Modeling the 802.11 Distributed Coordination Function in Non-saturated Heterogeneous Conditions

David Malone[1], Ken Duffy[1], Doug Leith

Abstract— **Analysis of the 802.11 CSMA/CA mechanism has received considerable attention recently. Bianchi [1] presents an analytic model under a saturated traffic assumption. Bianchi's model is accurate, but typical network conditions are non-saturated and heterogeneous. We present an extension of his model to a non-saturated environment. The model's predictions, validated against simulation, accurately capture many interesting features of non-saturated operation. For example, the model predicts that peak throughput occurs prior to saturation. Our model allows stations to have different traffic arrival rates, enabling us to address the question of fairness between competing flows. Although we use a specific arrival process, it encompasses a wide range interesting traffic types including, in particular, VoIP.**

Keywords— **802.11, CSMA/CA, non-saturated traffic, heterogeneous network.**

I. Introduction

The 802.11 wireless LAN standard has been widely deployed during recent years and has received considerable research attention. The 802.11 MAC layer uses a CSMA/CA algorithm with binary exponential back-off to regulate access to the shared wireless channel. While this CSMA/CA algorithm has been the subject of numerous empirical studies, an analytic framework for reasoning about its properties remains notably lacking. Developing analysis tools is desirable not only because of the wide deployment of 802.11 equipment but also because the CSMA/CA mechanism continues to play a central role in new standards proposals such as 802.11e. A key difficulty in the mathematical modeling of the 802.11 MAC lies in the large number of states that may exist (scaling exponentially with the number of stations). In his seminal paper, Bianchi [1] addressed this difficulty by assuming that (i) every station is saturated (i.e. always has a packet waiting to be transmitted), (ii) the packet collision probability is constant regardless of the state or station considered and (iii) transmission error is a result of packets colliding and is not caused by medium errors. Provided that every station is indeed saturated, the resulting model is remarkably accurate. However, the saturation assumption is unlikely to be valid in real 802.11 networks. Data traffic such as web and email is typically bursty in nature while streaming traffic such as voice operates at relatively low rates and often in an on-off manner. Hence, for most real traffic the demanded transmission rate is variable with significant idle periods, i.e. stations are usually far from being saturated. Indeed, to even determine if the network will be saturated for a given traffic load may require an understanding of non-saturated operation. Thus our aim in this paper is to derive a mathematical model of CSMA/CA that relaxes the restriction of saturated operation while retaining as much as possible of the attractive simplicity of Bianchi's model, in particular, the ability to obtain analytic relationships.

In Section II earlier approaches to non-saturated modeling are reviewed. In Section III the model is introduced and solved. In Section IV its predictions are verified through ns2 simulation for homogenous stations and heterogeneous stations that have one of two distinct arrival rates. In Section V, using the model, the scope for optimizing CWmin in the non-saturated context is investigated. As a case study, we consider voice-call pairs. In Section VI fairness in the heterogeneous case is analysed. In Section VII the model's scope is discussed, along with possible variations and extensions. Concluding remarks are in Section VIII.

II. Related Work

There are approaches to non-saturated modeling other than ours. In [2] a modification of [1] is considered where a probability of not transmitting is introduced that represents a station having no data to send. The model is not predictive as this probability is not known as a function of load and must be estimated from simulation. In [3] idle states are added after packet transmission to represent bursty arrivals. The number of idle states is distributed geometrically with a parameter λ, however no relationship is given between λ and the load on the system. This model also includes a full backoff before each packet transmission, which does now allow for packet inter-arrival and 802.11's post-backoff period to overlap. This model also considers multi-rate transmissions. In [4] a Markov model where states are of fixed real-time length is introduced. As observed in the paper, the derived throughput is a monotonic function of offered load, and so the model cannot predict a pre-saturation peak in throughput. In [5] a model focusing on multi-rate transmission is presented, including an infinite queue with Poisson arrivals. This model is not solved analytically and is subject to limited validation. In [6] a non-Markov model is developed, but is based on an unjustified assumption that the saturated setting provides good approximation to certain unsaturated quantities. It appears to produce inaccurate predictions. None of these previous models have considered fairness issues arising from

Work supported by Science Foundation Ireland grant IN3/03/I346.
The authors are with the Hamilton Institute, National University of Ireland, Maynooth, Co. Kildare, Ireland.
[1] Joint first authors.

