

# Asynchronous Reservation-Oriented Multiple Access for Wireless Networks (AROMA)

Mustafa Özdemir  
RWIN-Lab ECE Dept.  
Northeastern University  
Boston, MA 02115

Ram Ramanathan  
Internetworking Research Dept.  
BBN Technologies  
Cambridge, MA 02138

A. Bruce McDonald  
RWIN-Lab ECE Dept.  
Northeastern University  
Boston, MA 02115

*Abstract—Supporting multimedia streams (e.g. voice/video over IP), in a Wireless LAN requires guaranteed access to channel capacity, which is best provided using capacity reservations. Existing Medium Access Control (MAC) protocols for Wireless LANs either do not provide reserved access or do so in a synchronous (e.g. polling, TDMA) context. A novel mechanism for reserved access within an asynchronous or contention-based paradigm (for example CSMA/CA, or 802.11 DCF) is presented. The protocol, called Asynchronous Reservation Oriented Multiple Access (AROMA), allows a client to specify a capacity request using a moving window or leaky bucket descriptor. The server performs admission control, reserves the capacity and provides access to the capacity in a simple, efficient manner. AROMA is backward compatible with IEEE 802.11 DCF. The performance of AROMA using theoretical and experimental analysis is analyzed. The theoretical analysis shows that the AROMA parameters can be optimized using the analytical model in the case of saturated traffic. Experimentally, the comprehensive simulations using an enhancement of OPNET’s 802.11 model, shows that AROMA can support more VoIP calls with a better QoS than IEEE 802.11e EDCF.*

## I. Introduction

The recent proliferation in the deployment of wireless LANs (WLANs) has been unprecedented. Facilitated largely through widespread broadband Internet access and a robust consumer electronics market, demand for innovative WLAN systems has outpaced industry and standards bodies’ ability to support key applications. Most notable has been the parallel shift towards the use of packetized voice and video. Demand for faster Internet access and the availability of DVD players costing less than \$30 have brought MPEG video within reach of millions of consumers. Moreover, corporate users are increasingly turning to vendors for voice-over-IP (VoIP) solutions to replace legacy PBX systems. Given these trends, the need for real-time communications over WLANs will likely become increasingly pronounced in the near future.

The most common WLAN systems today are based on

the IEEE 802.11 protocol standards (e.g. IEEE 802.11 b/g/a), which lack support for real-time applications such as voice, video and multimedia. Specifically, 802.11 does not support differentiation in services to distinct application data streams. Hence, it is infeasible to use it for applications that require assured Quality-of-Service (QoS). In current WLAN systems users contend (statistically) for access to a shared, broadcast channel. The approach is adequate for non-real-time applications, however, an ongoing videoconference could be seriously disrupted in the event of a simultaneous file transfer or web access by other users. To prevent this QoS differentiation must be provided at the Medium Access Control (MAC) layer.

The objective of this paper is to present AROMA - The Asynchronous Reservation Oriented MAC for wireless networks; AROMA is the first wireless MAC mechanism that provides reserved access in an asynchronous protocol. AROMA enhances CSMA/CA in general and 802.11 DCF in particular to enable capacity request characterization, reservation and admission control. AROMA allows non-QoS traffic access to residual capacity in a best-effort manner. A specific instantiation of AROMA is described for IEEE 802.11 based networks, which is backward compatible with IEEE 802.11 DCF. AROMA is vastly simpler than synchronous-based protocols such as 802.11e HCCA. Finally, although the treatment of AROMA in this paper is restricted to access point (AP) based WLANs, it is extensible to ad hoc networks because it is based on a distributed protocol. Before describing AROMA, it is instructive to set in context current approaches to QoS and MAC, and the respective tradeoffs.

Approaches to QoS provision can be broadly classified as *prioritized* or *reserved*. In prioritized access, users are classified into multiple priority classes and the network services higher priority traffic before lower priority traffic. In reserved access, users request capacity that is dedicated and guaranteed to them. If the resources are available, while the tradeoffs between reserved and

prioritized access have and continue to be hotly debated, a few indisputable points stand out: With prioritized access, a node or session is provided better service than the competition, but is not guaranteed specific capacity. Thus, if other nodes/streams of equal or greater priority join the network, service will degrade. Reserved access not only provides capacity guarantees, but also admission control, and typically also request characterization. Multimedia streams such as VoIP are very sensitive to capacity fluctuations – unlike data flows, it is far better to refuse admission to a voice stream in the first place, rather than admit it and make it susceptible to capacity fluctuations [10].

Wireless MAC can be broadly classified as *asynchronous* (or contention-based, or unscheduled) or *synchronous* (or contention-free, or scheduled). Asynchronous protocols include ALOHA, CSMA/CA, 802.11 DCF etc. Synchronous protocols include TDMA, token-based access, 802.11 PCF etc. Synchronous techniques are harder to implement, are inefficient when traffic is bursty, and depend on centralized control – witness the failure of 802.11’s PCF to gain traction<sup>1</sup> and the limited popularity of Hiperlan.

The philosophy adopted in this work is that the best approach combines reserved access within an asynchronous context. Conventional thinking, unfortunately, has consistently considered prioritized QoS in the context of asynchronous, or contention based medium access, and relegated the development of reserved access to synchronous, or contention-free techniques. The 802.11 QoS extensions (IEEE 802.11e) have the same bias - prioritized access is an extension of the DCF and reserved access is an extension of the PCF. The reason for this is that contention free techniques such as scheduling (TDMA) and polling lend themselves more easily to resource allocation. For example, reserved access in TDMA (or polling) corresponds simply to assigning a certain set of slots per frame (or polls per unit time) to a particular user.

Based on the previous arguments the architecture for AROMA took shape: AROMA is a protocol between two logical entities - a *client* (reservation requester) and *server* (reservation granter). In a Wireless LAN, the client is typically the wireless terminal or host and the server is the base station or access point. In general, AROMA is applicable to any wireless network where a client-server relationship can be established.

The basic idea behind AROMA is to leverage the existing RTS-CTS handshake to control access to the

<sup>1</sup>Notably, they are optional, are not well understood, and are rarely available in commercial Wireless LAN products since PCF is not a requirement for WiFi compatibility.

channel by denying clients a CTS when they have sent more than their agreed-upon share of packets. Thus, CTSs double as resource-granting “tokens” allotted by the server to clients in accordance with the client’s reservation. The higher the bandwidth reservation of a client, the more CTSs that client is eligible to receive per unit time, hence, the more packets that client can send. An “account” is created upon the initial reservation and is “credited” over time as per reservation, and debited when the client sends packets. A new reservation is accepted when the sum total of reservations is below capacity.

AROMA adapts many ideas developed for traffic characterization and congestion control at the network layer [12], including traffic descriptors such as moving window and leaky bucket descriptors, and token bucket management. The identification of CTSs as resources that can be controlled and the realization of the correspondence between CTSs and tokens in this context is a key insight behind the ability to adapt these schemes. To the authors’ knowledge this is the first application of token bucket within a WLAN MAC context.

The remainder of the paper is organized as follows: Related work and the unique contributions of AROMA are discussed in section II. A detailed description of AROMA follows in section III. Theoretical Analysis for managing the admission control process is presented in section IV. The analytical model is validated and performance analysis based on discrete-event simulator follows in section V-B. Finally, conclusions are presented in section VI.

