Model-based Resource Prediction for Multi-hop Wireless Networks

Yuan Sun[†]   Xia Gao[‡]   Elizabeth M. Belding-Royer[†]   James Kempf[‡]

[†] Department of Computer Science   [‡]DoCoMo Communications Laboratories USA

University of California, Santa Barbara   {gao, kempf}@docomolabs-usa.com

{suny, ebelding}@cs.ucsb.edu

Abstract

Ad hoc networks have been proposed for a variety of applications where support for real time, multimedia services may be necessary. This requires that the network is able to offer quality of service (QoS) appropriate for the latency and throughput bounds needed to meet the real time constraint. An important component for QoS provisioning is resource estimation and quality prediction. This paper describes a model-based resource prediction (MBRP) mechanism to support real time communication in multi-hop wireless networks. Specifically, we develop an analytical model for differentiated MAC scheduling protocols. The model can predict per-flow and system-wide throughput and delivery latency, thereby enabling admission control of the flows and providing an efficient network management utility. After describing the basic model, we propose Enhanced MBRP (EMBRP) for realistic network environments. Our proposed quality prediction method is beneficial in the deployment of a real ad hoc network where knowledge of resource allocation and consumption is needed to meet the service requirements. Analytical and simulation results show that EMBRP provides accurate flow quality prediction. The results also demonstrate the effectiveness of EMBRP as an admission control solution in multi-hop ad hoc networks.

1   Introduction

Wireless networking and multimedia content are two rapidly emerging technological trends. Among types of wireless networks, multi-hop ad hoc networks provide a flexible means of communication when there is little or no infrastructure, or the existing infrastructure is inconvenient or expensive to use. They are also useful in conjunction with infrastructured wireless networks to extend the coverage area of access points. With the development of ad hoc networks, we can anticipate that multimedia applications will be popular in scenarios where these networks are used.

One challenge of providing multimedia services in wireless networks is that certain quality of service (QoS) metrics should be satisfied. There has been significant research on QoS provisioning in wired networks. For instance, Intserv [21] and Diffserv [10, 17] are two well-known approaches. In ad hoc networks, however, several unique characteristics make QoS provisioning more challenging. These characteristics include the shared wireless medium, mobility, and the distributed multi-hop communication.

Most QoS solutions for wired networks rely on the availability of precise resource utilization information for wired links. However, in ad hoc networks, all traffic within a mobile node's transmission range contends for media access; the shared na-

ture of wireless communication channels hence makes resource estimation more challenging. Multi-hop interference introduces further challenges to the problem, making it difficult to accurately determine the available resources. Without sufficiently accurate estimation of channel utilization and prediction of flow quality, i.e., throughput or transmission delay, it is difficult to provide multimedia services with satisfactory quality.

Service quality prediction is therefore an important building block for providing QoS in multi-hop wireless networks. It also enables effective admission control. The latter is important for ad hoc networks because these networks generally have limited resources, in terms of both device capabilities and available network bandwidth. If a flow has rigid QoS requirements, a prediction of the achievable quality will prevent the waste of resources at both the source node and in the whole network if the network cannot support the flow.

In this paper, we propose a model-based resource prediction (MBRP) scheme to provide flow quality prediction for ad hoc networks. Our targeted network environment is multi-hop wireless networks where support of multimedia services is desired. To help meet the real time constraints, priority scheduling mechanisms at the MAC layer can be utilized in this environment [1, 13, 22]. For instance, a Voice over IP (VoIP) traffic session has stringent real time constraints and therefore could be labeled high priority while other delay-tolerant traffic can be given lower priority. Under this context, our model supports various differentiated MAC schemes with multiple priorities and provides estimation of both per-flow and aggregated network-wide throughput and delay analysis. We further apply the basic MBRP analysis to a realistic network environment, i.e., unsaturated nodes and hidden terminal interference exist, and propose Enhanced MBRP (EMBRP) to improve the estimation accuracy.

Model-based prediction has two important benefits. First, it improves channel efficiency by avoiding the waste of the network resources due to unprovisioned traffic. By using the MBRP analysis, a flow can check whether the network can support the real time requirements before it starts. Consequently, serviced flows will meet the desired quality, thereby improving the channel efficiency. Secondly, MBRP enables flexible admission control with a wide range of quality policies. This is especially important when service differentiation is supported in the network. The network quality policy can then be, for instance, to maximize the network-wide through-

put, or to admit the maximum number of high quality flows that can be supported. However, no matter what policy is required, the admission of a new flow will affect ongoing traffic because of the shared wireless channel. Because our model predicts the impact of the new flow on both new and existing traffic, flexible admission control can be achieved.

The remainder of this paper is organized as follows. Section 2 describes related work. Section 3 presents our proposed basic MBRP analysis and its improvement, EMBRP, for a realistic network environment. We then describe how an estimation module using MBRP can be integrated with existing routing schemes for multi-hop wireless networks in section 4. The performance of our proposed approach is evaluated in section 5, and finally section 6 concludes the paper.

2 Related Work

Resource estimation has been studied extensively in wired networks [7, 18, 20]. The bandwidth or latency of a path can be estimated through end-to-end probing techniques. For instance, the *packet bunch* technique [7, 18] measures the available bandwidth between a node pair by dividing the receiver-ACKed probing packets with the time interval between the first and last received packets. Latency measurements between two nodes can be achieved through ping messages or any designated packets [20]. To deal with high network variability, multiple measurements are needed to achieve a better estimation.

In ad hoc networks, resource estimation in recently proposed QoS-aware routing protocols often takes advantage of statistical information provided by MAC layer. These solutions can be categorized into the following groups: active measurement, passive measurement, emulation-based and model-based approaches.

Active measurement methods from wireless networks inherit the basic techniques for wired networks; however, the primary difference is that they are typically conducted in a hop-by-hop fashion, due to the lack of information about the full path. Hence, they are often combined with the route acquisition process. For instance, SWAN [2] uses a *request/response* probe during route discovery to estimate bandwidth availability along a path. A ticket-based probing technique to measure link delay is proposed in [9]. Each probe accumulates the delay of the path it has traversed. One drawback of active measurements is that they are susceptible to network variability. Extra effort is required to dampen short-time variations and maintain measurement stability.

Passive measurement techniques leverage the unique characteristics of wireless networks through a collection of channel statistics at the MAC layer. For instance, Quiet Time Fraction is often suggested to predict the available bandwidth of a wireless channel by listening to the channel and measuring the fraction of time during which the channel is not in use [8, 24]. Packet forwarding latency is often measured by timestamps on RTS/CTS or DATA/ACK packets [2, 16]. Compared to active measurement, passive methods have the advantage of less control overhead. However, they often do not fully consider the contentious nature of 802.11-based MAC access.

