. Thus, when a new flow is admitted, A must have available bandwidth that equals to roughly 3 times the requested flow rate due to the consumption of its available bandwidth by node B and C. Similarly, B, C, and D need 3, 3, 2 times the requested flow rate available, respectively. Furthermore, suppose node E is located within the carrier sense range of A, B, C, and D. Then, when data transmission over any of the three links occurs, E senses busy channel and has to wait. If there are ongoing traffics going through E, those traffic may be significantly affected by the new flow due to the limited available bandwidth at E.

In single rate networks, all outgoing links of a node have basically the same physical link rate. However, when multi-rate capability is enabled, a node may have links with different possible rates. If we assume Two-ray propagation model [81], then link rate is solely dependent on node distance. For IEEE 802.11b, the possible link rates at a node A can be 2, 5.5, and 11 Mbps, depending on the perceived signal strength. Thus, different links starting from the same node may have quite different available bandwidth. To make appropriate admission decision, each node has to maintain per *link* available bandwidth information for all outgoing links. Usually, admission control schemes tend to choose high rate links for packet transmission. However, selecting high rate link usually yields longer path due to the relatively shorter transmission distance at high link rate. With more number of hops, transmission at one node may be affected by a higher number of transmissions within its CS range and the previously mentioned interference becomes more significant. Therefore, as far as the admission control decision is concerned, there is a tradeoff between choosing high rate links and short distance paths.

Node mobility also has high impact on network performance. In a MANET, the network topology is dynamically changing, leading to frequent link breaking. If a flow is being transmitted on a broken link, then route re-discovery has to be performed to smoothly forward packets along new paths. However, it is possible that the new path may not be able to support the QoS requirement of the flow. If multi-rate is enabled, node mobility affects signal strength and thus the channel transmission rate. Thus, link condition varies and a previously supported flow may not receive good QoS as it is promised. In both cases, a new path may have to be chosen for that flow.

In summary, QoS guarantee in IEEE 802.11 based MANET is a challenging issue due to its natural characteristic of shared medium access and node mobility. Successful resource management schemes demand three components:

- *Accurate available bandwidth estimation*: Given the network topology, channel status, traffic characteristics, and mobility, how to accurately estimate the available bandwidth for a specific flow along a certain path;
- *Route selection:* With the available bandwidth and other information, how to choose the best path to support a multimedia flow;
- *Mobility handling:* When nodes are moving frequently with medium or high speed, how to provide reliable and sufficient QoS support.

5.1. QoS Scheduling, Routing and Frameworks

Early research on QoS in MANETs basically focused on either extending of QoS strategies in single-hop wireless LANs such as MAC service differentiation, or investigating network layer strategies such as routing for resource reservation and path selection. These schemes usually assume low node mobility and no consideration is given for link interference. Thus, the performance improvements for multimedia applications are quite limited.

5.1.1 Priority Scheduling in Multi-hop Ad Hoc Networks

Kanodia and Li [35] proposed a distributed priority scheduling for end-to-end quality of service guarantee in multi-hop ad hoc networks. Due to the distributed nature of wireless networks, it is difficult to coordinate channel access in an ideal fashion. In their work, a deadline based priority backoff policy is devised and the priority information is piggybacked with RTS, CTS, and ACK to minimize the extra overhead. With the priority information obtained from neighbors, each node can rank the priority of all the nodes and decide the backoff counter as follows:

$$f_l(S_j) = \begin{cases} \text{Uniform}[0, 2^l CW_{\min} - 1], & r_j = 1, l < m \\ \alpha CW_{\min} + \text{Uniform}[0, \gamma CW_{\min} - 1], & r_j > 1, l = 0 \\ \text{Uniform}[0, 2^l \gamma CW_{\min} - 1], & r_j > 1, l \geq 1 \end{cases}$$

Where r_j is the rank of node j's packet in its own scheduling table S_j. α and γ are some constant and set as 1 and 2, respectively. With priority scheduling, even though a node may not be able to overhear all packets, significant contention reduce can still be achieved to improve the performance of high priority flows.

Then, to satisfy the end-to-end delay requirement of flows, downstream stations adjust the priority levels of packets based on their performance in the upstream by a technique known as multi-hop coordination. However, the reduction of average end-to-end delay is not significant. There are two reasons for this: (i) There is no approach to try to select the lowest delay path. If a high delay path is chosen, then it is very difficult to reduce the delay through priority change. (ii) Since every node enforces priority change to improve their end to end delay, the overall effect is not very desirable among all flows.

Thus, routing should play an important role in supporting QoS in multi-hop ad hoc networks. The basic idea of QoS routing is to find a path that has sufficient bandwidth for a new flow such that its throughput and delay requirements can be satisfied. From this aspects, it is important to integrate certain QoS measures with existing routing protocols such as Destination Sequenced Distance Vector (DSDV [55]), Dynamic Source Routing (DSR [32]), Ad Hoc On Demand Distance Vector (AODV [57]), and Optimized Link State Routing Protocol (OLSR [18]).

5.1.2. QoS Routing in MANETs

Perkins and Royer [56] proposed a QoS version of the earlier proposed on-demand routing protocol AODV [57]. The idea of AODV-QoS is to facilitate routing based on QoS metrics, not only the number of hops. AODV-QoS adds the following fields to each routing table entries corresponding to each destination: (i) *Maximum Delay*; (ii) *Minimum Available Bandwidth*; (iii) *List of Sources Requesting Delay Guarantees*; (iv) *List of Sources Requesting Bandwidth Guarantees*. The Maximum Delay Extension can be appended to a Route Request (RREQ) by a node requesting a QoS route in order to place a maximum bound on the acceptable time delay experienced on any acceptable path from the source to the destination. Before forwarding the RREQ, an intermediate node MUST compare NODE_TRSVERSAL_TIME to the remaining Delay indicated in the Maximum Delay Extension. If the remaining Delay is smaller than NODE_TRSVERSAL_TIME, i.e., the deadline has passed, the RREQ is simply discarded. Otherwise, the node subtracts NODE_TRSVERSAL_TIME from the Delay value in the extension and continues processing the RREQ as specified in AODV. Then, when a node forwards a Route Reply (RREP), it adds its own NODE_TRSVERSAL_TIME to the Delay field, which is initialized to zero at the destination. Also, this information is recorded in the local route table entry for that destination. With this information, an intermediate node may be able to reply a RREQ by comparing the Maximum Delay field in the route table entry and the requested Maximum Delay in the RREQ. Similarly, the Minimum Bandwidth Extension can also be appended to a RREQ in order to specify the minimum amount of bandwidth that must be made available along an acceptable path from the source to the destination. Before forwarding the RREQ, an intermediate node must compare its available link capacity to the Bandwidth field in the extension. If the available bandwidth is less the requested one, the RREQ is simply discarded. Each intermediate node forwarding RREP compares the Bandwidth field in the RREP and its own link capacity and maintains the minimum of the two in the Bandwidth field, which is initialized to infinity at the destination. Also, this information is recorded to the local route table entry for the destination. With this information, an intermediate node may be able to reply a RREQ by comparing the Minimum Bandwidth field in the route table entry and the requested Minimum Bandwidth in the RREQ. AODV-QoS provides a good framework for QoS support in mobile ad hoc networks for multimedia applications that demands certain bandwidth and delay constraints. However, there are three issues to be addressed in order to enforce QoS with AODV-QoS: (i) An accurate and efficient bandwidth estimation technique is required in order for an intermediate node to be able to update the Minimum Bandwidth field in RREP and the local route table; (ii) How to provision the maximum delay requirement among multiple hops to make sure that the probability that a RREQ is incorrectly dropped, i.e., the total delay is smaller than the Maximum Delay but RREQ is dropped at an intermediate node, is minimized.

Chen and Nahrstedt [17] proposed Ticket-based Probing (TBP) one of the first QoS routing protocols in mobile ad hoc networks. Authors assumed that the network topology does not change frequently such that soft QoS can be supported. In this work, a simple imprecision model is first proposed to estimate $\Delta D_i(t)$, the maximum change of $D_i(t)$, the delay from node i to node t before the next update, and $\Delta B_i(t)$, the maximum change of $B_i(t)$, the bandwidth from node i to node t before the next update. Then a smoothing technique is used to calculate the new $\Delta D_i(t)$ and $\Delta B_i(t)$ at each update as

$$\Delta D_i^{new}(t) = \alpha \times \Delta D_i^{old}(t) + (1-\alpha) \times \beta \times \left| D_i^{new}(t) - D_i^{old}(t) \right|$$

and

$$\Delta B_i^{new}(t) = \alpha \times \Delta B_i^{old}(t) + (1-\alpha) \times \beta \times \left| B_i^{new}(t) - B_i^{old}(t) \right|$$

Where α and β are some constants to make sure that it is highly probable that the actual delay and bandwidth fall into the ranges $[D_i(t)\text{-}\Delta D_i(t),\ D_i(t)+\Delta D_i(t)]$ and $[B_i(t)\text{-}\Delta B_i(t),\ B_i(t)+\Delta B_i(t)]$, respectively. With these delay and bandwidth measurement, the cost of paths can be obtained.

The ticket-based probing routing protocol works as follows. For a source s and destination t, two types of tickers are issued from s every time a packet sending requests arrives:

- *Yellow tickets* that maximize the probability of finding a feasible path. Thus, yellow tickets prefer paths with smaller delays to satisfy the delay requirement.