## II. Related Work

In the wired network, IETF’s Integrated Services [14] and Differentiated Services architectures [2], [18] are available to support guaranteed QoS and traffic prioritization respectively above the link layer. However, no mature standards for QoS provision exist for wireless LANs. The upcoming standard 802.11e [16] tries to deal with the QoS problems faced by wireless LANs based on 802.11 specifications.

The QoS support in 802.11e is provided in two forms. First, it supports a priority based best-effort service, termed EDCAF, that is based on DCF. Second, it supports parameterized QoS, termed HCCA, that is based on PCF, for the benefit of the applications requiring QoS for different flows. The two are integrated together using the Hybrid Coordination Function (HCF).

Considerable research exists outside of the standards bodies. For prioritized access in WLANs, a black-burst (BB) contention method was proposed to minimize delay for real-time traffic [1]. The main goal of BB is to minimize the delay for real-time traffic. It imposes certain

requirements on the high priority stations. The stations jam the medium with pulses of energy, denominated BB's. Other researchers have used the method of dividing different idle periods to allocate different priority for different data streams in order to reduce collision and delay and to improve the reliability and QoS data service. [13],[7],[3].

AROMA offers a unique combination of features not provided by previous work. In contrast to the PCF, "polling based" solutions such as 802.11e HCCA and Rether, or TDMA based solutions, AROMA operates in an asynchronous context. The majority of WLANs use the simpler asynchronous DCF mode of operation, hence, AROMA is more useful in practice given today's marketplace. Moreover, unlike the 802.11e EDCF, AROMA offers guaranteed service. Prioritized service can not guarantee a specific capacity, rather it attempts to assure "better" service. Further, AROMA provides the ability to describe capacity requests using the same descriptors that are used at the higher layers, in wireline networks. This facilitates vertical integration of QoS functionality. AROMA provides admission control which ensures the network is operated below the point of congestion. Existing solutions are either synchronous, or fail to provide reserved capacity. AROMA is the first WLAN MAC mechanism that provides guaranteed resources, yet is still asynchronous (contention-based). Finally, AROMA remains compatible with DCF mode of IEEE 802.11b/g, hence, can interoperate in a legacy environment.

### III. AROMA Description

AROMA is based on the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) approach, which is the basis for the IEEE 802.11 wireless LAN standard. Although AROMA applies to most variants of the CSMA/CA approach (such as MACA [11], FAMA [9], and MACAW [17]), it is presented in the context of 802.11 DCF as that is the most prevalent version. First, a brief summary of the 802.11 DCF is given, and then describe the AROMA enhancements is described.

The IEEE 802.11 DCF uses up to four frames for each data packet transfer. A sender first transmits a Request-to-Send (RTS), and the receiver responds with a Clear-to-Send (CTS). Then the sender sends the DATA and finally the receiver completes the transaction with an Acknowledgment (ACK). Both RTS and CTS contain the proposed duration of the data frame. Nodes located in the vicinity of the sender and the receiver that overhear one or both of the RTS/CTS store the duration information in a *network allocation vector (NAV)*, and defer transmission for the proposed duration. The IEEE 802.11 protocol senses the

medium and waits for it to be idle before transmitting. Collisions are resolved using a backoff mechanism. A contention window (CW) is maintained by each node and the random backoff interval is chosen from within this window. The CW is increased exponentially when frames are lost (as discerned by the non-receipt of CTS or ACK). The description is necessarily brief, and readers are referred to [5] for further details.

AROMA is a protocol between two logical entities - a client (reservation requester) and a server (reservation granter). In general, it is to any wireless network where a client-server relationship can be established. For example, in a cellular network the hand set contains the client and the base station the server. In an ad hoc network, every node and its immediate downstream node on a flow constitute, respectively, a client-server pair. Intermediate nodes are both a server for the upstream node and client for the downstream node. The AP based client-server interaction is illustrated in Figure 1, and summarized below. Sections following the summary elaborate on each aspect of AROMA.

When a client needs to reserve capacity, it sends a modified RTS referred to as the *Reservation-RTS (R-RTS)*. The purpose of the R-RTS is to let the server know that the ensuing data packet contains information about the reservation request of this client. The data packet following the R-RTS includes the requested capacity. Two well-known traffic descriptors are used for this purpose, which are elaborated on in section III-A. R-RTS is sent to the server. Retransmission rules for R-RTS are identical to the existing rules for regular RTS, as given in IEEE 802.11 standard. Upon receipt of an R-RTS, the server sends a CTS to the client and waits for the data packet, which includes the reservation request information.

Upon receiving the data packet, the server checks to see if the request can be accommodated by executing the *flow reservation* functionality, as described in subsection III-B. This determines, based upon the reservation request and existing reservations, whether this flow can be admitted or not. If flow is admitted, the reservation is successfully made and the server sends an ACK to the client. Both the server and client create "soft-state" regarding the reservation. If the request cannot be accommodated, the ACK is not sent and the client times out and follows base IEEE 802.11 rules.

Upon receipt of a RTS, the server first looks up the reservation state using the client identifier and decides whether or not to send a CTS by executing the *packet admission* functionality, as described in the subsection III-C. If the RTS is from a client that exceeded its reserved

capacity, or if it has no reservation, it is treated as “best-effort”, that is, a CTS is returned if there is residual capacity.

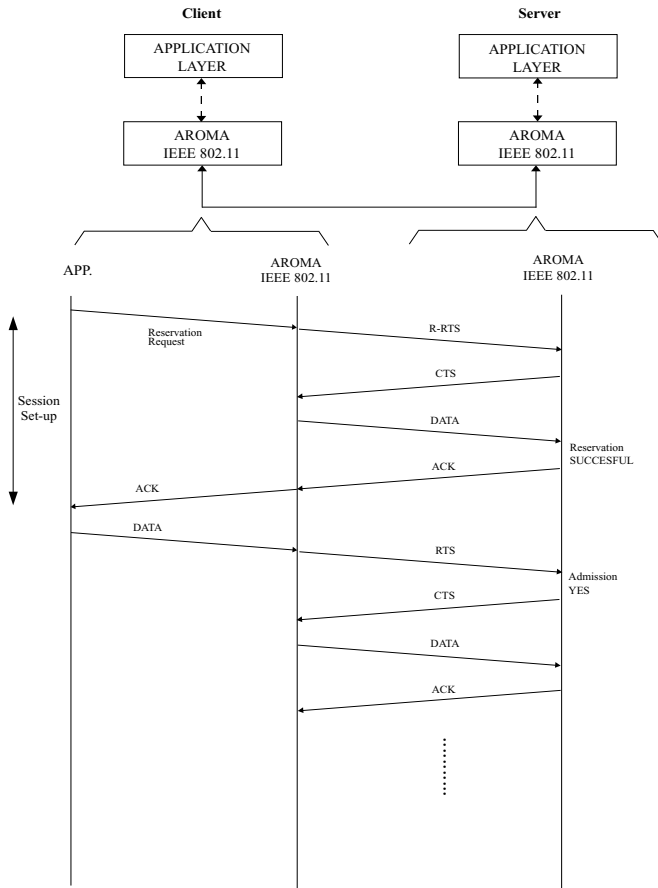


Fig. 1. AROMA Client-Server Interaction

Reservations are node-specific, identified by the client id. If there are multiple sessions at a node requiring reservations, a multiplexing and admission scheme is needed at the node to ensure sessions get the required service. This is beyond the scope of AROMA. At the server, the reservations are per-client, and therefore in case of multiple sessions, an aggregate reservation is made and tracked. (Although per session state is easy protocol-wise, it is more difficult to make it IEEE 802.11 compatible, and hence has been relegated for future work).