An emulation-based delay estimation method, Virtual MAC (VMAC) [3, 23], captures most of the aspects of a real MAC and operates in parallel to the real MAC protocol. However, no real packets are actually transmitted. Instead, the VMAC algorithm estimates the probability of collision if the real packet were transmitted. VMAC emulates real MAC behavior without introducing any communication overhead. However, one drawback of VMAC is that when multiple nodes simultaneously utilize VMAC estimation, they cannot detect the collisions that will occur because the packets are not actually transmitted.

Several analytical models for IEEE 802.11 have been proposed [2, 5, 15]. Bianchi [5] uses a discrete Markov chain model to capture the behavior of CSMA/CA channel multiplexing in IEEE 802.11 and derives the saturated throughput based on a constant and independent channel collision ratio p. However, it does not specify how to calculate p in a given network topology. It also does not support priority-based differentiation schemes. A delay model for IEEE 802.11 is derived in [2] by assuming that the channel contention of competing flows are Poisson distributed. The model also takes into account the effect of different CW_{min} and CW_{max} values. However, similar to [5], it does not specify detailed algorithms to estimate channel utilization and contention window size of each flow, which is the key parameter for delay estimation. The model also does not support other differentiation schemes, such as different backoff ratio or backoff policies. A more detailed model to estimate channel utilization (based on which our model is constructed) and network throughput is provided in [15]. The advantage of this model is that it includes greater detail of IEEE 802.11. One limitation of the model is that it does not provide flow-based estimation. It also does not support priority-based MAC schemes.

To summarize, these models do not fully consider different types of differentiation schemes used in IEEE 802.11. Because of their limitations, the schemes cannot provide a flow-level QoS estimation at the level of accuracy that is needed by admission control. These models also do not address issues such as random channel corruption, non-saturated nodes, or hidden terminals within carrier-sensing range. In contrast, MBRP provides per-flow quality analysis as well as consideration of the above issues.

3 Model-based Resource Prediction (MBRP)

In this section, we propose a model-based resource prediction mechanism for multi-hop ad hoc networks. The primary objective of our mechanism is to provide accurate resource prediction for both new and existing traffic. We first describe the basic MBRP model in section 3.1. Then in section 3.2, we explain how MBRP can be applied to a realistic network environment and propose Enhanced MBRP (EMBRP).

3.1 Basic MBRP Mechanism

The basic premise of MBRP is to provide quality prediction for both ongoing traffic and new flows so that a correct flow ad-

mission decision can be made according to the quality of service policy of the network. Our model is based on the model of basic IEEE 802.11 DCF described in [2, 5, 15]. However, we extend existing work and make the following contributions:

- Development of a differentiated MAC scheme model with multiple priorities.
- Development of models for different priority-based backoff schemes.
- Estimation of both per-fbw and aggregated system-wide throughput and delay.

Hence, our model is more generic and has wider applicability than the previously proposed models. It can be used by admission control schemes for both delay and bandwidth sensitive applications. The admission decision can be modified to either admit the maximum number of supportable high-priority fbws, or to achieve the maximum network throughput.

3.1.1 Priority Scheduling Model

To provide throughput and delay estimation, we first need to model the node backoff behavior and analyze the channel utilization. Similar to previous work [2, 15], we assume that the time interval between two adjacent transmission attempts is exponentially distributed. As a result, the channel attempt rate is assumed to follow a Poisson distribution with an average rate of λ_c. We also assume that the channel collision rate, p, is constant and only relates to the current competing traffic load [2, 5, 15].

Let $A = \{a_1, a_2, ..., a_s\}$ be a set of fbws with different priorities, where s denotes the total number of priority classes supported by the system, and $\forall a_i \in A$, a_i is the number of fbws of priority class i. To maintain the clarity of the derivation, we let fbws with the same priority level have the same average packet length F_i [1] and the same average backoff window size $L(a_i)$.

The current average channel attempt rate, λ_c, which indicates the number of transmission attempts in a time slot, can then be represented by

$$\lambda_c = \sum_{i=1}^{s} \sum_{j=1}^{a_i} \frac{1}{b_{i,j}} = \sum_{i=1}^{s} \frac{a_i}{\bar{b}_i} = \sum_{i=1}^{s} \frac{a_i}{L(a_i)} \quad (1)$$

where $b_{i,j}$ denotes the backoff window size of fbw j with priority class i and $\bar{b}_i = E_j[b_{i,j}] = L(a_i)$.

In contrast to previous work [5, 15], our intention is to support multiple priority levels. To this end, the channel attempt rate, collision rate, and backoff window size of different priorities must be differentiated.

The channel attempt rate λ_c in Eq. (1) includes the effect of all the fbws in the system. For each individual fbw with priority i, the attempt rate of the competing fbws is

$$\lambda_i = \lambda_c - 1/L(a_i) \quad (2)$$

The competing fbws include all other fbws except the given fbw itself. The transmission of the fbw is successful only

<hr/>

[1] Our derivation also hold for scenarios where fbws of the same priority have different packet sizes. Instead of forming equations for each priority, we need to form equations for each fbw. However, the basic principles are the same and the equations are still solvable.

<hr/>

when all competing fbws do not transmit. Hence, the collision probability of a fbw with priority i, p_i, is

$$p_i = 1 - e^{-\lambda_i} \quad (3)$$

Before calculating the average backoff window size, we need to first decide what type of backoff scheme to use. The generic form of a priority-based backoff scheme is $CW_{next} = f_{a_i}(p_i, CW_{curr}, i)$. In the basic exponential backoff scheme of DCF, $CW_{next} = 2 * CW_{curr}$. In our previous work [22], we proposed a series of priority-based backoff schemes that differentiate traffic with various priorities. One scheme is shown below as an example. A more detailed description of the schemes and a performance comparison can be found in [22].