- *Green tickets* that maximize the probability of finding a low-cost path. Thus, green tickets prefer paths with smaller costs but may have larger delay.

The strategy is to use the green tickets to find low-cost feasible path with relatively low success probability and to use yellow tickets as a backup to guarantee a high success probability of finding a feasible path. Then the total number of tickets issued at s, $N_0=Y_0+G_0$, where Y_0 and G_0 are the number of yellow tickets and green tickets, respectively. Basically, Y_0 depends on the delay requirement D. If D is larger than $D_s(t)+\Delta D_s(t)$, then only one yellow ticket will be sufficient to find a path to satisfy the D requirement. Otherwise, more yellow tickets are preferred to maximize the probability. However, if D is even smaller than the best expected end-to-end delay $D_s(t)-\Delta D_s(t)$, no ticket is issued and the connection request is simply rejected. G_0 also depends on the delay requirement D. Larger D usually gives a smaller G_0. However, to obtain low cost paths, $G_0=1$ if $D \geq \theta\times(D_s(t)+\Delta D_s(t))$ and then increase with smaller D. Similarly $G_0=1$ if $D \geq \theta\times(D_s(t)+\Delta D_s(t))$. A probe p from the source accumulates its delay, $delay(p)$ as it propagates along a path. When an intermediate node i receives p with a number of tickets from a node k, it determines the neighbors to forward the probes as follows. The set of candidate neighbors, $R_i^p(t)$, is defined as $\{j \mid delay(p)+delay(i,j)+D_j(t)-\Delta D_j(t)\leq D, j \in V_i-\{k\}\}$, where $delay(i,j)$ is the delay from node i to its neighbor j. and V_i is the set of neighbors of node i. Since stationary nodes provide more stable paths than mobile nodes, they are given priority for distributing probes. Among the stationary nodes, more yellow probes are sent along the paths with a smaller delay and more green tickets are sent along the paths with smaller cost.

TBP achieves desirable performance due to two advantages: (i) the proposed delay constrained routing and bandwidth constrained routing only choose qualified paths that satisfy the delay/bandwidth requirements of the applications, thus significantly improving the throughput of accepted flows; (ii) the adaptive ticket determination avoids query flooding and intelligently forwards probes along paths with high probability of satisfying the delay/bandwidth requirements.

5.1.3. QoS Frameworks in MANETs

Lee and Ahn [39] proposed INSIGNIA, a QoS framework for supporting adaptive services in mobile ad hoc networks. In INSIGNIA, routing, QoS signaling, and resource reservation are separated. The three interesting features in INSIGNIA are:

- *In-band Signaling*: To facilitate fast reservation, in-band signaling where the control messages are carried along the data packet is adopted. In contrast, with out-band signaling, control messages are sent as separate control packets and may be sent along different data paths. The advantage of in-band routing is that it is capable of operating close to packet transmission speed and can significantly reduce the resource reservation overhead.

- *Adaptive service support*: To support adaptive service, several fields are defined in the IP header: (i) the *reservation* (*RES*) *mode* bit indicate the connection request should go through admission control procedure. If it is not set, it is considered best-effort (BE) traffic. (ii) the *payload type* identifies whether the packet is a base QoS (BQ) or enhanced QoS (EQ) packet. While BQ requires only minimum bandwidth requirements to be met along the path between a source-destination pair, EQ requires maximum bandwidth requirements to be met. Thus, the payload type affects the admission control decision. (iii) the *bandwidth indicator* bit can be set to identify whether a max-reserved or a min-reserved should be enforced. (iv) The *bandwidth request* provides the information on how much maximum/minimum bandwidth is requested from the source. With these information included in the IP header of data packets, it is very convenient for intermediate nodes to make adaptive decision on providing sufficient QoS support for the flows.

The resource reservation protocol works as follows. Source nodes first initiate reservations by setting appropriate field of the IP option in data messages before forwarding "reservation request" towards destination nodes. Usually, such reservation request packet set service mode, payload, and bandwidth indicator to RES, BQ/EQ, and MAX/MIN, respectively. Upon the arrival of a reservation request at an intermediate node, the node will enforce admission control modules, allocate resources, and establish local state. This procedure continues until the destination receives the reservation request packet and sends back a QoS reporting message. At this time, the reservation is done. Sometimes, there are some situations that a flow is min-reserved but the payload type is EQ. To fix this conflict, all reserved packets with EQ type received at a destination will have their service level switched from RES to BE by the bottleneck node. As a result of this, all resource reserved earlier for the flow will be released. INSIGNIA also provide mechanisms to handle reservation restoration and soft state management when mobility is present.

There are three issues to be resolved in INSIGNIA: (i) managing the soft states at each intermediate node incurs significant overhead. When mobility is frequent, removing the state information requires additional signaling. (ii) how to accurately estimate bandwidth availability to avoid over-commitment or under-commitment is not mentioned. Without effective bandwidth estimation, it is difficult to guarantee the performance of accepted real-time flows. (iii) When both real-time and best-effort traffic coexist, how to provision the network resource such that the performance of admitted real-time flows will not be affected significantly by heavy best-effort traffic. This usually either requires priority based access control protocols at the MAC layer, or flow control protocols at the transport layer.

Anh and Campbell [3] proposed SWAN, a service differentiation framework in stateless wireless ad hoc networks. Unlike INSIGNIA, SWAN does not maintain per-flow state information at intermediate node and does not assume any prioritized MAC access protocol to do service differentiation for real-time and best-effort flows. In SWAN, three major control algorithms are implemented:

- *Rate control for best effort traffic*: Additive Increase and Multiplicative Decrease (AIMD) rate control algorithm based on measured MAC layer delay is used to regulate the sending rate of best-effort flows. Every T seconds, each mobile device increases its transmission rate (by *c* Kbps) gradually until the packet delays exceed a certain threshold. When the excessive delay is detected, the rate controller reduces the sending rate by *r*%. With this rate control, the effect of best-effort flows to real-time flows is minimized.

- *Source based admission control for real-time traffic:* An admission controller sends a probing request packet toward the destination node to estimate the end-to-end bandwidth availability. When the source receives the probing response message, it compares the measured available bandwidth and the bandwidth requirement of the new real-time session and makes the admission/rejection decision.

- *Dynamic regulation of real-time traffic*: When mobility is present, resource availability may change due to dynamic packet rerouting. In this case, each node periodically measures the local available bandwidth. If violation is detected, i.e., the bandwidth availability is smaller than the flow rate of the real-time traffic, the source will try to reestablish the real-time session based on its original bandwidth needs. If the session cannot be satisfied, it will be dropped.

SWAN is very practical and scalable due to its elimination of soft state of flows at intermediate nodes. Also, the regulation of best-effort traffic makes sure that good resource provisioning can be achieved for the support of real-time traffic. However, the bandwidth estimation of SWAN is not very reliable: (i) using probing packet does not provide very stable performance prediction of the real-time traffic; (ii) the multi-hop interference is not well captured, making over-commitment occur frequently.

5.2. Bandwidth Estimation Based QoS Routing

Despite of the effort of QoS frameworks and routing protocols, how to estimate the available bandwidth is the key for the success of resource management schemes. Without good knowledge about the resource availability, it is difficult to make decision on appropriate resource allocation and admission control. However, due to various factors such as network topology, multi-hop interference, rate adaptation, and traffic characteristics, it is impossible to have an accurate performance model for mobile ad hoc networks. The majority of the research in the literature focus on measurement based approach that provide efficient and reasonable prediction on how much bandwidth is available for a new flow on a specific path.

5.2.1. Interference Aware Bandwidth Estimation

Xue and Ganz [76] proposed Ad hoc QoS on-demand routing (AQOR), a QoS routing protocol with admission control enforced to support quality of service in multi-hop ad hoc networks. In their work, admission control is made based on the knowledge of the traffic at both a node itself and its neighboring nodes to account for the interference. In this approach, two different types of bandwidth are introduced:

- ***Available Bandwidth at a node i*** is the available bandwidth experienced locally at node *i* and calculated as $B_{avail,i} = B$ - sum($B_{self,j}$) for all neighbor *j* of *i* where *B* is the maximum transmission bandwidth and $B_{self,j}$ is the bandwidth consumed by traffic transmitted or received at node *j*.

- ***Consumed Bandwidth of a flow f at a node i***, $B_{consumed,i}(f)$, is the total channel bandwidth consumed by flow *f* due to traffic aggregation, i.e, node *i* also consumes bandwidth of neighbors transmitting/forwarding packets of flow *f*. $B_{consumed,i}$ is calculated as $B_{uplink,i}(f) + B_{consumed,i}(f)$, which are R_f if *i* is the source or destination and $2R_f$ otherwise.

By comparing the available bandwidth and the consumed bandwidth, an intermediate node can decide whether or not to accept or reject a flow.