All traffic does not require guaranteed capacity. Clients may send packets without having a prior reservation. There may also be clients that are not AROMA compatible which behave as per the 802.11 specification. Packets that do not have a reservation associated with them are treated on a “best effort” basis. That is, CTS is returned for RTS if and only if there is available capacity. Reservations

time out after a configured period of time. That is, a reservation is purged at the server if no packet arrives for a predetermined period of time. The client’s packets are then treated as “best effort”. If the purging of the reservation turned out to be premature, the client must determine to re-establish a reservation.

### A. Traffic Descriptors

A traffic descriptor is a set of parameters used to describe the behavior of a source of traffic [12]. AROMA accommodates two kinds of descriptors, both of which are popular in the wired networking literature:

- *Moving Window*. This descriptor has two parameters - Bits (B) and Time (T). The source is allowed to send at most B bits over all windows of length T seconds. The average rate of this source is B/T bits/sec.
- *Leaky Bucket*. Also called the *linear bounded arrival process (LBAP)*, this descriptor has two parameters: Rate (R) and Burst (B). The source is allowed to send at most  $R.t + B$  bits over any given interval of t seconds. The average rate of this source is R, with a deviation of B bits possible “occasionally”.

Reader are referred to [12] for a detailed description of these descriptors . In AROMA, the leaky bucket is used as the primary descriptor and the moving window is mapped to it, thus, both are accommodated. The rate R is further decomposed into token size and token rate. Thus, there are three parameters, Token Size (TS), Token Rate (TR), and Burst (Bu).  $TS*TR$ , which is R, corresponds roughly to long-term average rate allocated by server to the client, and Bu is the longest burst a client may send. Setting the token- bucket limit to one token and replenishing the bucket at the average rate makes it a moving-window descriptor.

### B. Flow Reservation

Upon initialization, and periodically, the server determines the aggregate *effective* capacity  $B_{eff}$  of the medium. This is the capacity seen by the application and represents the total rate of admissible flows. Estimation of  $B_{eff}$  can be achieved using theoretical analysis or simulation model of the wireless LAN. In subsection ??, a theoretical approach is developed. Estimation of the effective channel capacity is an interesting, challenging, problem, but one that is orthogonal to AROMA. Any QoS reservation system would face this problem and AROMA can utilize any mechanism developed for this purpose.

An AP also maintains a reservation list containing information about all active node reservations. Each element in the list is  $\langle client - id, reserved - capacity \rangle$  which identifies a reservation uniquely. Upon node activation or reset, the reservation list is empty. A small amount of

minimum capacity is reserved for all best-effort flows, in order to prevent their starvation.

When a reservation request arrives in the DATA frame following the R-RTS, the server extracts the requested capacity rate. If this plus the used capacity rate plus minimum best-effort capacity rate exceeds the total capacity  $C$ , the R-RTS is silently discarded. The client times out, notes that it does not have a reservation and may continue to use the channel as a best-effort node, or attempt a reservation at a later time.

If the request can be accommodated, a new entry for this reservation is added to the reservation list and an ACK is sent to the client. This tells the client that the request is accepted.

### C. Packet Admission

When the client needs to transmit a data packet, it first sends an RTS to the server. If this client has a valid reservation, the server associates the corresponding packet with the client. If it has not, the server considers it as best-effort traffic, which does not have a prior reservation. The server decides whether or not to send a CTS in response to the RTS based on the leaky bucket admission controller, which works as follows.

Recall from section III-A that a reservation request has three parameters: *token size* ( $TS$ ), *token rate* ( $TR$ ), and *burst size* ( $BU$ ). Each client  $i$  is allocated a token bucket of capacity  $BU_i$  that is replenished at a rate  $TR_i$ , with tokens of size  $TS_i$ . That is, every  $t$  seconds, the controller adds  $t * TR_i$  tokens to the bucket to produce an effective rate of  $R = TR_i * TS_i$  bits/sec. The bucket overflows if the number of tokens exceeds  $BU_i$ .

The server honors an RTS if and only if the sum of the token sizes in the bucket adds up to at least the packet's size (indicated in the RTS). The controller rejects a packet if it does not have sufficient tokens for transmission. That is, no CTS is sent in response to an RTS if the admission function fails. The client is forced into an EIFS backoff. If the sum of the token sizes is less than the packet size, then the admission controller treats it as if it is from a best-effort client. Hence, the packet still might be accepted from the best-effort reservation. On a packet admission, the controller removes tokens corresponding to the packet size from the token bucket.

Over the long term, the rate at which packets are accepted from the admission controller is limited by the rate at which tokens are added to the bucket, which in turn is the reserved rate.

### D. IEEE 802.11 Compatibility

This section presents the changes in IEEE 802.11 frame format for using AROMA. All changes are backward

compatible with IEEE 802.11. That is, upgrading the server (AP) and introducing AROMA-enhanced clients do not affect legacy clients in any way. In particular, all packet format modifications and algorithmic modifications are made in a manner transparent to the legacy clients. Thus, AROMA-enhanced nodes will co-exist with legacy nodes. Further, this allows us to “incrementally” deploy AROMA in an enterprise using WLANs.

AROMA introduces only one new frame, namely the R-RTS. The frame formats for the RTS and R-RTS are as shown in Figure 2. Consider the fields that are marked 0 in the figure. These fields are intended for DATA frames, and are ignored for RTS, CTS and ACK frames. One of these bits is used, namely the Order bit, to indicate an R-RTS. That is, if the bit is set to 1, the frame is treated as an R-RTS by an AROMA-capable node. A server checks if this field is set to 1 upon reception of RTS. A node which does not have AROMA capability sees the R-RTS as a regular RTS since it does not have to deal with the order bit. Such a node then defers on NAV. This is the only change in the frame format.

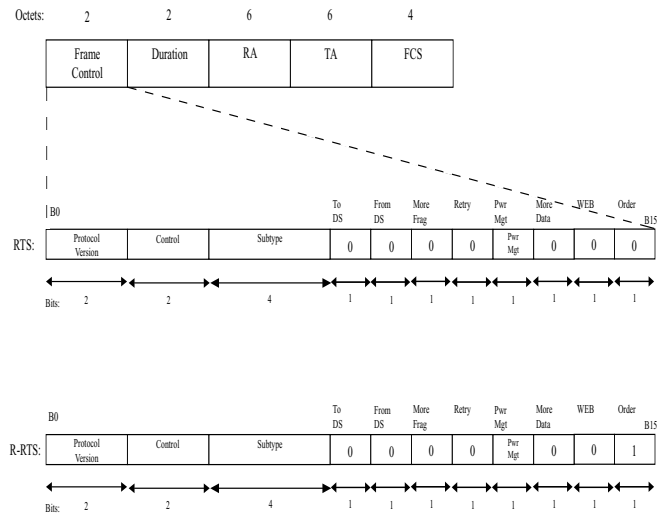


Fig. 2. RTS and R-RTS frame formats

The second modification is that the server has reservation and admission functionalities and processes the reservation request data packet without sending it to the upper layer. In a WLAN the AP is the only entity having these functionalities. The other nodes that need the reservation should be able to send a R-RTS and the reservation request to the AP.