Let m be the maximum number of retransmissions. For the exponential backoff scheme, the probability that the jth collision occurs is

$$c_j = \begin{cases} c_0 \cdot p_i^j & 1 \leq j \leq m-1, \\ c_0 \cdot \sum_{k=m}^{\infty} p_i^k = \frac{c_0 p_i^m}{1 - p_i} & j = m \end{cases} \quad (4)$$

where $\sum_{k=0}^{m} c_k = 1$, and we also have

$$c_0 \cdot (1 + p_i + p_i^2 + \cdots + \frac{p_i^m}{1 - p_i}) = 1 \Longrightarrow c_0 = 1 - p_i \quad (5)$$

Then for any priority class i, the average backoff window size during collisions is

$$L(a_i | \text{backoff}) = b_0 c_0 + b_1 c_1 + \cdots + b_m c_m$$
$$= \sum_{j=0}^{m-1} \frac{CW_j - 1}{2} p_i^j (1 - p_i) + \frac{CW_m - 1}{2} p_i^m \quad (6)$$

Now consider the following priority-based backoff scheme $CW_j = [2^j + p_i \alpha_i] CW_{min} (\forall j \in [0, m])$ where p_i is the collision rate and α_i is a constant associated with the fbw's priority class i to differentiate the backoff ratio of different priority fbws. We have

$$L(a_i \mid \text{backoff}) = \sum_{j=0}^{m-1} \frac{(2^j + p_i \alpha_i) CW_{min} - 1}{2} p_i^j (1 - p_i)$$
$$+ \frac{(2^m + p_i \alpha_i) CW_{min} - 1}{2} p_i^m$$
$$= \sum_{j=0}^{m-1} \frac{2^j CW_{min} - 1}{2} p_i^j (1 - p_i) + \frac{2^m CW_{min} - 1}{2} p_i^m$$
$$+ [\sum_{j=0}^{m-1} \frac{\alpha_i}{2} p_i^{j+1} (1 - p_i) + \frac{\alpha_i}{2} p_i^m] \cdot CW_{min}$$
$$= \frac{CW_{min}}{2(1 - 2p_i)} [1 - p_i - p_i (2p_i)^m] + \frac{\alpha_i}{2} p_i \cdot CW_{min} - \frac{1}{2} \quad (7)$$

where CW_{min} is the minimum contention window size. By adjusting α_i, we can adjust the sensitivity of the difference of different priority classes with respect to the collision rate p_i.

Continuing the average backoff window size, $L(a_i)$, calculation, Eq. (7) gives the average backoff window size under the condition that the channel is sensed busy. IEEE 802.11 specifies that the node transmits immediately without backoff if the channel is sensed idle for the DIFS period. Let $P_{free,i}$ denote the probability of a free channel when the node attempts a

transmission and all competing flows are in the backoff stage, and let $P_{busy,i}$ denote the probability of a busy channel when at least one competing flow is transmitting. Let $\bar{F}$ be the average packet transmission time:

$$\bar{F} = \frac{\sum_{i=1}^{s} \sum_{j=1}^{a_i} F_{ij}}{\sum_{i=1}^{s} a_i} = \frac{\sum_{i=1}^{s} a_i \cdot F_i}{\sum_{i=1}^{s} a_i} \qquad (8)$$

Using the same assumption that the transmission attempt of competing flows is of Poisson distribution with rate λ_i, and the inter-arrival time between two adjacent transmission attempts is of exponential distribution with average of $1/\lambda_i$, we have

$$P_{free,i} = \frac{1/\lambda_i}{\bar{F} + 1/\lambda_i}$$
$$P_{busy,i} = \frac{\bar{F}}{\bar{F} + 1/\lambda_i}. \qquad (9)$$

Then the average backoff window size $L(a_i)$ is
$$L(a_i) = P_{free,i}(1 - e^{-\lambda_i})L(a_i \mid \text{backoff})$$
$$+ P_{busy,i}L(a_i \mid \text{backoff}) \qquad (10)$$

Hence, we have derived the expression of the current channel attempt rate λ_c as a function of average backoff window size $L(a_i)$ in Eq. (1), contending traffic rate λ_i for a node with priority i as a function of λ_c and $L(a_i)$ in Eq. (2), the expression of collision possibility p_i as a function of λ_i in Eq. (3), and finally, the expression of the average backoff window size $L(a_i)$ as a function of the collision possibility p_i in Eq. (10).

For any given flow set $a_1, a_2, ..., a_s$, we can obtain a derived $\lambda_i^\star$, $p_i^\star$ and $L(a_i)^\star$ by using a similar iteration algorithm as in [15]. The difference is that we need to calculate the variables for each priority separately. Specifically, as proved in [12], the expected number of collisions in a binary backoff algorithm grows asymptotically with $O(\log M)$, where M is the number of active stations in the network. Hence, the initial backoff window size $L(a_i)^{(0)}$, in our scheme, is bounded by the following:
$$L(a_i)^{(0)} < CW_{min} \cdot 2^{K \cdot \log(\sum_{i=1}^{s} a_i)} \qquad (11)$$
where $K(> 0)$ is some arbitrary constant. Given $L(a_i)^{(0)}$ of each priority, which represents the largest backoff window size, we can first obtain the channel attempt rate λ_c by using Eq.(1) and based on that, calculate $\lambda_i^{(0)}$ and $p_i^{(0)}$ using Eq. (2) and (3), respectively. Then, by applying Eq. (10), we can calculate $L(a_i)^{(1)}$. The iteration repeats until the difference of two consecutive iteration values satisfies $|L(a_i)^{(j+1)} - L(a_i)^{(j)}| < \epsilon$, where ϵ denotes some predefined small value. The iterative algorithm always converges as proved in Theorem 1 in [15].

3.1.2 Throughput Model
Given the current traffic rate λ_c, the competing traffic rate for each priority λ_i, the collision possibility p_i, and the average backoff window size $L(a_i)$, we now derive the throughput calculation for each priority flow.

Let B denote the capacity of the wireless link (e.g., 2Mbps). Because the packet transmission attempt is of Poisson distribution with rate λ_c, the inter-arrival time (the portion of wasted

bandwidth) between two adjacent transmission attempts is of exponential distribution with average $1/\lambda_c$. Considering that the average packet length (the portion of utilized bandwidth) is $\bar{F}$, the total utilized bandwidth, B_u, expressed as the ratio of link capacity B is:

$$B_u = B \times \frac{\bar{F}}{\bar{F} + 1/\lambda_c} \qquad (12)$$

Let B_i be the bandwidth utilized by a flow of priority i and $\bar{B}_u$ be the sum of channel usage of all flows. Then,

$$\bar{B}_u = \sum_{i=1}^{s} a_i \cdot B_i \qquad (13)$$

Due to transmission collisions among competing flows, $\bar{B}_u \geq B_u$. Note that the portion of collided bandwidth is calculated multiple times in $\bar{B}_u$. More precisely, if there are i flows transmitting at the same time, this portion of the bandwidth is added i times in the $\bar{B}_u$. Because the probability of i simultaneous transmissions is $(\lambda_c^i/i!)e^{-\lambda_c}$, under the condition that there is a transmission,

$$\bar{B}_u = \frac{\lambda_c e^{-\lambda_c}}{(1 - e^{-\lambda_c})} \cdot B_u + \frac{2 \times (\lambda_c^2/2!)e^{-\lambda_c}}{(1 - e^{-\lambda_c})} \cdot B_u + ...$$
$$= \sum_{j=1}^{\infty} j \cdot \left(\frac{\lambda_c^j}{j!}\right) \cdot \frac{e^{-\lambda_c}}{1 - e^{-\lambda_c}} \cdot B_u \qquad (14)$$