Although AQOR is more accurate than earlier approaches [3][17] on the available bandwidth estimation by multi-hop flows, it only partially considers the multi-hop interference. To address this issue, Chen and Heinzelman [16] improved the AODV-QoS framework with a more accurate estimation of available bandwidth. In this approach, minimum bandwidth is first calculated locally with the similar way in [PAC], i.e., the available bandwidth is the product of channel idle ratio and the channel capacity. Then, messages are broadcasted to notify neighbors to obtain a *minimum* available bandwidth, $B_{avail,min}$, among all neighbors within the communication range. Finally, they adopted the result proposed by Li and Blake [40] to account for intra-flow interference as follows:

```
If ($HopCount < 4$)
    $B_{avail,min} = B_{avail,min}/HopCount$
Else
    $B_{avail,min} = B_{avail,min}/4$
```

This calculation provides an approximation of the upper bound of the available bandwidth for a flow along a certain path. However, it is not accurate enough, especially when multi-rate is enabled. With multi-rate MAC, the distance of a link may be much smaller due to transmission with high data rate, resulting in more intra-flow interference.

5.2.2. Intra-flow Interference Based Bandwidth Estimation

Since transmission at a node is potentially affected by transmissions of all nodes within its carrier sensing range, it is desirable that the list of those nodes on the same path that affect its transmission can be obtained at the time the available bandwidth is estimated. Li and Prabhakaran [45] proposed two metrics:

- ***Intra-flow Interference Set of a flow f at a node i on path p*** is a list of nodes (including *i* itself and excluding the destination) on *p* that reside within the carrier sensing range of *i*.
- ***Route Available Bandwidth of a flow f over path p*** is an index that indicates how much throughput a flow *f* can potentially receive if it is transmitted along a specific path *p* and is calculated based on important factors such as *intra-flow interference*, *effective link capacity*, and *channel busy time* in the MAC layer.

5.2.2.1. Intra-flow Interference Set

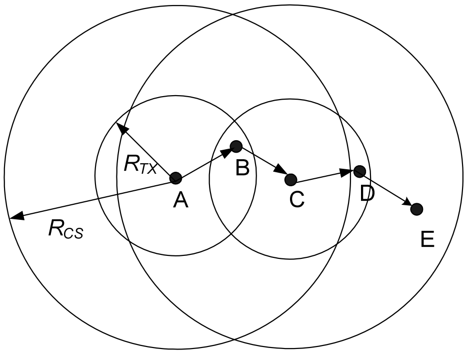

Figure 10. Illustration of intra-flow interference set

Intra-flow interference set of a flow *f* at a node *N*, $IS_{N,f}$, can be obtained as follows. Let $P=\{K_1, K_2, \ldots, K_{L-1}\}$ with L be the number of hops of *f*. Here K_1, K_2, through K_{L-1} are listed in the order of hops from source to destination. Note that the destination node is excluded since it does not send data packet. Let P' be the list of nodes within the carrier sensing range of node N and $P' =\{O_1, O_2, \ldots, O_{L'}\}$ with L' being the total number of those nodes. Then, the list of nodes on the path that affect node N is $P \cap P'$. Thus, the size of $IS_{i,f}$, $|IS_{N,f}| = |\ P \cap P'\ |$. Figure 10 illustrates intra-flow interference sets of nodes at different locations. CS range (R_{CS}) and transmission range (R_{TX}) of node A and C are represented by small and large circles, respectively. We can see that for the flow from node A to E along B, C, and D, the corresponding sizes of intra-flow interference set at node A and C are 3 and 4, respectively. While only node B and C are within R_{CS} of A, all nodes on the flow path are within R_{CS} of node C.

Obviously, the key component in intra-flow interference set is to identify all the nodes within the carrier sensing range. Although IDs of these neighbors are very easy to obtain, they consist of only a small portion of the interference set. In most network configurations, R_{CS} is more than 2 times R_{TX}, indicating that most nodes in IS are outside of the transmission range. Therefore, more nodes in the interference set must be discovered to achieve high accuracy on the estimation of the interference set. Three approaches have been proposed in the literature:

- *Two-hop approach*: Yang and Kravets [75] proposed to collect all neighboring nodes within two hops from a node as the intra-flow interference set. This approach is easy to implement, but it is not very accurate. It is possible that many nodes within 3 or 4 hops away from a node may still reside within its carrier sensing, especially when node density is low and nodes are not uniformly distributed in the network. It is also possible to adopt three-hop approach to increase the accuracy. However, the overhead will increase significantly and it is possible that some nodes outside of the carrier sensing range may be incorrectly included in the intra-flow interference set.
- *High power approach*: Yang and Kravets [75] proposed to estimate the interference set at a node by temporarily increasing the power so that it can communicate with all nodes within the carrier sensing range and then obtain the corresponding IDs. This is a quite innovative idea. However, it is impractical for a node to increase the power to such a high level in order to facilitate the communication. If this power is available, then normal data transmission would have taken advantage of this high power to increase the channel bandwidth significantly.
- *Carrier sensing approach*: Sanzgiri and Chakeres [63] proposed to calculate intra-flow interference set by monitoring the channel and estimate the sizes of packets broadcasted by nodes on a path. According to their approach, when a transmission occurs at a node within carrier sensing range of the receiver node, the received signal strength is greater than the carrier-sensing threshold (CS_{thresh}) and the receiving node is able to detect the packet. Furthermore, even though the receiving node may not be able to decode the packet, it can still determine the length of the packet by the sensing transmission duration. This is a more elegant way to estimate interference set without incurring too much overhead. To distinguish the packets from different nodes so that the number of nodes within carrier sensing range can be obtained, each node can simply broadcast a very short beacon with a *unique* length at basic transmission rate (2Mbps for IEEE 802.11b) periodically. If a node hears a packet from a neighbor or senses an error packet due to carrier sensing, it can obtain the corresponding ID of the source node and add it in its interference set.

5.2.2.2. Route Available Bandwidth

Table 2. B_e for different physical link rates

Link Rates (Mbps)	MTM Weight	B_e Weight	B_e (Mbps)
2	3	1	1.2
5.5	1.44	2.083	2.5
11	1	3	3.6

Let $b_{ava,e}$ be the experienced channel available bandwidth at station i along outgoing link e. Then, $b_{ava,e}$ can be approximated as

$$b_{ava,e} = B_e \cdot (1 - BT_e) .$$

Where B_e and BT_e are the maximal achievable bandwidth and the estimated channel busy time at a link e, respectively. It should be noted that it is not easy to calculate the effective link capacity. First, for different packet sizes, B_e is different. Second, the accurate backoff time is very difficult to estimate. In [22], bandwidth is estimated using probing packets. We conduct simulations and fix B_e to be 1.2Mbps for physical link rate of 2Mbps. Then, we use the weight factors proposed in MTM [5] and then decide the corresponding B_e for different physical link rates (Table 2). It should be noted that MTM weight is proportional to channel time consumption and thus is inversely proportional to B_e weight.

Let f be a new flow that will be transmitted through link e. Then, the available bandwidth at node i along link e will be consumed by packet transmissions of flow f at all nodes belonging to intra-flow interference set of node i for flow f. Since all links transmit with the same flow rate (we assume constant bit rate for flow f and $b_{ava,e}$ is *fairly* allocated to all contending links on path p), the *local* available bandwidth of flow f at link e, $b_{ava,e,f}$, can be expressed as

$$b_{ava,e,f} = \frac{b_{ava,e}}{|IS|_{e,f}} = \frac{B_e \cdot (1 - BT_e)}{|IS|_{e,f}} .$$

Where $|IS|_{e,f}$ is the size of intra-flow interference set of flow f at link e. Then, let p be a path of flow f, we have that route available bandwidth

$$RAB(f, p) = \min_{e \in p}(b_{ava,e,f}) .$$

With the estimated RAB of each path, admission control can be easily enforced at the source node by comparing the flow rate request and RAB. If RAB is sufficient, the flow is accepted. Otherwise, the flow is rejected.

5.3. Mobility Aware QoS Routing

Even though the intra-flow interference set based approach well captures the bandwidth consumption characteristics in multi-hop ad hoc networks, reliable QoS still may not be achieved when mobility is present. In most existing routing protocols such as DSDV, DSR, and AODV, whenever a link fails due to node moving out of the transmission range, an error message is sent back to the source node. Upon the receipt of the error message, the source node initiates another round of route discovery and sends packet along new path. During the transient time from link failure to the new path establishment, many packets are queued and even dropped at intermediate nodes, leading to significant throughput fluctuation and performance degradation.

Preemptive routing only chooses use the power threshold to alleviate route rediscovery and does not help much on the regular route discovery. However, when high mobility is present, most of links may not be very stable, leading to frequent packet rerouting and unreliable QoS due to the possibility of insufficient bandwidth on the new path. To address this issue, Goff and Abu-Ghazaleh [27] proposed Preemptive Routing where a warning message is sent to the source node whenever a link has high probability of getting broken due to mobility. In PR, each node detects the probability of link failure by measuring the power of packets received from neighbors. If the received packet power is lower than certain power threshold corresponding to a distance very close to the transmission range, the link is considered highly instable. When a source node receives a warning message, it initiates route discovery immediately but eliminates all paths where at least one of the link has power below the threshold. Since the new route is discovered while an existing path is still alive, the delay due to link failure can be minimized, if not completely eliminated.