To see that the changes are backward compatible, consider what a non-AROMA capable 802.11 client  $N$  in a WLAN with a AROMA-capable AP and one or more

AROMA clients perceives. An R-RTS from an AROMA-enhanced client  $C$  to server  $S$  is treated as an RTS and the NAV deferral is employed as it should be. The ensuing reservation request appears as just another DATA packet exchange between two nodes and is ignored. When the server runs out of capacity, client  $N$  does not receive a CTS in response to its RTS. The client then goes into backoff mode and tries it again, and so on. When the capacity becomes available again, the client can get a response to its CTS.

#### IV. Theoretical Analysis

Previous discussion describes how AROMA is integrated and fully compatible with IEEE 802.11 DCF. This section presents analysis of the probabilistic functioning of the packet admission controller based on a service differentiation model; the objective is to service packets associated with reserved sessions at a minimum of the agreed service level, and to service best effort requests and excess demand from reserved sessions without exceeding the network's saturation capacity. This goal is achieved by computing a service class-based rejection probabilities based on the saturation capacity assignments to each class.

Based on DCF operation the channel must be monitored in the idle state for DIFS (Distributed Inter-Frame Space) before an eligible node may contend for channel access. If the node was previously busy it must begin in the first backoff stage - delaying its initial contention a random number of time slots ( $\sigma$ ); as long as the channel remains idle the backoff counter is decremented each  $\sigma$ . If, however, activity is sensed on the channel to backoff process must be deferred - resuming only after the channel is once again idle for DIFS. Transmission of RTS or R-RTS proceeds when the backoff counter reaches zero. Permission is granted by the AP to transmit a packet if CTS is received prior to the expiration of a timer equivalent to the EIFS (Extended Inter-Frame Space) time.

In AROMA a station will exponentially increase its backoff window size, thus, failing in its bid to acquire the channel in the event of (1) a collision, or (2) rejection of the RTS by the AP resulting in failure to receive CTS. The rejection probability,  $P_{rej}$ , is a fundamental aspect of AROMA's control mechanism. The objection is to modulate traffic of each service class using  $P_{rej}$  and bounding total traffic of the saturation level to avoid congestion and maximize throughput.

The following assumptions are adopted in the ensuing analysis. The network consists of  $n$  stations divided among  $N$  pre-defined (by traffic descriptors) service classes. The number of stations contending in the  $i^{th}$  service class is fixed at  $n_i$ ,  $i \in [1, N]$ ,  $\sum_{i=1}^N n_i = n$ .

To determine the upper bound all stations are assumed to be continuously backlogged, ensuring constant saturation conditions. Finally, since all packets of a given class are treated equivalently at all times, the collision-rejection probability,  $p$ , of a transmitted packet is memoryless and, hence, stationary.

The models from [4] [15] [8] provide the basis for the case of AROMA with multiple service classes. The following analysis consists of four components: (1) Evaluation of the stationary transmission probability,  $\tau_i$ , used by a "tagged" station in the  $i^{th}$  service class. This determines the probability that such a station will transmit during a randomly selected time slot - it is a function of the collision-rejection probability,  $p_i$ . (2) Using the results of the first part of the analysis the class-based saturation throughput is defined and evaluated. (3) The required packet admission 'test', the rejection probability is derived based on the preceding intermediate results. (4) Finally, the analytical model is validated via simulation.

##### A. The Markov Chain State Probabilities

In this subsection the stationary probabilities for packet transmission for a given service class is determined as a non-linear function of the collision-rejection probabilities and the rejection probabilities. The Markov chain formulation first proposed in [4] and used for the multi-hop saturation capacity analysis in [8] and M/MMGI/1/K queueing analysis in [15] is extended for the multi-class case in AROMA. For a given station in service class  $i$ , let  $b(i, t)$  denote the state of the backoff time counter at the beginning of time slot  $t$  ( $t = 0, 1, \dots$ ). Note that  $b(i, t)$  is a non-Markovian Process because of its dependence on transmission history. Let  $s(i, t)$  be a stochastic process which represents the backoff stage  $[0, \dots, m]$  of a station in the  $i^{th}$  service class at time  $t$ . In this model, the 2-dimension process  $(b(i, t), s(i, t))$  forms a discrete time Markov chain whose non-zero one-step transition probabilities are given by:<sup>2</sup>:

$$\left\{ \begin{array}{ll} P\{j, k|j, k+1\} = 1 & k \in [0, W_j - 2], \\ & j \in [0, m] \\ P\{0, k|j, 0\} = (1 - p_i)/W_0 & k \in [0, W_0 - 1], \\ & j \in (0, m) \\ P\{j, k|j-1, 0\} = p_i/W_j & k \in [0, W_j - 1], \\ & j \in (1, m) \\ P\{m, k|m, 0\} = p_i/W_m & k \in [0, W_m - 1] \end{array} \right. \quad (1)$$

where  $W_j = 2^j W$ . These transition probabilities account, respectively, for: (1) the decrements of the backoff timer; (2) the backoff timer of the new packet starts from the

<sup>2</sup> $P\{j_1, k_1|j_0, k_0\} = P\{s(i, t+1) = j_1, b(i, t+1) = k_1, |s(i, t) = j_0, b(i, t) = k_0\}$

backoff stage 0 after a successful transmission; (3) the unsuccessful transmission causing the backoff stage to increase; (4) the resetting of the contention window after unsuccessful transmission or restart the backoff process for a new packet when the transmission is successful.

Let  $b_{i,j,k} = \lim_{t \rightarrow \infty} P\{s(i,t) = j, b(i,t) = k\}$ ,  $i \in [1, N]$ ,  $j \in [0, m]$ ,  $k \in [0, W_j - 1]$  be the stationary distribution of the chain. The following holds:

$$b_{i,j,0} = p_i b_{i,j-1,0}, \quad 0 < i \leq m \quad (2)$$

The chain is regular, for each  $k \in [0, W_j - 1]$ , hence:

$$b_{i,j,k} = \frac{W_j - k}{W_j} \begin{cases} (1 - p_i) \sum_{k=0}^{m-1} b_{i,k,0} + b_{i,m,0} & j = 0 \\ p_i b_{i,j-1,0} & 0 < j \leq m \end{cases} \quad (3)$$

All stationary probabilities  $b_{i,j,k}$  can be expressed as functions of  $b_{i,0,0}$ . The value of  $b_{i,0,0}$  is determined using the normalization condition:

$$\begin{aligned} 1 &= \sum_{j=0}^m \sum_{k=0}^{W_j-1} b_{i,j,k} = \sum_{j=0}^m b_{i,j,0} \sum_{k=0}^{W_j-1} \frac{W_j - k}{W_j} \\ &= \sum_{j=0}^m b_{i,j,0} \frac{W_j + 1}{2} \end{aligned} \quad (4)$$

from (3) and (4),  $b_{i,0,0}$  is found:

$$b_{i,0,0} = \frac{2(1 - 2p_i)(1 - p_i)}{W(1 - (2p_i)^{m+1})(1 - p_i) + (1 - 2p_i)(1 - p_i^{m+1})} \quad (5)$$