Notice that the first item in Eq. (14) is the no-collision portion of B_u, which is also the overall throughput, T, of the wireless link. T can be given by

$$T = \frac{\lambda_c e^{-\lambda_c}}{(1 - e^{-\lambda_c})} \cdot B_u \qquad (15)$$

Now let's consider per-flow throughput. For each priority flow i, the channel utilization is proportional to its average packet length $\bar{F}_i$ and inversely proportional to its average waiting time between two adjacent transmissions. The average waiting time is the sum of its own backoff window size $L(a_i)$ and the total transmission time of other competing nodes during $L(a_i)$. Hence, the ratio of channel utilization between two different priorities i and j can be presented as:

$$\frac{B_i}{B_j} = \frac{(L(a_j)\lambda_j \bar{F} + L(a_j)) \cdot F_i}{(L(a_i)\lambda_i \bar{F} + L(a_i)) \cdot F_j} \qquad (16)$$

For a total of s priority flows, $(s - 1)$ independent equations of type of Eq.(16) can be listed. Together with Eq.(13), this set of equations can be used to calculate B_i of each priority.

Finally, considering the collision loss of each priority, the effective throughput of each priority traffic is then

$$\bar{B}_i = (1 - p_i) \cdot B_i = e^{-\lambda_i} \cdot B_i \qquad (17)$$

Figure 1 shows a comparison between the analytical and simulation results for the throughput (both per-flow and aggregated) versus the number of flows in a single broadcast region. The lines are the numerical results calculated using Eq.(15) and (17), whereas the symbols are the values obtained from simulation. We can see that the simulation results closely match the analysis, thereby verifying our model.
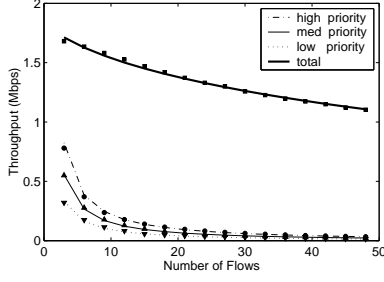
Figure 1: Throughput Model Verification. CW_{min} is set to 32, m is set to 5, and the packet size is set to 1000 bytes. The α values for different priorities are set to 1, 8 and 16.

3.1.3 Delay Model

We now derive the delay model based on the competing flow traffic rate λ_i and the collision possibility p_i, as calculated in section 3.1.1.

Following the same analysis as in [23], let $d_j(a_i)$ denote the total deferred time during the jth backoff for priority i. Because the backoff timer only decreases when the channel is idle, we have

$$d_j(a_i) = \begin{cases} \bar{F}' + k_j\bar{F} + b_j & j = 1, \\ k_j\bar{F} + b_j + \bar{F} & j > 1 \end{cases} \quad (18)$$

where b_j is the backoff time of the jth collision. k_j is a Poisson random variable with average $\lambda_i \cdot b_j$ and denotes the number of packets that are sent during the jth collision. $\bar{F}$ is the average packet length of the traffic and $\bar{F}' = F/2$ is the residual packet length that caused the collision on the first try.

Hence, given the current attempt rate λ_i and the collision possibility p_i calculated using Eq. (2) and (3), the average value of the total accumulated deferred time for priority i, denoted as d_l, can be estimated as

$$d_l = E[\sum d_j]$$
$$= \sum_{l=0}^{\infty} E \sum_{j=0}^{l} [d_j | l\_backoffs](1 - p_i)p_i^l$$
$$= \sum_{l=1}^{\infty} (\sum_{j=1}^{l} E[(k_j + 1)\bar{F} + b_j | l\_backoffs]$$
$$+ E[(k_1 + \frac{1}{2})\bar{F} + b_1 | l\_backoffs])(1 - p_i)p_i^l$$
$$+ E[(k_1 + \frac{1}{2})\bar{F} + b_1 | l\_backoffs](1 - p_i) \quad (19)$$
$$= \sum_{l=0}^{\infty} \sum_{j=0}^{l} E[(\lambda\bar{F} + 1)b_j + \bar{F}](1 - p_i)p_i^l$$
$$- \frac{\bar{F}}{2} \sum_{l=0}^{\infty} (1 - p_i)p_i^l$$
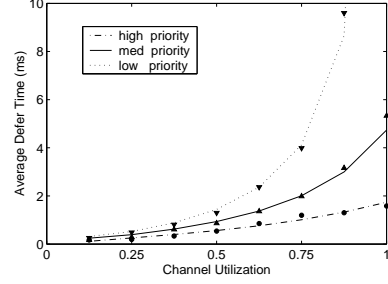
For the basic exponential backoff scheme, where $E[b_j] =$

Figure 2: Delay Model Verification (parameters are the same as in figure 1).

$(2^j \cdot CW_{min} - 1)/2$, we have

$$d_l(basic) = \frac{\lambda_i F + 1}{2} CW_{min}[\frac{2^m p_i^{m+1}}{1 - p_i}$$
$$+ \frac{1 + p_i^m - (2^{m+1} + 3)p_i^{m+1} - 2p_i^{m+2}}{1 - 2p_i}]$$
$$+ [F - \frac{\lambda F + 1}{2}]\frac{1}{1 - p_i} - \frac{\bar{F}}{2} \quad (20)$$

For our priority-based backoff scheme,

$$E[b_j] = \frac{(2^j + p_i\alpha_i)CW_{min} - 1}{2} \quad (21)$$

Hence, we have

$$d_l = d_l(basic) + [\sum_{l=m+1}^{\infty} \sum_{j=0}^{m} p_i\alpha_i + \sum_{l=0}^{m} \sum_{j=0}^{l} p_i\alpha_i]$$
$$\times (\frac{\lambda_i\bar{F} + 1}{2})(1 - p_i)p_i^l \quad (22)$$
$$= d_l(basic) + p_i\alpha_i \frac{\lambda_i\bar{F} + 1}{2} \frac{p_i(1 - p_i^m)}{1 - p_i} \cdot CW_{min}$$

Eq. (22) gives the average of the total accumulated deferred time for a packet transmission when backoff occurs. When the channel is sensed free and the flow transmits without backoff, the deferred time is zero if transmission succeeds or $(\bar{F} + d_l)$ if a collision happens. Similar to Eq. (10),

$$d_{defer,i} = P_{free,i}(1 - e^{-\lambda_i})(\bar{F} + d_l) + P_{busy,i}d_l \quad (23)$$

Let $d_{transmission,i}$ denote the transmission time of a packet, whose average is F_i, then the average service delay is

$$d_{service,i} = d_{defer,i} + d_{transmission,i} \quad (24)$$

Eq. (24) is the result for head-of-line packets. When queuing delay is considered, the total delay can be obtained by utilizing the delay results of an M/M/1 queue [4]. Specifically, suppose the traffic arrival rate is Poisson distributed with average rate of $R_{arrival,i}$, and the service rate is also Poisson distributed with average rate of $R_{service,i} = 1/d_{service,i}$. Then the total delay is

$$d_i = 1/(R_{service,i} - R_{arrival,i}) \quad (25)$$

Eq. (25) can be used for admission control to check whether the delay bound of the flow is satisfied.