To overcome the effect of mobility on QoS in MANETs, Li and Prabhakaran [45] also proposed the metric of Route Reliability (RR) that predicts how long an existing path may survive under a certain moving speed. First, let v_e be the *relative* moving speed between two nodes i, and j on a link e and L_e be the link distance, we can calculate reliability, LR_e, the estimated lifetime of link e as $(rang\text{-}L_e)/v_e$ where *rang* is the transmission range of the starting node of link e. Accordingly, let p be a path of flow f, then *route reliability* $RR=\min(LR_e)$ for all e on p. Then, similar to [27], it is assumed that signal strength is solely dependent on distance and the minimum power (P_{range}) receivable by the device, i.e., the power received at the maximum transmission range, which in [27] is indicated in $3.652 \cdot 10^{-10}$ W. Similarly, the power threshold δ is defined as the ratio between the minimum allowed power level for the required RR and is calculated as

$$\delta = \frac{P_{\min}}{P_{range}} = \left(\frac{range}{L_{e,\max}} \right)^4$$

Table 3. Power threshold δ for different maximum moving speed ($RR_{\min}$=1.5 second)

Moving speed (m/s)	5	10	15	20
v_{max} (m/s)	10	20	30	40
$v_{max} \cdot RR_{\min}$ (m)	15	30	45	60
δ	1.281	1.667	2.212	2.997

Then, let us assume 250 meters maximum transmission range, then the power threshold for a path becomes

$$\delta = \left(\frac{250}{250 - v_{\max} \cdot RR_{\min}} \right)^4 .$$

Table 3 gives the appropriate power threshold for different maximum moving speeds. It is possible that both nodes on a link move in opposite direction, so v_{max} is two times the maximum speed. RR_{min} is set to 1.5 seconds to eliminate highly unstable links. We could set this value higher but high RR_{min} might lead to the difficulty of finding an eligible path.

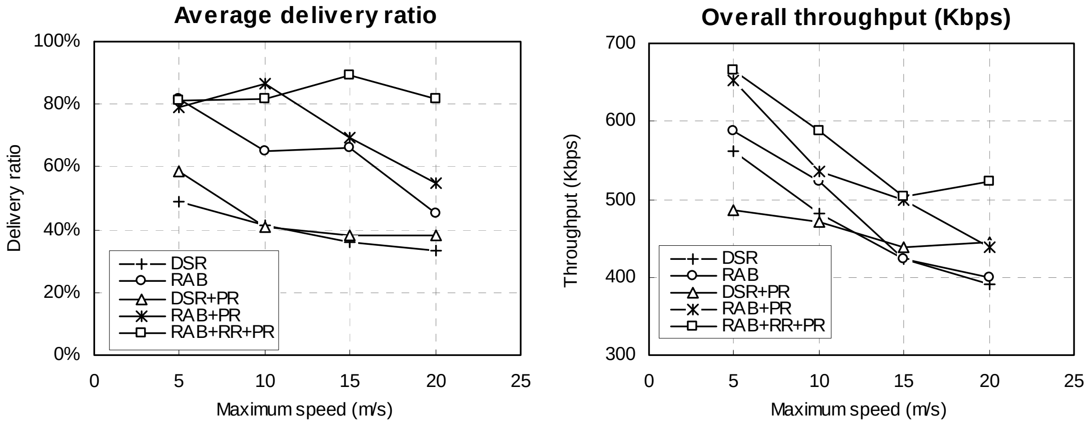

Figure 11. Performance comparison of various QoS schemes

The work in [45] integrates RAB and RR together with the Dynamic Source Routing (DSR) for reliable QoS support in MANETs. In this approach, only paths with power higher than the corresponding threshold and RAB higher than the requested bandwidth requirement are considered for data transmission. Then, the path with highest RAB is chosen for actual data transmission. When mobility is present, Preemptive Routing (PR) is adopted to avoid the high cost of route reestablishment. Figure 11 shows the average delivery ratio of admitted flows for various schemes. It can be seen that with low mobility, usually 60-90% throughput can be achieved with admission control. However, “DSR” and “DSR+PR” only receive 40-60% throughput on average. With the increase of the moving speed, the benefit of using RAB alone for admission control decreases due to frequent route re-discovery and severe channel contention. With preemptive routing, better performance can be achieved due to sending of early warning messages. With “RAB+RR+PR”, best performance is achieved in the case of high mobility (speed of 15 m/s and 20 m/s). Using route reliability usually results in less frequent route breaking and thus maintains higher per-flow throughput. Fig. 7 compares the overall throughput of all schemes. With higher mobility, overall throughput decreases due to more frequent link failure and route re-discovery. We can see that admission control does not reduce the channel utilization, though the improvement on overall throughput is marginal when only RAB is enforced for admission. However, with “RAB+RR+PR”, significant improvement (18% to 33%) on overall throughput is achieved, compared to the basic DSR protocol. This improvement should be attributed to less channel collision and link failure.

6. Resource management in WMNs

Wireless Mesh Networks (WMN) are perceived as the most promising commercial incarnation of the IEEE802.11 standard in the multi-hop domain. As presented in Figure 12: a wireless mesh networks is an access network, where the fixed infrastructure is reached through a gateway and the intermediate hops are sustained by fixed wireless routers. Of course, the same WMN can encompass several gateways. A WMN works in between layer 2 and layer 3, since it implements a local addressing strategy which is basically layer 2 routing, buy it is transparent to all other network trunks as it was any other AP based IEEE802.x LAN. In fact, the layer 2 addressing from external networks is not visible and, as in Ethernet LANs, all devices in the mesh appear directly connected to the gateway.

In brief, compared to ad-hoc networks, WMNs are IP-stack based, have a hierarchical structure and mobility at the infrastructure level is not possible. Thus, it is clear that for WMNs, mobility and power consumption issues are less relevant, but network scalability is still a primary concern. Network scalability, in particular, was a premiere research field in ad-hoc networks, since the seminal work from [29]. In the case of WMNs, some authors [4] argue that the operating hypotheses in WMNs are different, and this gives hope for tailored solutions where milder throughput decaying rates in the number of nodes are possible. Nevertheless, some other authors confirm the trend [34], making also the remark that, due to the concentration of traffic at the gateways, WMNs would suffer even worse scalability. It is worth remarking, that, for commercial purposes, larger number of hops means also larger coverage and higher income for the wireless operator. Furthermore, from the practical standpoint, different requirements in terms of services that must be supported by the WMNs are indeed true compared to ad-hoc networking. In fact, since WMNs are essentially IP networks, apart from traditional applications like FTP and Web Browsing are still present, current trends in Internet services show an increasing interest in multimedia application. Such applications are characterized by strict requirements in terms of delay, jitter and packet loss. Thus, in order to make WMNs a viable solution for wireless operator, the request to be addressed by the research community is the fundamental scalability issue. In fact, as the number of users increases, network performance degrades rapidly with end-to-end data rate lower than 1 Mbps for a five hops path over a IEEE 802.11-based WMN lowering the economic convenience of WMNs.

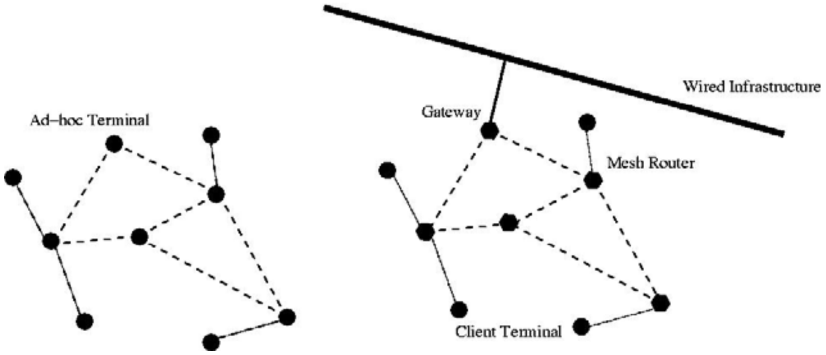

Figure 12. (a) Ad-hoc Network. (b) Mesh network.

In sight of the previous discussion, in WMNs the use of the IEEE 802.11 technology over several hops has to be modulated by suitable resource allocation policies in order to obtain reasonable performance. In literature, there exist some proposals. Two of the most promising techniques are the use of a multi-channel mesh architecture and cross-layer coupling of the routing layer to the links state, respectively.

6.1. Multi-channel WMNs

As concerns frequency allocation, the 12 available channels of IEEE802.11 can be used in order to increase the network throughput. In particular, in the first generation WMNs, all devices, routers, gateways and clients were adopting the same RF interface and leveraged the same channel. Later on, manufacturers and network designers, due to the availability of inexpensive IEEE802.11 cards, equipped WMN routers with one interface to connect to stations and one interface towards other routers. Thus, in the upcoming WMNs, multi-channel communications will be possible within the mesh routers backbone [84] As described in the emerging literature on multi-channel radio networks [52][67][2], the idea is that a common channel will enable multi-channel operation of devices with a single and multiple radio interfaces. Of course, non-multi-channel enabled stations or access points can be integrated in the WMN using a common channel, so that backward compatibility is guaranteed. The problem of frequency allocation is a traditional issue, but, the domain is to some extent new, as proved by the interest for the general problem of the frequency allocation over multi-channel CSMA/CA networks. In fact, the IEEE802.11 standard makes available 12 channels. Thus, the scarcity in the number of channels would represent a major problem for the allocations of channels. In particular, when the number of non-overlapping channels is smaller than the number of neighbors of a terminal, the problem can be recasted to a soft edge-coloring problem [59], with the aim of minimizing the number of collisions, i.e., the number of communicating nodes within 2-hops having the same assigned channel. In general, such a problem is NP-hard, but approximation algorithms exist that solve the problem with various degrees of optimality and in a distributed fashion. The rationale being that, the less conflicting channels, the less the collisions in the network. In general, the problem of efficient channel allocation in 802.11 WMNs remains hard even when only to the set of non-overlapping channels (3 out of 12) are considered [64]. Other authors [65] consider only two channels available in the network, and provide topology construction algorithms with bounded convergence time

6.2. Link Aware Routing

In order to improve scalability of WMNs, the nature of the wireless medium has to be accounted for in order to exploit efficiently the available capacity, To this respect, as already showed in ad-hoc networks literature, cross-layer techniques are able to match routing and MAC layer parameters related to the link status. But, routing metric developed for ad hoc networks barely took into account link quality and/or bandwidth exploiting instead hop count or for example residual power at each node in order to minimize battery consumption. However, WMNs are characterized by totally different requirements in terms of power efficiency and mobility. In fact, in WMNs, relaying is performed by dedicated nodes, e.g mesh routers, with limited or no mobility and no constraints on battery life being directly plugged to the power line.