Let  $\tau_i$  denote the probability that a station subscribing to the  $i^{th}$  service class transmits a packet in a slot time. Since transmission occurs when the backoff counter reaches zero, independent of the backoff stage, it is given as:

$$\tau_i = \sum_{j=0}^m b_{i,j,0} = \frac{1 - p_i^{m+1}}{1 - p_i} b_{i,0,0} \quad (6)$$

The probabilities  $\tau_i$  are functions of  $p_i$  which is yet to be determined. In steady state, each remaining station in the  $i^{th}$  service class transmits a packet with probability  $\tau_i$ , therefore a failure occurs when at least one other station, from any service class, transmits at the same time or a rejection occurs due to the admission controller rejects the request. Thus,  $p_i$  can be expressed as the probability that a transmitted packet encounters a collision or is rejected by the AP:

$$p_i = 1 - \left[ (1 - \tau_i)^{(n_i-1)} \prod_{j=1, j \neq i}^N (1 - \tau_j)^{n_j} \right] \cdot (1 - P_{rej}(i)) \quad (7)$$

The set of equations (6) and (7) ( $i = 1, \dots, N$ ) represent a nonlinear system of equations with  $2N$  unknowns  $\tau_i$  and

$p_i$  assuming that the rejection probability  $P_{rej}$ , is known. It can be shown that they have a unique solution and can be solved by using numerical techniques [4].

## B. Saturation Throughput by Service Class

Saturation throughput occurs when all relevant stations always have a packet to send. It represents to point before network congestion is reached, when peak capacity is attainable. Let  $S_i$  represent the saturation throughput of a tagged station in the  $i^{th}$  service class. It can be defined as the ratio of payload in a randomly selected slot time to the mean duration of the slot time.

Let  $P_{tr}$  be the probability that there is at least one transmission in the considered slot time. Since  $n$  stations contend for the channel, and each transmits with probability  $\tau_i$  in the  $i^{th}$  service class.  $P_{tr}$  is given by:

$$P_{tr} = 1 - \prod_{j=1}^N (1 - \tau_j)^{n_j}$$

Let  $P_s(i)$  be the probability that an attempted transmission is successful. It is given by the probability that a station in the  $i^{th}$  service class is transmitting and the remaining  $n - 1$  stations remain silent, conditioned on the fact that at least one station transmits:

$$P_s(i) = \frac{[n_i \tau_i (1 - \tau_i)^{n_i-1} \prod_{j=1, j \neq i}^N (1 - \tau_j)^{n_j}] (1 - P_{rej}(i))}{P_{tr}} \quad (8)$$

and

$$P_s = \sum_{i=1}^N P_s(i)$$

Let  $E[P]$  be the mean packet length. The average amount of payload information successfully transmitted in a slot time can be expressed as  $P_{tr} P_s E[P]$ . The average length of a slot time can be determined by considering that, with probability  $1 - P_{tr}$  the channel is empty; with probability  $P_{tr} P_s$  it having a successful transmission, and with probability  $P_{tr}(1 - P_s)$  it is having a collision or a rejection. Hence,  $S_i$  can be expressed as:

$$S_i = \frac{P_s(i) P_{tr} E[P]}{(1 - P_{tr} \sigma) + P_{tr} P_s T_s + P_{tr} (1 - P_s) T_c} \quad (9)$$

where,  $T_s$  is the average time that the medium is sensed busy due a successful transmission and  $T_c$  is the average time that the medium is sensed busy by each station when a collision or a rejection occurs and  $\sigma$  is the duration of an empty slot. The values of  $T_s$  and  $T_c$  depend on the channel access mechanism of IEEE 802.11. Assuming that all stations use the same channel access mechanism,  $T_s$

and  $T_c$  are defined as follows, assuming the RTS/CTS access mechanism is employed:

$$\begin{aligned} T_s &= DIFS + RTS + SIFS + CTS + SIFS + \\ &\quad H + E[P] + SIFS + ACK + \sigma \\ T_c &= DIFS + RTS + \sigma \end{aligned}$$

Where  $H = MAC_{hdr} + PHY_{hdr}$ .

### C. Probability of Admission Rejection

Based on the AROMA architecture, the server process in the AP determines if a packet transmission request will be accepted with an RTS. The decision is based on the reservation (if any) made by the client (MAC node), the load presented by the client (policed via the leaky bucket), and the current residual capacity. If demand exceeds the reservation and excess capacity is not available, the request will be rejected. Adaptive mechanisms at a higher layer could be used to enable a client to request more bandwidth.

For the present analysis, it is assumed, without loss of generality that there are two service classes: Quality of Service (QoS) and Best-Effort (BE). Reservation of resources only applies to QoS class nodes. Nodes in BE service class compete for residual capacity,  $B_{res}$ , without exceeding saturation.  $B_{res}$  is determined by the difference between the effective capacity  $B_{eff}$  and the reserved capacity  $B_{rsv}$  :

$$B_{res} = B_{eff} - B_{rsv} \quad (10)$$

Initially  $P_{rej}$  is initialized to 0 for each class. It is easy to find  $\tau$  and  $S$  for QoS and BE stations using the previous analysis. Assuming ergodic conditions how much capacity each station is demanding  $B_{dmd}$  and  $P_{rej}$  are found as follows:

$$B_{dmd}(QoS) = S_{QoS} \cdot R \quad (11)$$

$$B_{dmd}(BE) = S_{BE} \cdot R \quad (12)$$

$$P_{rej}(QoS) = 1 - \frac{B_{rsv}}{B_{dmd}(QoS)} \quad (13)$$

$$P_{rej}(BE) = 1 - \frac{B_{res}}{B_{dmd}(BE)} \quad (14)$$

where  $R$  is the channel rate. Using these new values for  $P_{rej}$  and the Markov Chain model the new  $\tau$  values are calculated. How much capacity the station is demanding given the new  $\tau$  is to be found as follows: the successful transmission probability is first decided considering no rejection probability.

$$P_s(i|P_{rej}(i) = 0) = \frac{n_i \tau_i (1 - \tau_i)^{n_i - 1} \prod_{j=1, j \neq i}^N (1 - \tau_j)^{n_j}}{P_{tr}} \quad (15)$$

From (9) and (15) the new  $S_i$  is found for each QoS and BE service class which reflects the throughput demand and using (11) and (12) the capacity demand is calculated. From (13) and (14) the  $P_{rej}$  values are found. These steps are iterated until the difference between a new and previous value for  $P_{rej}$  is small. Finally, the saturation throughput values are calculated from (9) using the final values of  $P_{rej}$ .

The following section presents validation of the analytical results and simulation based performance analysis. Conclusions and discussion of future work involving AROMA, including support for multiple-hop configurations and incorporation of related analytical work [15] into the admission control algorithm to achieve statistically bounded QoS without peek allocation appear in the final section.

## V. Simulation Model and Analysis

This section presents results from discrete event simulation modeling that achieves the following goals: (1) validation of the analytical results presented in Section-IV emphasizing the impact of  $B_{eff}$  on throughput and demonstrating how the analytical model can be utilized to obtain the optimal value for  $B_{eff}$ ; and (2) performance analysis of AROMA in the access-point (AP) configuration using one QoS traffic class for VoIP traffic and a "default" class for best-effort (BE) traffic. The performance analysis focuses on delay and packet-loss metrics while varying the ratio of QoS to BE stations given a fixed number of nodes. The experiments include comparison of AROMA to the most significant proposed alternatives.