Figure 2 shows the comparison between the analytical model and the simulation results for the average packet service latency (excluding queuing delay) as the traffic load increases. The lines represent the numerical results calculated using the model, while the symbols indicate the simulation results. We can see that the simulation and analytical results are close to each other, thereby verifying our analysis.
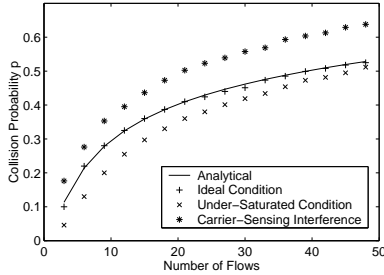
5

Figure 3: Impact of under-saturated nodes and hidden terminals.

3.2 MBRP in Real-world Environments

3.2.1 Impact of Under-Saturated Nodes

The basic MBRP mechanism described in section 3.1 assumes nodes are saturated, i.e., all nodes in the network have packets in their queue for transmission. When nodes are not saturated, our analysis may overestimate the actual number of collisions in the network. Figure 3 shows the impact of under-saturated nodes. The line represents the collision possibility calculated through the analysis in section 3.1.1 versus the number of competing flows, assuming all nodes are saturated. The symbols that match the line are the simulation results with all nodes saturated. The lower symbols are the measurement results for simulations with a certain percentage (30%) of unsaturated nodes, while the remaining nodes are operated in saturated conditions. The difference indicates that when nodes are not saturated, the basic MBRP analysis overestimates the collision rate in the network. This leads to lower throughput and higher calculated delay.

3.2.2 Impact of Hidden Terminals

Another assumption used by MBRP is ideal channel conditions, i.e., no packet corruption, and no hidden terminals. It does not consider the fact that a node's carrier-sensing neighbors can also interfere with its transmission, even though the node cannot correctly decode the interfering packets. The impact of interfering nodes is thus not reflected in the flow set. Hence, the modeling analysis may underestimate the actual collisions in the network. This is especially true in a multi-hop network where the hidden terminal problem and carrier sensing interference become more significant.

Consider a simple topology as shown in figure 4, where node 3's transmission interferes with node 2's packet reception because the nodes are with carrier-sensing range of each other. Flows f_1 and f_2 then conflict. Because nodes 2 and 3 cannot decode each other's packets correctly, node 1 is unaware of the existence of node 3. Consequently node 1 will not include flow f_2 in its flow set. [2] The upper symbols in figure 3 show the impact of the interference from the nodes in the carrier sensing range but out of the transmission range of a node. With the interference from carrier-sensing neighbors, the basic MBRP analysis underestimates the collisions in the network.

[2]If nodes 2 and 3 are direct neighbors, node 2 will include flow f_2 in its flow set. Node 1 will thereby learn of the interference.
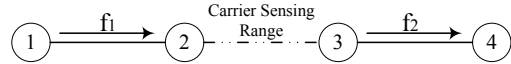

Figure 4: Scenario with "hidden terminals".

3.2.3 Adjustment using Measurement Feedback

In addition to the above described assumptions, unexpected collisions, such as those caused by control packet transmissions and random interference (e.g., microwave or other wireless transmissions), will also affect the results. As shown in figure 3, the presence of unsaturated nodes makes our model overestimate the channel usage, while unexpected packet loss due to hidden terminal, collision or interference makes our model underestimate the channel usage. Depending on the real network topology and traffic distribution, these issues will result in a discrepancy between the model-based output and the actual measurement results.

To mitigate these effects, we improve MBRP by utilizing the difference between the measured value and the model output as run-time feedback to improve the accuracy of our model. The improved analysis is called Enhanced MBRP (EMBRP).

Suppose at the time when the n-th flow is admitted, the channel collision ratio estimated by new EMBRP model is $p_{cal}(n)$. When the (n+1)-th flow is requested, the real channel collision ratio measurement is $p_{measure}(n)$. The difference between the model result and actual value is then calculated:

$$\Delta p(n) = p_{measure}(n) - p_{cal}(n) \qquad (26)$$

Hence, the difference $\Delta p(n)$ during the n-th flow request can then be used by the EMBRP model as feedback to make a better decision for the (n+1)-th request.

First, the basic MBRP model predicts the collision ratio $p_{cal}(n+1)'$ based on current flow information using the process in section 3.1. Then EMBRP combines the output from MBRP and $\Delta p(n)$ to produce the final estimation $p_{cal}(n+1)$ as the following:

$$p_{cal}(n+1) = p_{cal}(n+1)' + k_1 \Delta p(n) + k_2 \sum_{i=1}^{n} \Delta p(i) \qquad (27)$$

where k_1 is the proportionality coefficient that controls the response to changes of collision rate. A larger k_1 improves the response rate, but it may lead to systematic error or oscillation. k_2 is the integration coefficient that decreases the response rate but diminishes system error. The adjustment of the values for k_1 and k_2 is important for the accuracy of our prediction. Because we do not expect the network state to dramatically change, previous measurements are weighted more heavily, and k_1 is generally smaller than k_2. We will explain in detail the parameter selection in the experiments in section 5. Section 5 discusses estimation of the parameters in more detail.

Based on this adjusted collision rate, an adjusted λ can be calculated. Specifically, from Eq.(3), we have

$$\lambda_i = -\log(1 - p_i) \qquad (28)$$

6

Consequently, adjusted $L(a_i)$ can be calculated using Eq.(10). Adjusted throughput and delay prediction can further be obtained.

The collision rate that a node experiences is measured as the probability that a packet transmission by the node in question fails using a standalone measurement process at the MAC layer. Given a measurement interval, the process continuously measures the collision rate without being triggered by the analytical model. When the latter needs the measurement result to adjust the calculation, it obtains the results from the measurement unit.

The collision rate is calculated as following: we count the number of failed transmissions, i.e., number of packets that do not receive ACK packets, and divide it by the total number of data transmissions in a given measurement duration. Hence, we have

$$\bar{p} = N_{unacked}/N_{transmitted} \qquad (29)$$

We further use an ARMA (Auto Regressive Moving Average) [14] filter to provide run-time estimation, considering the previous results, to smooth the measurement. β is the smoothing factor.

$$\bar{p}' = \beta\bar{p} + \frac{(1-\beta)}{n}\sum_{i=0}^{n-1}\bar{p}_i \qquad (30)$$

Note that the collision rate measurement is passive and it does not incur extra communication overhead, i.e., no packet transmission is needed. Additionally, the modeling and measurement results are only calculated at the source and intermediate nodes along a flow path, i.e., the destination is not involved. This is because the interference at the reception node is included in the unacknowledged packet measurement of the node's upstream neighbor.