A detailed description of the most relevant routing metrics in WMNs can be found in [21][22]. Here we briefly summarize their key features:

- *Hop Count*: Hop count is the simplest routing metric available. Exploiting a single bit of information for describing each link, this routing metric can either tell if the link exists or if it does not. Such an approach lead to poor performance especially in wireless network where a two-hops path over reliable links can easily outperforms a single hop path over a lossy link.

- *Per-hop Round Trip Time (RTT)*: This metric computes the RTT between neighbor nodes. In order to compute the RTT a node sends unicast probes to each neighbors that immediately responds with a probe ACK. Such a design is able to take into account both link quality and network congestion. In fact probes are delayed in each node's queue and by the IEEE 802.11 contention mechanism. As expected, over lossy links the metric will increases due to the IEEE 802.11 ARQ mechanism. It is worth noting that this metric does not take into account the link's data rate, moreover this metric require to send probes between any each pair of neighboring nodes and thus may not be suitable for dense networks.

- *Per-hop packet pair (PktPair)*: This metric aims at measuring the delay between a pair of unicast probe sent to a neighbor. The to probes have different sizes (the first is Small , while the second is large). The destination computes the delays between the two probes at report the result to the sender. Such a design is capable of taking into account both link's data rate and the loss rate. In fact an high loss rate or a slow link will increase the metric due to respectively the ARQ mechanism or the longer time required to send the large packet. The first drawback of this metric is network overhead in that two probes must be transmitted for each pair of neighbors.

- *Expected Transmission Count (ETX)*: This metric estimate the number of retransmission required to send unicast packets. ETX [20] sends periodically a broadcast probe. This probe contains the number of probes received by each neighbors during a certain observation window. By exploiting this information each node can compute the forward and reverse loss rate for each link (the IEEE 802.11 MAC does not retransmit broadcast packets). The main advantage of ETX is that it exploits broadcast probes thus lowering the network overhead. However, according to the IEEE 802.11 standard, broadcast packets are always sent at the lowest rate (11 Mbps for IEEE 802.11b/g and 6 Mbps for IEEE 802.11a). As a result since data packets are sent typically at higher rates, they may experience higher loss rate. Moreover the ETX metric does not take into account the link rate.

- *Expected Transmission Time (ETT)*: ETT is defined in [21] as a “bandwidth adjusted ETX”. In fact ETT is computed starting from ETX using this formula: $ETT = ETX \cdot S/B$ where S is the size of the probe packet and B is the link bandwidth. However, as reported in [21], computing the link bandwidth may not be trivial in real-world deployments.

- *Wighted Cumulative ETT (WCETT)*: WCETT aims at minimizing interference in multi-channel WMNs. In order to achieve this goal, WCETT assumes that two hops on the same channel will always interfere (this may not be true for long routes). According to WCETT, for a n-hop path it holds that $WCETT = (1-\beta)\sum_{i=1}^{n} ETT_i + \beta \max_{1 \le j \le k}(X_j)$ where X_j is the sum of ETTs of all hops on channel j and β is a tunable parameter between 0 and 1. Thus, WCETT is a weighted average of the total ETT over all channels and the maximum total ETT on each individual channel. By considering these two factors, WCETT considers both the bandwidth availability and channel diversity.

A performance comparison is described in [22] for a wide set of routing metrics are analyzed. Along this line, the Link Quality Source Routing (LQSR) is proposed as an extension to the Dynamic Source Routing (DSR) protocol. The LQSR selects the routing path according to the link quality. Three different performance metrics are analyzed: Expected Transmission Count (ETX) [20], per-hop RTT, and per-hop packet pair. Such performance metrics are compared with the traditional hop-count metric. Result shows that when WMNs node mobility is required, hop-count performs better than the other metrics. However, when WMNs nodes are fixed, and this is the case of WMN, more advanced performance metric taking into account link quality outperform hop-count.

However, exploiting link quality as only performance metric is not enough to properly support multimedia services. Instead, multiple routing metric involving delay, jitter and packet loss should be used. In this context a significant improvement of network scalability is expected as result of cross-layering techniques aimed at merging Network Layer and MAC layer functionalities. This is particularly true since in IEEE 802.11-based WMNs where medium access is not transparent to routing, being the transmission at a node affecting the transmission more than one-hop away.

Specific to metropolitan networks operated by Wireless ISP is the need for traffic differentiation. In such a scenario the network operator is interested in providing SLA tailored to specific customer's classes (i.e. business and residential). Differentiated Services (DiffServ) [11] and Virtual LANs (VLAN) [85] are two reference, complementary approaches aimed at satisfying QoS requirements. DiffServ is an architecture that attempts to provide service differentiation by using a class-based approach where individual application flows with similar quality requirements are aggregated. Of course, the crucial issue in WMNs is how to decide the Per-Hop-Behavior in order to describe the treatment of aggregated traffic to ensures the quality guarantees to the corresponding service class. Conversely, VLANs can be used in order to provide security by isolating traffic between different users. Down to the IEEE802.11x MAC, different services can be tagged and mapped into different priority queues to achieve statistical differentiation, in the spirit of the IEEE802.11e specification described before. Despite all the components are available, a suitable architecture for exploiting such approaches in WMNs is an open issue.

Table 4. Key features of the G.729.3 Codec

Codec	Packet Interval	Bit-rate	Payload
G.729.3	30 ms	8 kbps	30 bytes

6.3. Channel Aware Aggregation for VoIP in WMNs

In order to make a concrete example of how the above mentioned techniques increase the sustainable QoS of WMNs, a simple case study will clarify the expected improvement. In particular, in the following an example reporting of a the real testbed experiment operated for VoIP applications will demonstrate the (voice) capacity gain. Being very strict in terms of QoS guarantees for delay and jitter, in particular, VoIP services are an interesting benchmark case. Beside the probing aspect, moreover, it is well known that VoIP is per se a very well known issue in IEEE 802.11 WLANs [26][28][48], due to protocol overhead and to the carrier sense strategy, based on which transmission and collisions elongate backoff waiting times at each station. This translated in sever limitations in the number of sustained VoIP sessions, basically 2 orders of magnitude less than expected from the nominal link rate over VoIP session rate ratio. Table 4 summarizes the key features of the G.729.3 VoIP codec: compared to standard data transmissions, a typical VoIP source sends typically a large number of small packets with a large penalty in terms of protocol overhead.

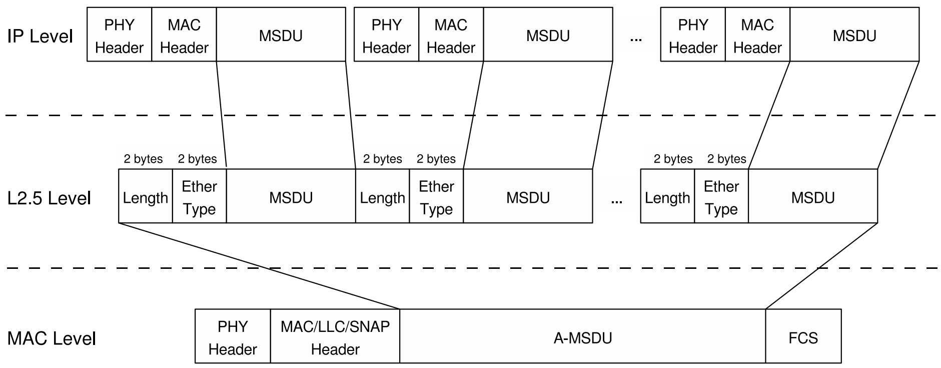

Figure 13. Aggregated MSDU (A-MSDU) frame format.

In order to mitigate the detrimental effect of the IEEE802.11 MAC overhead, a natural choice is to concatenate several MAC Service Data Units (MSDUs) to form the data payload of a larger MAC Protocol Data Unit (MPDU) [28][48]. The PHY header and the MAC header, together with the Frame Check Sequence (FCS) are then appended in order to build the Physical Service Data Unit (PSDU). The frame format for an Aggregated MSDU (A-MSDU) is sketched in Figure 13: the computational cost of the aggregation procedure is largely within the reach of commercial devices.

A typical drawback of any packet aggregation scheme is that it increases the processing delay at each node invalidating its suitability for VoIP applications: compared to the single hop scenarios, in a WMN, the aggregation policy is applied in a hop-by-hop fashion. Thus, all VoIP packets _owing towards the gateway will be aggregated and de-aggregated at each intermediate mesh router. An upper bound (20ms) to the aggregation time is introduced in order to limit the processing delay at each node.