### A. Model Validation

The discrete-event simulation engine OPNET provided the tools to build an effective model of AROMA for analytical validation and performance analysis. All simulation results reflect statistically significant analysis based on a 95% confidence level and relative precision of 0.05. Sensitivity of throughput for QoS, BE and combined traffic classes was conducted as the value of  $B_{eff}$  was varied. Results show no statistical difference between the analytical results of the previous section and the simulation for the network scenario described above and the system parameter values given in table-I.

Investigation of the impact of effective channel capacity on the various throughput values was chosen for experimental validation of the analytical model. Results are shown in Figure-3, wherein the number of QoS nodes and BE nodes remained fixed at 10 each, with the AROMA packet size fixed at 200 bytes. The figure includes vertical bars that delineate the desired 95% confidence interval

802.11 Parameters	Values
PYH Layer Specification	DSSS
Channel Transmission Rate	11Mbps/sec
Packet Length	1600bits
$CW_{min}$	32
$CW_{max}$	1024
m	4
$\sigma$	20 $\mu$ s
H	6Bytes
DIFS	50 $\mu$ s
SIFS	10 $\mu$ s
RTS	44Bytes
CTS	38Bytes
ACK	38Bytes

TABLE I

IEEE 802.11 SYSTEM PARAMETER VALUES

values obtained from the simulation. The independent variable is AROMA Server's value for  $B_{eff}$  and the dependent variable indicates the achieved throughput: The curves on the top represent the QoS traffic, the curves on the bottom reflect BE traffic and the central curve is the aggregate throughput.

Two upward jumps in QoS throughput are balanced by equal downward jumps in BE throughput. Prior to the initial jump at  $B_{eff} = 810$  Kbps offered by the AROMA Server there were 8 QoS nodes admitted and active. The increase threshold for BE at 810 Kbps enabled the admission of one additional QoS session. This explains the initial jump in QoS throughput which leave less capacity available for the BE traffic. Increasing  $B_{eff}$  at levels insufficient for the admission of additional QoS traffic shows a predicted linear drop in QoS throughput with accompanied by an equivalently increasing throughput for BE traffic. So long as the QoS throughput is sufficient for the application the optimal value for  $B_{eff}$  occurs at the corresponding peak value for BE throughput. The final jump is at 880 Kbps, the point at which the tenth and maximum number of QoS sessions is admitted.

## B. Performance Analysis

Three main objectives were identified for the performance analysis of AROMA. The first was to perform sensitivity analysis of the QoS metrics with respect to the number of QoS nodes under the condition that the remaining nodes produced a constant stream of background best effort traffic. The question to be answered is how many QoS sessions can be admitted while retaining their required service level? The second objective was compare the performance of AROMA under the above conditions to the service available by the existing DCF function of IEEE 802.11 and the QoS extensions provided by the EDCF function of IEEE 802.11. Finally, in a mixed environment where there may be large numbers of best-effort and/or VoIP stations in competition, hence, it becomes critical

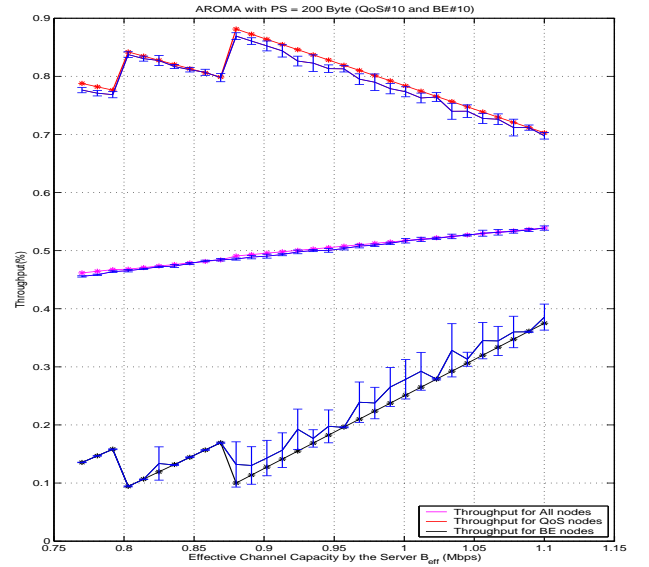


Fig. 3. Throughput vs.  $B_{eff}$ : Analytical and simulation results.

to understand the impact of simultaneously varying the number of admitted QoS sessions and best-effort nodes. The simulation model was designed to isolate the effects directly attributable to the MAC layer. Hence, enabling the direct investigation of AROMA's ability to deliver QoS and resolve channel contention.

## C. Simulation Parameters

System parameters were chosen to reflect typical installations of IEEE 802.11b; they are the same parameters used for model validation as listed in table-I. The traffic parameters were selected to model the behavior of the G.711 codec using 20ms packetization intervals for VoIP. Peak allocation was used by the server since the achievable throughput limits can be estimated accurately by the analytical model. All traffic sessions assumed a wired server, hence, traffic was not generated between wireless nodes. Based on the G.711 specification the raw packet lengths for voice were fixed at 160 bytes. However, Table II indicates the overhead required by the underlying protocol layers: RTP, UTP, IP and MAC; the aggregate frame lengths came to 258 bytes (significant protocol overhead). Each VoIP session was held for 3 minutes, consisting of unidirectional traffic from the wireless client. The background load was varied by changing the number of best-effort clients. Data traffic was generated by wireless nodes sending fixed 512 byte frames at a mean interarrival rate of 20 ms. The aggregate data load was set at 0.0, 0.8 and 1.6 Mbps for each fixed number of VoIP nodes. The number of voice stations was varied from 2 to 18.

VoIP Traffic Parameters	Values
Packet Interarrival time	20ms
Voice packet length	160 bytes
RTP layer overhead	12 bytes
UDP layer overhead	8 bytes
IP layer overhead	20 bytes
MAC layer overhead	34 bytes
PHY layer overhead	24 bytes

TABLE II  
VOICE TRAFFIC PARAMETERS

As the table shows the simulation models the 11 Mbps version of IEEE 802.11; this was also implicit in the analytical model. Experimental results and theoretical calculations show that the maximum achievable throughput is approximately 6.2 Mbps for data traffic. The maximum throughput when using only VoIP traffic can be as low as 960 Kbps using a typical VoIP codec (G711 with 20ms audio data per 200 bytes packet). Consequently, the maximum number of simultaneous VoIP calls that can be placed in the network cannot exceed 12.

AROMA requires specification of two additional parameters. The values for these parameters are determined by the analytical model. The first is the value for maximum effective channel capacity  $B_{eff}$ . The second is the maximum number of allowable VoIP admissions, i.e., the maximum capacity that can be reserved for voice stations. The upper bound for  $B_{eff}$  is 960 Kbps for the VoIP applications. However, considering a bounded reservation capacity for best-effort traffic and the performance metrics the optimal value for  $B_{eff}$  can be found analytically. Figure-3 uses the analytical model to plot packet loss percentage versus the number of admitted VoIP calls and the number of best-effort stations. Throughput given 10 VoIP stations is at maximum when  $B_{eff}$  is 880Kbps. Thus, the AROMA parameter for the minimum bound on best-effort capacity and  $B_{eff}$  are determined. The values shown depend on the precise network conditions and system parameters. Development of an adaptive mechanism for dynamically adjusting these parameters remains as future work. Finally, it was found experimentally that maintaining the best-effort bound at 9% of  $B_{eff}$  in all simulation and analysis delivers the best results for the traffic models used here.