4 Integration with Routing Protocols

Our resource prediction model can be integrated with ad hoc routing protocols as a module sitting in between the IP routing layer and the MAC layer, as shown in figure 5.
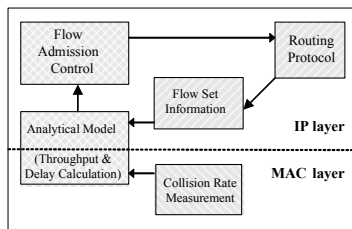

Figure 5: Integration with routing protocols at the IP and MAC layers.

MBRP provides channel statistics of a node's local contending area. However, because packet delivery often occurs in a multi-hop fashion, a local decision is not sufficient for the setup of an entire transmission path. The interference among neighboring nodes makes the estimation of channel utilization more difficult. For example, in figure 6, the circles indicate the transmission range of each node. Node A's neighborhood includes B, B's neighbors include A and C, and both B and D are C's neighbors. Suppose node A requests a new flow using

the path $A \rightarrow B \rightarrow C \rightarrow D$ to reach the destination. If the bandwidth consumption of the flow is x, then the bandwidth consumption is actually $2 \times x$ for nodes A and C, and $3 \times x$ at node B. This is because nodes within transmission range of each other contend for the shared medium. Therefore, a new flow will consume the resources in the neighborhood of all the nodes along the transmission path.
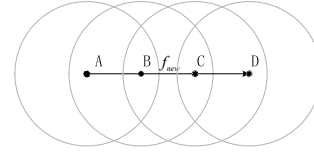

Figure 6: An example topology.

The routing process can be augmented to analyze the interference relationship among the nodes on the potential transmission path, as well as to disseminate the flow information along the path. Then, based on the potential flow set information, the estimated throughput or delay can be calculated using the analytical model described in the previous section. Finally, the source can choose the path that best meets the flow's QoS requirement.

We briefly describe how the model is integrated with reactive routing protocols, using AODV [19] as an example. The basic flow setup process can be divided into a Request and a Reply phase. In the request phase, the source node sends RREQ messages for the new flow, including QoS information such as the traffic class of the flow, the required quality, and the minimum throughput or accumulated delay through previous hops. Upon reception of the RREQ packet, each intermediate node adds a pending record for this flow and rebroadcasts the RREQ if the flow is locally admissible. This indicates that the predicted quality of the new flow is within an acceptable range, i.e., the minimum available bandwidth along the path is larger than the flow's throughput requirement, or the accumulated delay is smaller than the latency requirement. If the flow is not locally admissible, the RREQ packet is dropped. After the propagation of RREQ packets, intermediate nodes use Neighbor Reply messages (NREP) to notify neighbors about the potential load, including the new flow information. The updated flow set information, disseminated by NREP packets, serves as the input of the analytical model as described in section 3.1.1. The RREQ packet reaches the destination if a path with satisfied quality exists.

If a RREQ message is received, the destination node sends a Route Reply message (RREP) along the reverse path to the source node during the reply phase. Intermediate nodes obtain updated neighbor load information through the NREP packets in the Request phase. They now recompute the quality of service that can be provided to the flow and forward the RREP if the new flow is locally admissible. The source node selects an optimal path based on the available levels of service. Once data packet transmission begins, the nodes along the propagation path also send NREP packets to notify their neighbors that the flow has been admitted. Therefore, all nodes that are affected by the new flow obtain updated channel utilization information.

When path breaks due to node movement, route maintenance is performed so that the source node can re-discover a new valid route. In this case, flow information is updated through the new setup process. In a highly mobile environment, frequent broken paths and neighbor nodes changes will result in stale flow information and, consequently, inaccurate flow quality estimation. However, it is likely that a valid and stable route rarely exists with the high mobility. We foresee that the integration of MBRP with routing protocols will be most effective in environments with low or controlled mobility. For instance, backhaul networks consisting of wireless routers that provide multi-hop communication for mobile nodes are well-suited for our model.

The above discussion describes how MBRP can be combined with reactive routing protocols. For proactive routing protocols, flow set information can be exchanged between neighboring nodes through Hello messages or any other periodic neighbor link update messages. However, an extra call setup process is needed to accomplish the quality prediction. This is because the interference that will be caused by the new flow cannot be determined by neighbor exchanges alone. The proposed MBRP mechanism can also be integrated with QoS routing protocols in a similar manner.

5 Experimental Results

We implemented our approach in the NS-2 [11] simulator with the Monarch mobility extensions [6]. A modified MAC protocol is used to provide differentiated scheduling as described in [22]. $CWmin$ is set to 32 and m is 5. The link bandwidth is set to 2 Mbps. The value of β in the collision measurement (Eq.(30)) is set to 0.8 to place greater weight on recent measurements. The AODV routing protocol, modified as described in section 4, is utilized for multi-hop communication in the second and third sets of simulations.
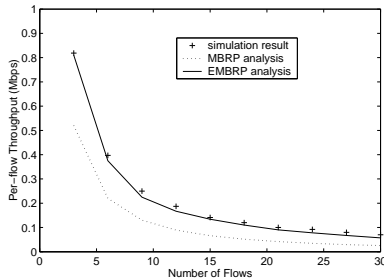
5.1 Single Broadcast Region

Figure 7: Throughput prediction with under-saturated nodes.

The first set of simulations explores MBRP performance in a single-hop scenario, where a group of source and destination pairs are all within the same broadcast region. No interference from the carrier-sensing range neighbors or hidden terminals occurs in this scenario. All the flows have the same priority where the priority adjustment parameter α equals 0. The packet size is 1000 bytes. 1/3 of all flows are operated at non-saturated condition, with a packet sending rate of 10 pkt/s, while the remaining flows all have sending rate of 100 pkt/s. MBRP is not used for admission control in this set of experiments so that results with a large number of flows can be shown.

Figure 7 shows the results of bandwidth prediction for each flow. The symbol represents the simulation results for the average per-flow throughput. The dotted line is the numerical result calculated using the basic MBRP model. The model results are lower than the actual throughput because of the overestimation of collision possibility as shown in figure 3. By using the measured collision rate to adjust the model calculation, EMBRP obtains results close to the simulation values.