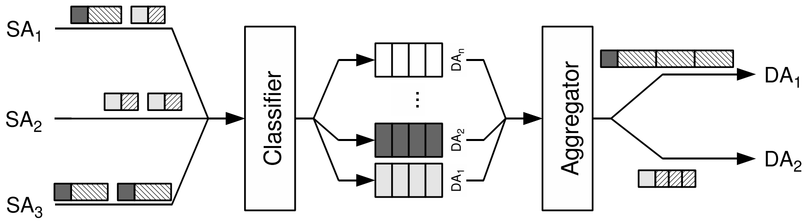

Figure 14. MSDU Aggregator architecture.

The building blocks of the MSDU aggregator and the relationships between them are sketched in Figure 14. The operations are as follows. Incoming MAC frames are first processed by the *Classifier* according to the destination address. Each flow is then fed to a different *Aggregation Queue*. For each queue, an A-MSDU is generated by the *Aggregator* when either an aggregation timer (50 ms) is expired or at least a burst of length B_{Opt} can be generated. Choosing B_{Opt} appropriately, the above scheme can be made aware of the link status. In fact, it is clear that the aggregation trades off packet length with overhead. But, over noisy channels, longer packets need to be retransmitted more due to channel errors. Thus, an optimal aggregation threshold exists: letting B_{Opt} equal to such threshold, the whole effect is to adapt the aggregation mechanism to the link conditions.

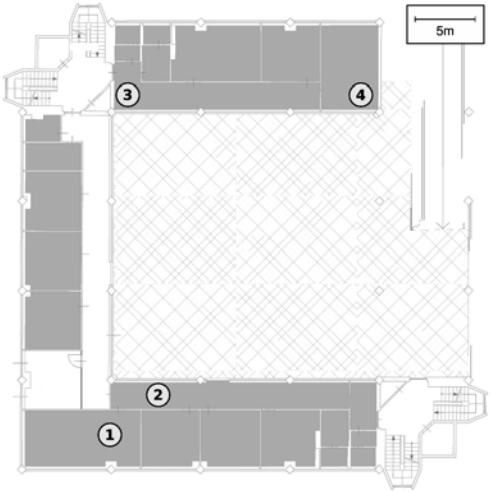

Figure 15. Planimetry. Node number one acts as gateway providing Internet connectivity to the WMN

The testbed consist of a 4 nodes deployment in a typical office environment, with a two-tier structure, as sketched in Figure 15. Testbed's nodes are based on the PCEngines 1E processor board. Each node is equipped with a 233MHz CPU, 128MB of RAM, 128MB of compact flash and one IEEE 802.11a/b/g wireless interfaces with RTC/CTS disabled (the board supports up to two wireless interfaces). The testbed planimetry is illustrated in Figure 15. Node number one acts as gateway providing Internet connectivity to the WMN. All measurements are run with the IEEE 802.11 interfaces operating in .g. mode. The testbed nodes run MIT Roofnet [50], an experimental IEEE 802.11-based WMN. Roofnet routes packets using a modified version of DSR [32] called SrcRR [10] exploiting the ETX routing metric. Routing is implemented using the Click modular router [37], developed at MIT; a Click router is built by assembling several packet processing modules, called *elements*, forming a directed graph. Each *element* is in charge of a specific function such as packet classification, queuing, and interfacing with networking devices. Click comes with an extensive library of elements supporting various types of packet manipulations. Such a library enables easy router configuration by simply choosing the elements to be used and the connections among them. In the testbed described here, the default Roofnet configuration was added two additional elements for packet aggregation and de-aggregation[1].

[1] All the developed software has been released under the BSD License (http://www.wing-project.org/).

Through measurements, one can assess the voice capacity [48], i.e. the maximum number of sustained VoIP calls with high quality and related parameters. Mean Opinion Score (MOS) tests are the traditional choice to assess the quality of conversations sustained by the network. But, determining the voice capacity would require a prohibitive amount of work if assessed via MOS, and evaluating the MOS rate for a VoIP solution besides being a time consuming process, is inherently not reproducible. For this reason, the experiments can be conducted via a simple and reproducible technique. The components of the tests are a synthetic traffic probe procedure: the probes are generated using Jugi's Traffic Generator (JTG), a freely available synthetic traffic generator [33] incorporating the typical characteristics of the traffic pattern of a G729.3 codec. In addition to being able of generating and injecting different traffic patterns over TCP and/or UDP sockets, in fact, JTG can read the information about packet transmission intervals and sizes from _les, allowing us to create an exact duplicate of a trace starting from a pre-recorder stream. Traffic is then collected at the receiver side where suitable tools are available for analysis. The analysis technique is the second component of the experimental deployment. Performance can be assessed via the E-Model [86], which provides an objective method to evaluate speech quality in VoIP systems. The outcome of an E-Model evaluation is called R-factor (R). The R-factor is a numerical measure of voice quality, ranging from 0 to 100, where the scale is basically a logarithmic mapping of objective parameters to the quality perceived by the average listeners, namely the e-model for voice quality assessment. Parameters of interest for the R-factors are collected in dedicated ITU-T recommendations [86][87][88]. According to such recommendations, R = 70 is the minimum tolerated value to obtain a VoIP call with acceptable quality.

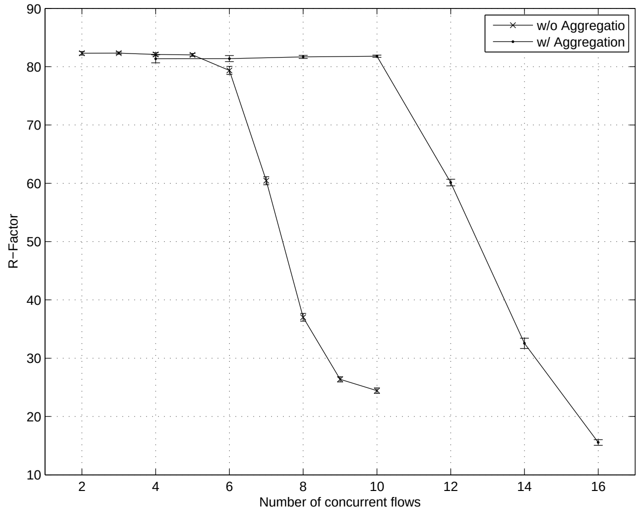

Figure 16. Performance measurements results

The testbed is a typical configuration for a WMN: several concurrent VoIP flows exploit the mesh nodes as their gateways towards the Internet. VoIP calls are thus UDP flows modeled according to the parameters of the G.729.3 VoIP codec (see Table. I). All mesh nodes sustain the same traffic, consisting in an increasing number of VoIP sessions. All measurements were performed over 3 minutes intervals; results are averaged over 10 runs. The results in Figure 16 show clearly aggregation improves the quality of the services sustained over WMNs. The increment on the number of served VoIP flows is indicated in is almost 100% for the 4 nodes deployed. Hence, even a simple design as the one proposed above doubles the number of sustained voice sessions. This measurement is in line with predictions obtained from simulations [48], where aggregation techniques were applied to similar but single-hop scenarios.

7. Summary and Future Trends

Resource management in IEEE 802.11 based wireless networks has been extensively studied. Many QoS schemes have been proposed to support bandwidth/delay constrained multimedia applications such as voice/video in Wi-Fi based wireless LANs, heterogeneous wired/wireless networks, mobile ad hoc networks, and wireless mesh networks. These schemes significantly improved flow performance as well as network capacity by exploring either MAC layer information such as link quality, channel diversity, and link interference, or network layer information such as neighboring nodes and collected paths, or both. Due to the contention based channel access protocol in IEEE 802.11 DCF MAC and the broadcast nature of wireless medium, it is imperative that MAC layer information should be tuned as well as incorporated to assist higher layer decisions such as route selection for maximum network performance. From this aspect, cross-layer approaches that combine the information at both MAC layer and network layer have good potentials. On the other hand, wireless technologies have been advanced tremendously in recent years. New wireless access technologies such as Bluetooth, Wi-Max, UMTS, and UWB have gained a lot of applications and deployments. With such ubiquitous wireless network environments, how to integrate QoS schemes of different wireless network domains together so that multimedia application can be supported seamlessly is critical. Furthermore, many protocols are proposed to solve a specific issue based on certain protocols (such as AODV-QoS). Due to this type protocol dependence, the benefit to the wireless users are very limited since most protocols are designed for one specific layer or issue without considering the needs of other layers or issues. In order to maximize the contribution from the research community, it is imperative that existing protocols are combined effectively and efficiently. Being aware of the heterogeneity of wireless technologies and traffic characteristics, we feel that it is impossible to propose a unique solution for this purpose. Instead, middleware technology should be adopted with careful design.

Therefore, we expect that following research works on the topic of resource management in IEEE 802.11 based wireless networks:

- *QoS in the emerging WMNs:* Efficient schemes to provide guaranteed QoS to VoIP and video applications will be of paramount importance. More in general, several techniques are expected to mitigate the impact of lack of scalability on QoS provisioning in WMNs.
- *QoS in ubiquitous wireless access networks:* QoS and Mobility Management in hybrid mobile/wireless networks where WLAN coexists with other networks such as MANET, WMN, Bluetooth, Wi-Max, UWB, and 3G networks.
- *QoS middleware:* Given a rich set of related, architecturally similar, and loosely connected QoS protocols proposed already by the research community, a cross-layer middleware architecture will (i) significantly increase the ease of design, implementation, validation, and comparison of existing and new QoS protocols. (ii) stimulate collaboration among researchers in the same field with a generic platform. (iii) provide a built-in resource management module in the future wireless networks.