Figure-4 shows the number of admitted voice stations varied from 10 to 40 and the number of BE nodes from 1 to 20, thus  $B_{eff}$  is adjusted considering the 9% minimum best-effort capacity. Admitting 11 voice stations is feasible if no best-effort stations are active in the system. This quality of this session would be excessively sensitive to best-effort traffic. Hence, a maximum of 10 voice stations is enforced in the AROMA simulations to ensure minimal acceptable voice performance. Simulation results provide

validation for this conclusion.

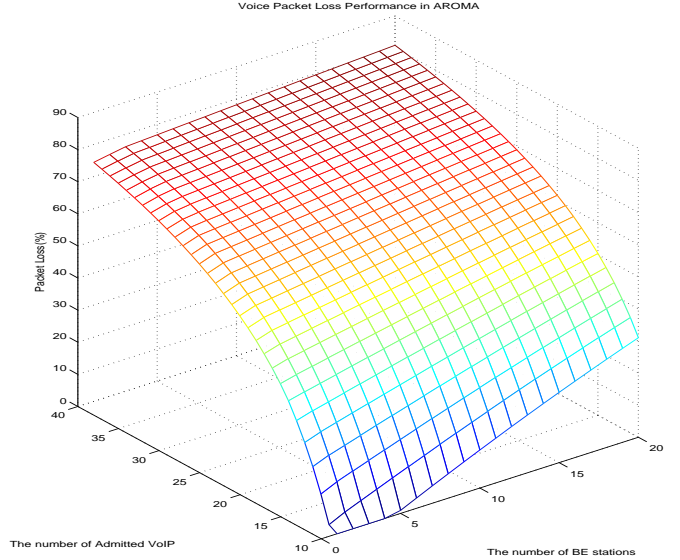


Fig. 4. Analytical results for voice packet loss

#### D. Performance Metrics

The two most important metrics reflecting the delivered QoS for VoIP are packet-loss and end-to-end delay. Packet loss may result from any of the following: (1) buffer overrun, (2) excessive MAC-layer collisions or (3) access denial (rejection) by the AROMA Server. Delay may be incurred in numerous ways; here we consider queuing delay at the source station and MAC delay incurred by the frame in service at the source station. Studies have shown that for acceptable voice quality VoIP can tolerate 200 ms delays with as much as 2% packet loss. The jitter in the delay caused due to steady variance in traffic patterns and transient situations can be remedied through dynamically adapting the playout buffer size [15]. Note that packet loss is represented as a percentage according to the following equation:

$$PL = 100 \cdot \left(1 - \frac{Pkts_{rcvd}}{Pkts_{sent}}\right) \quad (16)$$

where the fraction shows the ratio of received packets to the total number of packets generated by the source.

#### E. Simulation Results

Figure-5 depicts the mean packet loss and end-to-end delay for AROMA, IEEE 802.11 DCF and IEEE 802.11e EDCF with no best-effort traffic. As expected, packet loss and delay increase with the number of voice stations. There are two curves for AROMA case in the figures. The circled curve represents only admitted voice stations, which is limited to 10 even though active voice station number is increasing. The crossed curve shows the result

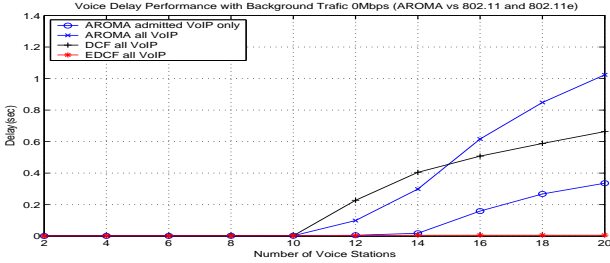
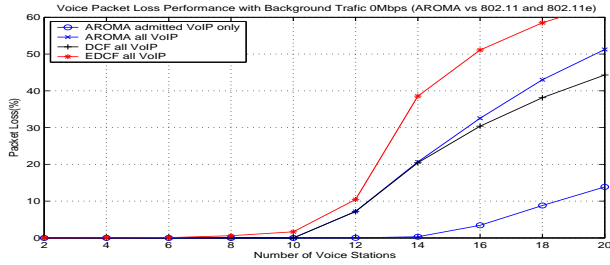


Fig. 5. Voice packet Loss and Delay Performance when the Background Traffic is 0Mbps

for all voice stations. The figures clearly demonstrate that AROMA outperforms DCF and EDCF by supporting 10 VoIPs even if there are 15 active voice stations. This also validates the analytical results.

Results for delay using EDCF appear to significantly outperform DCF and AROMA. This result is due largely to effects of system parameter settings required in EDCF: the minimum contention window size ( $CW_{min}$ ) is set to  $\frac{W}{4}$  and the maximum contention window size ( $CW_{max}$ ) to  $\frac{W}{2}$ , whereas  $CW_{min}$  and  $CW_{max}$  are set to  $W$  and  $2^m W$ , respectively in the AROMA and DCF simulations. Moreover, EDCF performs so poorly with respect to packet-loss that its improved delay performance has no application level relevance. When the wireless network is used only for VoIP traffic, after maximum effective 10 VoIPs, additional VoIP-streams lead the network to congestion collapse. Thus, all voice traffic cannot be sent through the network under DCF and EDCF. However, AROMA still provides 10 acceptable quality VoIPs given as many as 15 active voice stations.

Figure-6 shows the simulation results given 0.4Mbps of background traffic, which represents a light traffic scenario. AROMA and DCF support 10 VoIPs up to 13 active stations and 10 VoIPs up to 10 active active voice stations respectively. EDCF performs poorly supporting a maximum of 8 VoIPs.

Results given heavy background traffic of 1.6Mbps are shown in Figure-7. AROMA supports 9 VoIPs when compared with 8 and 7 for DCF and EDCF, respectively. When

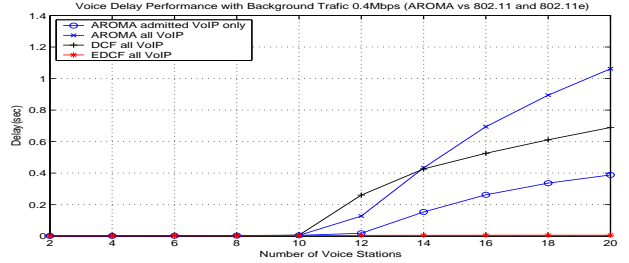
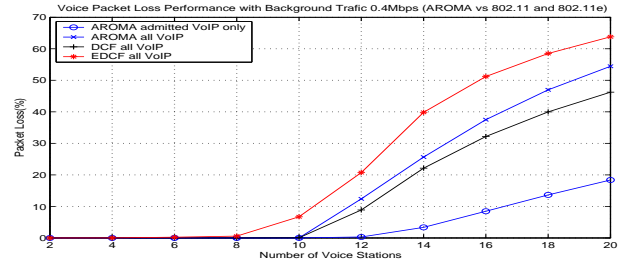


Fig. 6. Voice packet Loss and Delay Performance when the Background Traffic is 0.4Mbps

the network is shared between voice and data, additional voice clients increase network access contention causing DCF and EDCF performance to become much more sensitive to the additional VoIP-streams than AROMA.