The parameter k_1 in Eq.(27) is set to 0.2, and k_2 is set to 0.8. As explained in section 3.2, $k_2 > k_1$ is because more weight is given to previous results to achieve stability. As the results become more stable, the impact of $\Delta(p)$ approaches zero, while the impact of $\sum \Delta(p)$ becomes smaller because $k_2 < 1$. As more flows are admitted, nodes experience longer service delay, and consequently longer queuing latency. Hence, the impact of the understaurated condition is reduced.
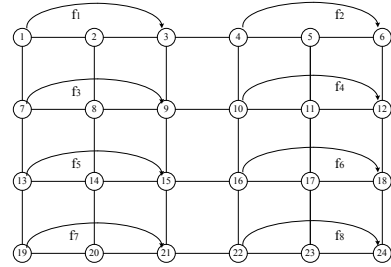
5.2 Grid Topology

Figure 8: Grid topology.

This set of simulations examines the effectiveness of applying MBRP to admission control in a multi-hop network. Specifically, the nodes form a 4×6 grid with inter-node spacing of 200m, as shown in figure 8. The parameters of the flows are indicated in table 1. For high priority VoIP flows, we require delivery delay less than 100 ms as indicated by $delay_{req}$. The minimum bandwidth requirement for high and low priority traffic is 64 kbps and 100 kbps, respectively, as indicated by bw_{req}. This traffic pattern represents a network environment where background traffic is delay-tolerant while higher priority is given to traffic with stringent real time constraints. In this set of experiments, we use the following admission policy: admit the maximum number of flows while ensuring that their received quality of the service meets their service needs. All the flows in figure 8 are started sequentially at 10 second intervals. Among them, flows 2, 4, 6 and 8 are high priority, while the rest are low priority.

Table 1: Priority traffic parameters.

Priority Class	Packet Size (bytes)	$Rate$ (Kbps)	bw_{req} (Kbps)	$delay_{req}$ (ms)
High (G.711 VoIP)	160	64	64	100
Low (CBR)	500	200	100	-

When there is no admission control, all flows start at their scheduled time. When MBRP is applied, i.e., no measurement feedback is utilized, the first 7 flows are all admitted, while the 8th flow is rejected. When EMBRP is utilized, the

8

Table 2: Throughput for flows without admission control (Kbps).

Start time	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
$t_0 + 50$	200	64	132	*57*	198	64	-	-
$t_0 + 60$	200	64	112	*56*	197	64	200	-
$t_0 + 70$	190	64	**98**	**51**	114	**50**	198	64

Table 3: Throughput for MBRP admitted flows (Kbps).

Start time	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
$t_0 + 50$	200	64	132	*57*	200	64	-	-
$t_0 + 60$	194	64	112	**55.7**	194	64	197	-
$t_0 + 70$	194	64	112	**55.7**	193	64	197	-

Table 4: Throughput for EMBRP admitted flows (Kbps).

Start time	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
$t_0 + 50$	198	64	186	64	198	-	-	-
$t_0 + 60$	194	64	174	64	196	-	200	-
$t_0 + 70$	190	64	166	64	186	-	200	64

6th flow is rejected, while the rest are all admitted. Tables 2-4 show the throughput of the admitted flows with each of the different schemes. The underlined values indicate that the flow does not receive its required quality. When there is no admission control (as shown in table 2), the throughput after time $t_0 + 50$ decreases significantly for flows that experience more contention, i.e., f_3 through f_6. When admission control is utilized, seven flows are admitted; however, the flows receive different quality, as indicated in tables 3 and 4. Flows admitted by EMBRP in general have higher throughput than that of MBRP. In particular, when f_6 is admitted using MBRP, the throughput of f_4 falls below 64 kbps, resulting in poor quality. When the measured collision rate is fed back into the model calculation, EMBRP achieves better performance because the impact of hidden terminal and carrier-sensing neighbors are included in the calculation.

Figure 9 shows the average packet latency of the high priority flows after all flows started. When there is no admission control, both f_4 and f_6 fail to meet their delay requirement. When admission control with MBRP is used, f_4 does not receive required quality. When EMBRP is utilized, all admitted flows have average latency below 100 ms. Figure 10 further illustrates the packet delivery latency for the admitted high priority flows using EMBRP. All the admitted flows have average latency of less than 100 ms. The occasional surge of latency for the flows is caused by the temporary flooding of routing control packets. Because we place more weight on the previous collision rate than the current rate, this occurrence is short-lived. Specifically, the feedback parameter k_1 is set to 0.2 as in the first set of experiments, and k_2 is set to 1.1. The intuition of choosing this value is that as more flows are admitted in the network, more interference from carrier-sensing neighbors and hidden terminals occur. The difference between the measurement and analytical results therefore increases. We also examine the value of k_2 with different topology and traf-
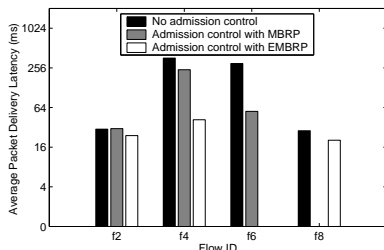
fic scenarios, for positive Δp, $k_2 = 1.1$ in general can achieve fairly accurate and stable prediction.

5.3 Random Topology

To better understand the performance of MBRP, we now examine random topologies. In this set of simulations, we generate a random topology in a $1000m \times 1000m$ area with 50 nodes. Flows are randomly chosen between node pairs and the traffic parameters are the same as described in table 1. Ten low priority flows are started at the beginning of the simulations and are used as background traffic. We then increase the number of high priority flows at 10 second intervals. The average path length is 2.7. Because the impact of hidden terminals and carrier-sensing interference is more significant than in the under-saturated conditions in a multi-hop environment, k_1 and k_2 are set to 0.2 and 1.1, respectively.

Table 5: Admission results with high priority flows.

Flow ID	1	2	3	4	5	6	7	8	9	10
accept	✓	✓	✓	✓	✓	✓	✓	×	✓	×

Table 5 shows the flow admission results using EMBRP. The 8th high priority flow is rejected because its delay requirement cannot be satisfied. The 10th flow is rejected because, if it was admitted, the quality of service of the other flows would degrade unacceptably.

Figure 11 shows the average packet delivery latency (in logarithmic scale) for the high priority flows with and without admission control. The data points represent the average delay of all the flows, while the error bars indicate the maximum and minimum delay among the flows. For instance, at 80 seconds, the average delay of the 7 admitted flows (f_8 is rejected) is 41.2 ms, while the maximum delay of one flow (f_4) is 55.2 ms, and the lowest of a flow (f_5) is 30 ms. When no admission control is performed, at time 80s the average delay of the 8 flows is 82.7 ms. However, the service needs of one of the flows (f_8) cannot be met; its delay is 127.7 ms. Similar events occur at time 100s, where f_{10} is rejected when EMBRP is used. This indicates that by predicting the per-flow quality, EMBRP assists the flow admission decision so as to meet the needed service quality constraints. Note that we shifted the

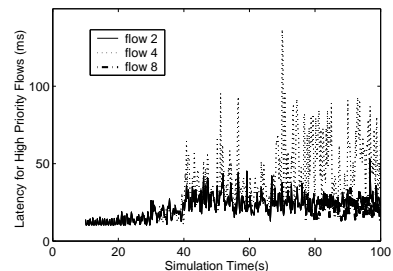**Figure 9:** Average packet delivery latency.