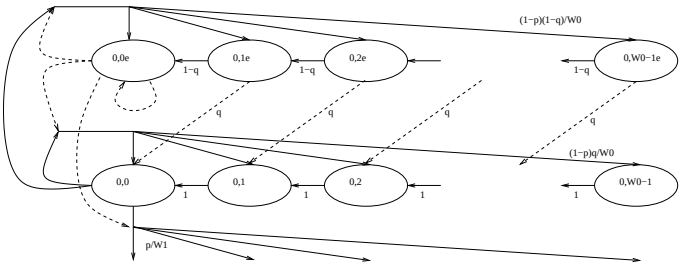


Fig. 1. Non-saturated Markov Chain.

different traffic arrival rates. The p-persistent approach of the 802.11 MAC has also been studied extensively, for recent work see [7] and the references therein.

III. Model of non-saturated heterogeneous stations

Following the seminal paper of Bianchi [1], much of the analytic work on 802.11 MAC performance has focused on saturated networks where each station always has a packet to send. For notable examples, see [8], [9]. The saturation assumption enables queueing dynamics to be neglected and avoids the need for detailed modeling of traffic characteristics, making these networks particularly tractable.

Networks do not typically operate in saturated conditions. Internet applications, such as web-browsing, e-mail and voice over IP exhibit bursty or on-off traffic characteristics. Creating an analytic model that includes fine detail of traffic-arrivals and queueing behavior, as well as 802.11 MAC operation, presents a significant challenge. We introduce a model with traffic and buffering assumptions that make it sufficiently simple to give explicit expressions for the quantities of interest (throughput per station, delay, collision probabilities), but still capture key effects of non-saturated operation. Although our traffic assumptions form only a subset of the possible arrival processes, we will see they are useful in modeling a wide range of traffic, including voice conversations. As in [1], our fundamental assumption is that each station has a fixed probability of collision when it attempts to transmit, irrespective of its history.

A. Per-station Markov Model

Bianchi [1] presents a Markov model where each station is modeled by a pair of integers (i, k). The back-off stage, i, starts at 0 at the first attempt to transmit a packet and is increased by 1 every time a transmission attempt results in a collision, up to a maximum value m. It is reset after a successful transmission. The counter, k is initially chosen uniformly between $[0, W_i - 1]$, where typically $W_i = 2^i W_0$ is the range of the counter and W_0 is the 802.11 parameter CWmin. While the medium is idle, the counter is decremented. Transmission is attempted when $k = 0$.

We introduce new states $(0, k)_e$ for $k \in [0, W_0 - 1]$, representing a station which has transmitted a packet, but has none waiting. This is called post-backoff. The first two stages of the new chain are depicted in Figure 1. Note that $i = 0$ in all such states, because if $i > 0$ then a collision has occurred, so we must have a packet awaiting transmission.

We assume that for each station there is a constant probability $1 - q$ that the station's buffer has no packets awaiting transmission at the start of each counter decrement[1]. This enables us to derive relationships between the per-station quantities: q, the probability of at least one packet awaiting transmission at the start of a counter decrement; m, the maximum backoff stage; p, the probability of collision given the station is attempting transmission; P, the Markov chain's transition matrix; b, the chain's stationary distribution; and τ, the stationary distribution's probability that the station transmits in a slot. These relationships can be solved for p and τ, and network throughput predicted. It is important to note that the Markov chain's evolution is not real-time, and so the estimation of throughput requires an estimate of the average state duration. Later, when we discuss multiple stations, we will subscript each of these per-station quantities with a station label.