These efforts, combined, will further improve the QoS performance and make the next generation wireless access networks more suitable for the support of multimedia applications.

REFERENCES

[1] Aad, I., Ni, Q., & Castelluccia, C. (2002). Enhancing IEEE 802.11 performance with slow CW decrease. IEEE 802.11e working group document 802.11-02/674r0.

[2] Adya A., Bahl, P., & Padhye, J. (2004). A multi-radio unification protocol for IEEE 802.11 wireless networks. In Proceedings of BroadNets, San Jose, California, USA.

[3] Ahn, G., Campbell, A. T., & Veres, A. (2002). Supporting Service Differentiation for Real-Time and Best Effort Traffic in Stateless Wireless Ad Hoc Networks (SWAN), IEEE Transactions on Mobile Computing, Volume 1, Issue 3.

[4] Akyildiz, I. F., Wang, X., & Wang, W. (2005). Wireless mesh networks: a survey. Computer Networks and ISDN Systems, Volume 47, Issue 4.

[5] Awerbuch, B., Holmer, D., & Rubens, H. (2004). High Throughput Route Selection in Multi-Rate Ad Hoc Wireless Networks. In Proceedings of 1st Working Conference on Wireless On-demand Network Systems (WONS 2004), Madonna di Campiglio, Italy.

[6] Banchs, A., & Pérez, X. (2002). Providing throughput guarantees in IEEE 802.11 wireless LAN. In Proceedings of IEEE Wireless Communications and Networking Conference (WCNC2002), Vol. 1, pp. 130-138.

[7] Banchs, A., & Pérez, X., & Qiao, D. (2003). Providing Throughput Guarantees in IEEE 802.11e Wireless LANs. In Proceedings of 18th International Teletraffic Congress, Berlin, Germany.

[8] Barry, M., Campell, A. T., & Veres, A. (2001). Distributed control algorithms for service differentiation in wireless packet networks. In Proceedings of IEEE INFOCOM.

[9] Bianchi, G. (2000). Performance analysis of the IEEE 802.11 distributed coordination function. IEEE Journal of Selected Areas on Commununications, Volume 18, Issue 3, pp. 535–547.

[10] Bicket, J., Aguayo, D., & Biswas, S. (2005). Architecture and Evaluation of an Unplanned 802.11b Mesh Network. In Proceedings of ACM MOBICOM, Cologne, Germany.

[11] Blake, S., Black, D., & Carlson, M. (1998). An Architecture for Differentiated Service. RFC 2475.

[12] Braden, R., Clark, D., & Shenker, S. (1994). Integrated Services in the Internet Architecture: an Overview. RFC 1633.

[13] Braden, R., Zhang, L., & Berson, S. (1997). Resource Reservation Protocol (RSVP)-Version 1 Functional Specification, RFC 2205.

[14] Buddhikot, M., Chandranmenon, G., & Han, S. (2003). Integration of 802.11 and Third-Generation Wireless Data Networks. In Proceedings of IEEE INFOCOM.

[15] Chakeres, I. D., & Belding-Royer, E. M. (2004). PAC: Perceptive Admission Control for Mobile Wireless Networks. In Proceedings of the 1st International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (QShine), Dallas, TX, USA.

[16] Chen, L., & Heinzelman, W. (2005). QoS-aware routing based on bandwidth estimation for mobile ad hoc networks. IEEE Journal on Selected Areas of Communication, Vol. 23, No. 3.

[17] Chen, S., & Nahrstedt, K. (1999). Distributed Quality-of-Service Routing in Ad-Hoc Networks. IEEE Journal on Selected Areas in Communications, Special Issue on Ad-Hoc Networks, Vol. 17, No. 8, pp. 1-18.

[18] Clausen, T. & Jacquet, P. (2003). Optimized Link State Routing Protocol (OLSR)", RFC 3626.

[19] Conti, A., Dardari, D., & Pasolini, G. (2003). Bluetooth and IEEE 802.11b Coexistence: Analytical Performance Evaluation in Fading Channels. IEEE Journal on the Selected Areas in Communications, Vol. 21, pp. 259-269.

[20] De Couto, D., Aguayo, D., & Bicket , J. (2003). High-throughput path metric for multi-hop wireless routing. In ACM MOBICOM, San Diego.

[21] Draves, R., Padhye, J., & Zill, B. (2004). Comparison of Routing Metrics for Static Multi-Hop Wireless Networks. In Proceedings of ACM SIGCOMM, Portland, Oregon, USA.

[22] Draves, R., Padhye, J., & Zill, B. (2004). Routing in multi-radio, multi-hop wireless mesh networks. In Proceedings of ACM MobiCom, Philadelphia, PA, USA.

[23] Ergen, M., & Varaiya. P. (2005). Throughput Analysis and Admission Control for IEEE 802.11a. Mobile Networks and Applications. 10, pp. 705-716.

[24] Gao, D., Cai, J., & Ngan, K. N. (2005). Admission control in IEEE 802.11e wireless LANs. IEEE Network, special issue on Wireless Local Area Networking: QoS provision & Resource Management, Vol. 19, No. 4, pp. 6-13.

[25] García-Macías, J. Rousseau, A., F., & Berger-Sabbatel, G. (2003). Quality of Service and Mobility for the Wireless Internet. Wireless Networks, Vol. 9, pp. 341-352.

[26] Garg, S., & Kappes, M. (2003). Can I add a VoIP Call? In IEEE International Conference on Communications, Anchorage, USA.

[27] Goff, T., Abu-Ghazaleh, N. B., & Phatak, D. S. (2001). Preemptive Routing in Ad Hoc Networks. In Proceedings of ACM MobiCom.

[28] Ganguly, S., Kim, N. V, & Kashyap, K. (2006). Performance optimization for deploying voip services in mesh networks. IEEE Journal on Selected Areas in Communications, Vol. 24, No. 11, pp. 2147-2158.

[29] Gupta, P., & Kumar, P. R. (2000). The capacity of wireless networks. IEEE Transactions on Information Theory, Vol. IT-46, No. 2, pp. 388-404.

[30] Havinga, P. J. M., & Wu, G. (2001). Wireless Internet on Heterogeneous Networks. In Proceedings of Workshop on Mobile Communications in Perspective, Enschede, the Netherlands.

[31] Jaseemuddin, M. (2003). An Architecture for Integrating UMTS and 802.11 WLAN Networks. IEEE Symposium on Computers and Communications (ISCC' 03), pp. 716 -723.

[32] Johnson, D. B., Maltz, D.A., & Broch, J. (2001). DSR: The Dynamic Source Routing Protocol for Multi-Hop Wireless Ad Hoc Networks. In C. E. Perkins (Ed.), Ad Hoc Networking, pp. 139-172, Addison-Wesley Press.

[33] Jugi. From https://hoslab.cs.helsinki.fi/savane/projects/jtg/

[34] Jun, J., & Sichitiu, M. L. (2003). The nominal capacity of wireless mesh networks. IEEE Wireless Communications, vol. 10, no. 5, pp. 8-14.

[35] Kanodia, V., Li, C., & Sabharwal, A. (2002). Distributed Priority Scheduling and Medium Access in Ad Hoc Networks. ACM Wireless Networks Journal, Vol.8, pp. 455-466.

[36] Kazantzidis, M., Gerla, M., & Lee, S. J. (2001). Permissible throughput network feedback for adaptive multimedia in AODV MANETs. IEEE International Conference on Communications (ICC'01), Vol. 5, pp. 1352-1356, June 11-14.

[37] Kohler, E., Morris, R., & Jannotti, J. (2000). The Click modular router. ACM Transaction on Computer System, Vol. 18, No. 3, pp. 263-297.

[38] Lamont, L., Wang, M., & Villasenor, L. (2003). Integrating WLANs & MANETs to the IPv6 Based Internet. In Proceedings of IEEE International Conference on Communications (ICC'03), Vol. 2, pp. 1090-1095.

[39] Lee, S. B., Ahn, G., & Zhang, X. (2000). INSIGNIA: AnIP-Based Quality of Service Framework for Mobile Ad Hoc Networks. Journal of Parallel and Distributed Computing, special issue on Wireless and Mobile Computing and Communications, Vol. 60, No. 4pg. 374-406.

[40] Li, J., Blake, C., & De Couto, D. S. J. (2001). Capacity of ad hoc wireless networks. ACM MobiCom, pp. 61-69.

[41] Li, M., Prabhakaran, B., & Sathyamurthy, S. (2003). On Flow Reservation and Admission Control for Distributed Scheduling Strategies in IEEE 802.11 Wireless LAN, Proceeding of the 6th ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM'03). San Diego, CA, USA.

[42] Li, M., & Prabhakaran, B. (2004). Dynamic Priority Re-allocation Scheme for Providing Quality of Service in IEEE 802.11e WLANs. In Proceeding of SPIE/ACM Conference on Multimedia Computing and Networking (MMCN'04), Santa Clara, CA, USA.