Figure 10: Packet delivery latency for admitted high priority flows.
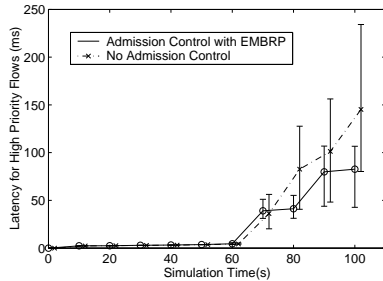
Figure 11: Average Latency for all admitted flows.

dotted lines for the results without admission control so that the error bars do not overlap.
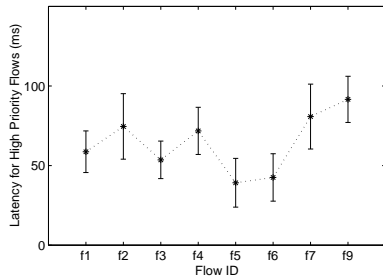

Figure 12: Packet delivery latency for admitted high priority flows.

Figure 12 further shows the packet delivery latency for each admitted high priority flow when admission control is applied. The data points are the average delay of each flow after time 100s; no new flow is admitted after that time. The error bars indicate the 95% confidence interval of the delay for each individual flow. The results show that the quality of service needs of each flow is met, thereby verifying the effectiveness of EM-BRP as an admission control solution.

6   Conclusion

This paper proposes a model-based resource prediction mechanism that supports real time communication in multi-hop wireless networks. An analytical model for differentiated MAC scheduling protocol is given, with adjustments for the multi-hop environment. The model can predict per-flow and system-wide throughput and delivery latency, thereby enabling admission control of the flows and providing an efficient network management utility. This is beneficial in the deployment of a real ad hoc network where knowledge of resource allocation and consumption is needed to meet the service requirements.

References

[1] I. Ada and C. Castelluccia. Differentiation Mechanisms for IEEE 802.11. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, Anchorage, Alaska, April 2001.

[2] G.-S. Ahn, A. Campbell, A. Veres, and L.-H. Sun. Supporting Service Differentiation for Real-Time and Best-Effort Traffic in Stateless Wireless Ad Hoc Networks (SWAN). *IEEE Transactions on Mobile Computing*, 1(3):192–207, July-Septemeber 2002.

[3] M. Barry, A. T. Campbell, and A. Veres. Distributed Control Algorithms for Service Differentiation in Wireless Packet Networks. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, Anchorage, Alaska, 2001.

[4] O. Bertsekas and R. Gallager. *Data Networks, 2nd Edition*. Prentice-Hall, 1992.

[5] G. Bianchi. Performance Analysis of the IEEE 802.11 Distributed Coordination Function. *IEEE Journal on Selected Areas in Communications*, 18(3):535–547, March 2000.

[6] J. Broch, D. A. Maltz, D. B. Johnson, Y.-C. Hu, and J. Jetcheva. A Performance Comparison of Multihop Wireless Ad Hoc Network Routing Protocols. In *Proceedings of the 4th ACM/IEEE International Conference on Mobile Computing and Networking (MobiCOM'98)*, pages 85–97, Dallas, TX, October 1998.

[7] R. Carter and M. Crovella. Measuring Bottleneck Link Speed in Packet-switched Network. Technical Report BU-CS-96-006, Computer Science Department, Boston University, March 1996.

[8] I. Chakares and E. Belding-Royer. PAC: Perceptive Admission Control for Mobile Wireless Networks. Submitted for publication.

[9] S. Chen and K. Nahrstedt. Distributed Quality-of-Service Routing in Ad-Hoc Networks. *IEEE Journal of Selected Areas in Communications*, 17(8), August 1999.

[10] D.Grossman. New Terminology and Clarifications for Diffserv. Request For Comments (Draft Stardard) 3260, Internet Engineering Task Force, April 2002.

[11] K. Fall and K. Varadhan. ns Manual. *http://www.isi.edu/nsnam/ns/doc/*. The VINT Project.

[12] J. Goodman and A. Greenberg. Stability of Binary Exponential Backoff. *Journal of the ACM*, 35(3), March 1998.

[13] IEEE. 802.11e Draft 3.1, May 2002.

[14] S. M. Kay. *Modern Spectral Estimation*. Prentice-Hall, 1988.

[15] H. Kim and J. C. Hou. Improving Protocol Capacity with Model-based Frame Scheduling in IEEE 802.11-operated WLANs. In *Proceedings of the Ninth Annual International Conference on Mobile Computing and Networking (MobiCOM'03)*, pages 190–204, San Diego, CA, September 2003.

[16] D. Maltz. Resource management in multi-hop ad hoc networks. Technical Report CMU CS 00-150, School of Computer Science, Carnegie Mellon University, July 2000.

[17] K. Nichols and B. Carpenter. Definition of Differentiated Services Per Domain Behaviors and Rules for Their Specification. Request For Comments (Draft Stardard) 3086, Internet Engineering Task Force, April 2001.

[18] V. Paxson. End-to-End Internet Packet Dynamics. In *Proceedings of the ACM SIGCOMM Conference on Communications Architectures, Protocols and Applications*, pages 138–152, Cannes, France, September 1997.

[19] C. E. Perkins and E. M. Royer. Ad-hoc On-Demand Distance Vector Routing. In *Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications*, pages 90–100, New Orleans, LA, February 1999.

[20] R. S. Prasad, M. Murray, C. Dovrolis, and K. Claffy. Bandwidth Estimation: Metrics, Measurement Techniques and Tools. *IEEE Network*, 17(6), November 2003.

[21] S. Shenker, C.Partridge, and R. Guerin. Specification of Guaranteed Quality of Service. Request For Comments (Draft Stardard) 2212, Internet Engineering Task Force, September 1997.

[22] Y. Sun, E. Belding-Royer, X. Gao, and J. Kempf. A Priority-based Distributed Call Amission Protocol for Multi-hop Wireless Ad hoc Networks. Technical report, Computer Science Department, UC Santa Barbara, April 2004.

[23] A. Veres, A. T. Campbell, M. Barry, and L.-H. Sun. Supporting Service Differentiation in Wireless Packet Networks Using Distributed Control. *IEEE Journal of Selected Areas in Communications*, 19(10), October 2001.

[24] Y. Yang and R. Kravets. Contention-aware Admission Control for Ad hoc Networks. Technical Report 2003-2337, Computer Science Department, University of Illinois at Urbana-Champaign, April 2003.