## VI. Conclusions

Existing solutions to the problem of QoS provision in Wireless LANs are either priority-based within asynchronous (contention-based) protocols or reservation-based in synchronous (contention-free) protocols. We have presented the first protocol – AROMA – that provides reservations within an asynchronous context. In particular, AROMA enhances CSMA/CA in general and IEEE 802.11 Distributed Coordination Function (DCF) in particular to enable traffic description, reservation requests, and admission control. AROMA is backward-compatible with IEEE 802.11 DCF and therefore can co-exist with non-AROMA-enhanced clients. This allows for incremental deployment in an existing Wireless LAN. AROMA uses well-known traffic descriptors from network layer QoS concepts, such as leaky-bucket and moving window descriptors which facilitates cross-layer integration for end-to-end QoS solutions.

Theoretical and experimental analysis using simulation of AROMA were both presented in this paper. The theoretical analysis is well validated with the simulation results and can be used to find optimal values for AROMA parameters. The experimental analysis is based on enhancing the IEEE 802.11 model in OPNET to include AROMA;

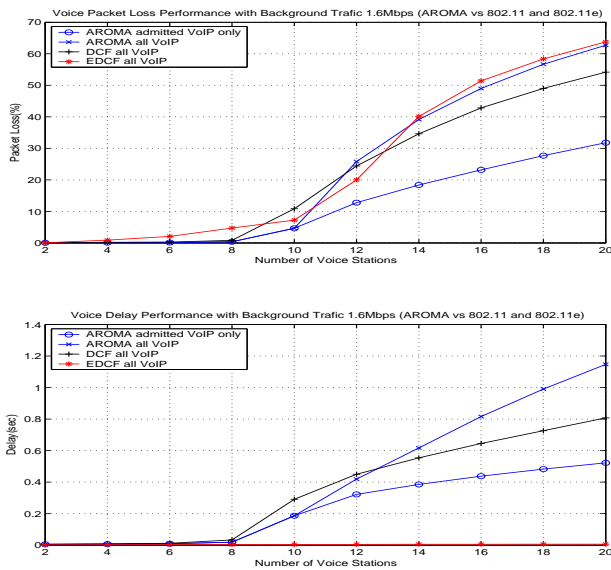


Fig. 7. Voice packet Loss and Delay Performance when the Background Traffic is 1.6Mbps

the model is of very high fidelity. AROMA's performance has been compared to both traditional 802.11 DCF and the emerging 802.11e priority-based EDCF standard using a realistic voice-over-IP (VoIP) scenario. Results show that compared to IEEE 802.11 EDCF AROMA can support 50% more active stations when there is no background traffic and 30% more active stations when there is background traffic.

### A. Future Work

At the time of this research 802.11e HCCA was not available in the OPNET simulator modeler. When it becomes available further analysis will be conducted to compare HCCA with AROMA. Additionally, future analysis will include more varied traffic scenarios and relax the peak allocation assumption. Variable Bit Rate (VBR) VoIP traffic models such as Brady's model [6] are of particular interest and importance.

As seen from figures, AROMA scales better with increasing numbers of BE users compared to IEEE 802.11 DCF and IEEE 802.11e EDCF. However, since all the nodes use the same CW and back-off appropriately in AROMA, it does not scale well enough when the number of BE nodes is very high. The design philosophy of AROMA was to keep it as simple as possible. Implementing AROMA over IEEE 802.11e EDCF is expected to provide far better scalability. Moreover, we plan to investigate adaptive CW and back-off algorithms in order to reduce the channel access times at high load for only BE nodes to further improve scalability.

A promising future research avenue is the use of AROMA in mobile ad hoc wireless networks. As an extension of the 802.11 DCF, which is the MAC layer used in a number of ad hoc network prototypes and simulation models, AROMA is an ideal choice for reservation-based QoS in ad hoc networks.

### References

- [1] J. L. Sobrinho and A. S. Krishnakumar, *Real-time traffic over the IEEE 802.11 medium access control layer*, Bell Labs Technical Journal (1996).
- [2] P. Almquist, *Type of service in the internet protocol suite*, RFC 1349, July 1992.
- [3] M. Barry and A. T. Campbell, *Distributed control algorithms for service differentiation in wireless packet networks*, IEEE INFOCOM Proceedings, vol. 1, 2001.
- [4] G. Bianchi, *Performance analysis of the IEEE 802.11 distributed coordination function*, IEEE Journal on Selected Areas in Communications **vol. 18** (2000), no. 3, pp. 535–547.
- [5] IEEE Standards Board, *IEEE standard 802.11 - wireless LAN medium access control (MAC) and physical layer (PHY) specifications*, The Institute of Electrical and Electronics Engineers, Inc, 345 East 47th Street, New York, NY 10017-2394, USA, June 1997.
- [6] P. Brady, *A model for generating on-off speech patterns in two-way conversation*, Bell Syst. Tech. Journal **vol. 48** (1969), no. 7, pp. 2245–2272.
- [7] D. J. Deng and R. S. Chang, *A priority scheme for IEEE 802.11 DCF access method*, IEICE Transactions on Communications (1999).
- [8] Y. Fang and A. B. McDonald, *Theoretical channel capacity in multihop ad hoc networks*, IEEE LANMAN'04 (San Francisco), April 2004.
- [9] C.L. Fullmer and J.J. Garcia-Luna-Aceves, *Floor acquisition multiple access (fama) for packet-radio networks*, Proceedings of the Conference on Applications, Technologies, Architectures and Protocols for Computer Communication (SIGCOMM), 1995, pp. 262–273.
- [10] S. Garg and M. Kappes, *Admission control for VoIP traffic in IEEE 802.11 networks*, IEEE GLOBECOM, 2001.
- [11] P. Karn, *MACA - a new channel access protocol for packet radio*, Proceedings of ARRL/CRRL Amateur Radio Ninth Computer Networking Conference, 1990, pp. 234–140.
- [12] S. Keshav, *An engineering approach to computer networking*, 7th ed., Addison Wesley, 2000.
- [13] K. Kim and S. Shin, *A novel MAC scheme for prioritized services in IEEE 802.11a wireless LAN*, ATM (ICATM 2001) and High Speed Intelligent Symposium, 2001. Joint 4th IEEE International Conference on, 2001.
- [14] S. Herzog, L. Zhang, S. Berson and S. Jamin, *Resource reservation protocol (RSVP)*, IETF RFC 2205, September 1997.
- [15] M. Ozdemir and A. B. McDonald, *An M/M/GI/1/K queueing model for IEEE 802.11 ad hoc networks*, IEEE/ACM PE-WASUN'04 (Venice, Italy), October 2004.
- [16] P. May, O. Klein, G. Hiertz, S. Mangold, S. Choi and L. Stibor, *IEEE 802.11e wireless LAN for quality of service*, In Proceedings of the European Wireless (Florence, Italy), vol. 1, February 2002, pp. 32–39.
- [17] S. Shenker, V. Bharghavan, A. J. Demers and L. Zhang, *Macaw: A media access protocol for wireless LAN's*, SIGCOMM, 1994, pp. 212–225.
- [18] D. Saha, W. Feng, K. Kandlur and K. Shin, *Adaptive packet marking for providing differential services on the internet*, In Intl. Conf. On Network Protocols, Oct 1998.