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

[44] Li, M., Zhu, H., & Chlamtac, I. (2006). End-to-end QoS Framework for Heterogeneous Wired-cum-Wireless Networks. ACM/Baltzer WINET, Volume 12, Issue 4.

[45] Li, M., & Prabhakaran, B. (2005). On Supporting Reliable QoS in Multi-hop Multi-rate Mobile Ad Hoc Networks. In Proceedings of the First IEEE International Workshop on Next Generation Wireless Networks (WoNGeN'05), Goa, India.

[46] Lo, C.-C., & Lin, M.-H. (1998). QoS provisioning in handoff algorithms for wireless LAN. In Proceedings of the International Zurich Seminar on Broadband Communications, Accessing, Transmission, Networking, Zurich.

[47] Lou, X., Li, B., & Thng, I. (2002). An Adaptive Measured-based Pre-Assignment Scheme with Connection-level QoS Support for Mobile Networks. IEEE Trans. on Wireless Communication, Vol. 1, Issue 3, pp. 521-530.

[48] Maguolo, F., Pellegrini, F. D., & Zanella, A. (2006). Cross-layer solutions to performance problems in VoIP over WLAN. In Proceedings of EURASIP EUSIPCO, Florence, Italy.

[49] Mishra, M. S. A., & Arbaugh, W. (2003). An Empirical analysis of the IEEE 802.11 MAC Layer Handoff Process. ACM SIGCOMM Computer Communication Review, 33(2):93-102.

[50] MIT, Roofnet Project, from http://pdos.csail.mit.edu/roofnet/

[51] Moon, B., & Aghvami, A. H. (2004). Quality of Service Mechanisms in All-IP Wireless Access Networks. IEEE Journal on Selected Areas in Communications, Vol.22, No.5, pp.873 -888.

[52] Nasipuri, A., Zhuang, J., & Das, S. R., A Multichannel CSMA MAC Protocol for Multihop Wireless Networks. In Proceedings of IEEE Wireless Communications and Networking Conference (WCNC).

[53] Park, H., Yoon, S., & Kim, T. (2002). Vertical Handoff Procedure and Algorithm between IEEE 802.11 WLAN and CDMA Cellular Network. Mobile Communications: 7th CDMA International Conference (CIC), Seoul, Korea.

[54] Park, S., Kim, K., & Kim, D. C. (2003). Collaborative QoS Architecture between DiffServ and 802.11e Wireless LAN. In Proceedings of IEEE VTC'03-Spring, Jeju, Korea.

[55] Perkins, C. E., & Bhagwat, P. (1994). Highly dynamic destination sequenced distance vector routing (DSDV) for mobile computers. ACM SIGCOMM.

[56] Perkins, C. E., & Belding-Royer, E. M. (2003). Quality of Service for Ad hoc On-Demand Distance Vector Routing. Internet Draft, draft-perkins-manet-aodvqos-02.txt.

[57] Perkins, C. E., Belding-Royer, E. M., & Chakeres, I. (2003). Ad hoc on demand distance vector (AODV) routing," IETF Internet draft, draft-perkins-manet-aodvbis-00.txt.

[58] Pong, D., & Moor T. (2003). Call admission control for IEEE 802.11 contention access mechanism. In Proceedings of IEEE Globecom.

[59] Raman, B. (2005). Channel Allocation in 802.11-based Mesh Networks. In Proceedings of IEEE INFOCOM, Miami, USA,.

[60] Romdhani, L., Ni, Q., & Turletti, T. (2003). Adaptive EDCF: Enhanced Service Differentiation for IEEE 802.11 Wireless Ad Hoc Networks. IEEE Wireless Communications and Networking Conference (WCNC), New Orleans, USA.

[61] Shah, S. H., Chen, K., & Nahrstedt, K. (2004). Dynamic Bandwidth Management for Single-hop Ad Hoc Wireless Networks. ACM/Kluwer MONET, Special Issue on Algorithmic Solutions for Wireless, Mobile, Ad Hoc and Sensor Networks.

[62] Shankar, S., & Choi, S. (2002). QoS Signaling for Parameterized Traffic in IEEE 802.11e Wireless LANS. *Lecture Notes in Computer Science*, Vol. 2402, pp. 67-84.

[63] Sanzgiri, K., Chakeres, I. D., & Belding-Royer, E. M. (2004). Determining Intra-Flow Contention along Multihop Paths in Wireless Networks. In proceedings of IEEE Broadnets Wireless Networking Symposium, San Jose, CA.

[64] Shin, M., Lee, S., & Kim, Y. (2006). Distributed Channel Assignment for Multi-radio Wireless Networks, In Proceedings of International Conference on Mobile Adhoc and Sensor Systems (MASS), Vancouver, Canada.

[65] Ju, H.-J., & Rubin, I. (2006). Backbone Topology Synthesis for Multiradio Mesh Networks. IEEE Journal on Selected Areas in Communications, Vol. 24, No. 11.

[66] Shin, S., Forte, A., & Rawat, A. (2004). Reducing MAC Layer Handoff Latency in IEEE 802.11 Wireless LANs. In Proceeding of ACM MobiWAC.

[67] So, J., & Vaodya, N. (2004). Multi-channel mac for ad hoc networks: Handling multi-channel hidden terminals using a single transceiver. In Proceedings of ACM Mobihoc, Roppongi, Japan.

[68] Vaidya, N. H., Bahl, P., & Gupta, S. (2000). Distributed Fair Scheduling in Wireless LAN. In Proceedings of ACM MOBICOM, pp. 167-178, Boston, USA.

[69] Valaee, S., & Li, B. (2002). Distributed call admission control in wireless ad hoc networks. In Proceedings of IEEE Vehicular Technology Conference (VTC), Vancouver, British Columbia.

[70] Wu, H., Peng, Y., & Long, K. (2002). Performance of reliable transport protocol over IEEE 802.11 wireless LANs: Analysis and enhancement. In Proceedings of IEEE INFOCOM, pp. 599–607, New York, USA.

[71] Xiao, Y. (2005). Performance analysis of priority schemes for IEEE 802.11 and IEEE 802.11e wireless LANs. IEEE Transactions on Wireless Communications, Vol. 4, No. 4, pp. 1506–1515.

[72] Xiao, Y., Li, H., & Choi, S. (2004). Protection and Guarantee for Voice and Video Traffic in IEEE 802.11e Wireless LANs. In Proceedings of IEEE INFOCOM, pp. 2153-2163.

[73] Xiao Y., & Li, H. (2004). Local Data Control and Admission Control for Ad Hoc Wireless Networks. IEEE Transactions on Vehicular Technology, Vol. 53, No. 5, pp.1558-1572.

[74] Xiao Y., & Li, H. (2004). Voice and Video Transmissions with Global Data Parameter Control for the IEEE 802.11e Enhance Distributed Channel Access. IEEE Transactions on Parallel and Distributed Systems, Vol. 15, No. 11, pp.1041-1053.

[75] Yang, Y., & Kravets, R. (2005). Contention-Aware Admission Control for Ad Hoc Networks. IEEE Transactions on Mobile Computing, Vol. 4, Issue 4, pp. 363-338.

[76] Xue, Q., & Ganz, A. (2003). Ad hoc QoS on-demand routing (AQOR) in mobile ad hoc networks. Journal of Parallel Distributed Computing, Vol. 63, pp. 154-165.

[77] Yavatkar, R., Hoffman, D., & Bernet, Y. (2000). SBM (Subnet Bandwidth Manager): A Protocol for RSVP-based Admission Control over IEEE 802-style networks. RFC2814.

[78] Zhai, H., Chen, X., & Fang, Y. (2006). A Call Admission and Rate Control Scheme for Multimedia Support over IEEE 802.11 Wireless LANs. ACM Wireless Networks, Vol. 12, No.4, pp. 451-463.

[79] Zhu, H., & Chlamtac, I. (2003). An analytical model for IEEE 802.11e EDCF differential services. In Proceedings of the 12th International Conference on Computer Communications and Networks (ICCCN), Dallas, TX, USA.

[80] Zhu, H., Li, M., & Chlamtac, I. (2004). Survey of Quality of Service in IEEE 802.11 Networks. IEEE Wireless Communications, Special Issue on Mobility and Resource Management, Vol. 11, No. 4, pp. 6-14.

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

[82] IEEE 802.11b (2003). Part 11: Wireless LAN Medium Access Control (MAC) and physical layer (PHY) specifications: Medium Access Control (MAC) Enhancements for Quality of Service (QoS), IEEE Std 802.11e/D4.3.

[83] IEEE 802.11f (2003). IEEE Trial-Use Recommended Practice for Multi-Vendor Access Point Interoperability via an Inter-Access Point Protocol Across Distribution Systems Supporting IEEE 802.11 ™ Operation.

[84] IEEE 802.11s (1998). Draft Amendment to Standard for Information Technology, Telecommunications and Information Exchange Between Systems - LAN/MAN Specific Requirements - Part 11: Wireless Medium Access Control (MAC) and physical layer (PHY) specifications: Amendment: ESS Mesh Networking, IEEE Std.

[85] IEEE 802.1Q (2003). Virtual Bridged Local Area Networks, IEEE Std..

[86] Recommendation G.107 (2005). The E-model, a computational model for use in transmission planning, ITU-T Std..

[87] Recommendation G.113 (2005). Transmission impairments due to speech processing, ITU-T Std..

[88] ITU-T Recommendation G.729 (1996). Annex B, A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70.