Under our assumptions, we have for $0 < k < W_i$

$$\begin{array}{rrcl} 0 < i \le m, & P[(i,k-1)|(i,k)] & = & 1, \\ & P[(0,k-1)_e|(0,k)_e] & = & 1-q, \\ & P[(0,k-1)|(0,k)_e] & = & q. \end{array}$$

If the counter reaches 0 and a packet is queued, then we begin a transmission. We assume there is a station-dependent probability p that other stations transmit at the same time, resulting in a collision. In the case of a collision we must increase the backoff stage (or discard). In the case of a successful transmission we return to backoff stage 0 and the station's buffer is empty with probability $1 - q$. In the case with infinitely many retransmission attempts we need introduce no extra per-station parameters and for $0 \le i \le m$ and $k \ge 0$ we have

$$\begin{array}{rcl} P[(0,k)_e|(i,0)] & = & \frac{(1-p)(1-q)}{W_0}, \\ P[(0,k)|(i,0)] & = & \frac{(1-p)q}{W_0}, \\ P[(\min(i+1,m),k)|(i,0)] & = & \frac{p}{W_{\min(i+1,m)}}. \end{array}$$

Naturally, these transitions could be adapted to allow discards after a certain number of transmission attempts.

The final transitions are from the $(0,0)_e$ state, where post-backoff is complete, but the station's buffer is empty. In this case we remain in this state if the station's buffer stays empty. If a packet arrives we have three possibilities: successful transmission, collision or, if the medium is busy, the 802.11 MAC begins another stage-0 backoff, now with a packet. With P_{idle} denoting the probability that the medium is sensed idle during a typical slot, the transitions from the $(0,0)_e$ state are:

$$\begin{array}{rrcl} & P[(0,0)_e|(0,0)_e] & = & 1-q+\frac{qP_{\text{idle}}(1-p)}{W_0}, \\ k>0, & P[(0,k)_e|(0,0)_e] & = & \frac{qP_{\text{idle}}(1-p)}{W_0}, \\ k\ge 0, & P[(1,k)|(0,0)_e] & = & \frac{qP_{\text{idle}}p}{W_1}, \\ k\ge 0, & P[(0,k)|(0,0)_e] & = & \frac{q(1-P_{\text{idle}})}{W_0}. \end{array}$$

[1]We discuss this assumption further in Section III-D and Section VII.

Given the collision probability p, the idle probability P_{idle} and per-station parameters q, W_i and m we may solve for a stationary distribution of this Markov chain. This will enable us to determine the probability, τ, that this station is attempting transmission in a typical slot.

First we make observations that aid in the deduction of the stationary distribution. With $b(i,k)$ and $b(0,k)_e$ denoting the stationary probability of being in states (i,k) and $(0,k)_e$, as b is a probability distribution we have

$$\sum_{i=0}^{m}\sum_{k=0}^{W_i-1} b(i,k) + \sum_{k=0}^{W_0-1} b(0,k)_e = 1. \tag{1}$$

We will write all probabilities in term of $b(0,0)_e$ and use the normalization in equation (1) to determine $b(0,0)_e$. We have the following relations. To be in the sub-chain $(1,k)$, a collision must have occurred from state $(0,0)$ or an arrival to state $(0,0)_e$ followed by detection of an idle medium and then a collision, so that $b(1,0) = b(0,0)p + b(0,0)_e q(1-p)p$. Neglecting packet discard, for $i > 1$ we have $b(i,0) = p^{i-1}b(1,0)$ and so

$$\sum_{i\geq 1} b(i,0) = \frac{b(1,0)}{1-p} = \frac{b(0,0)p + b(0,0)_e qpP_{\text{idle}}}{1-p}. \tag{2}$$

The keystone in the calculation is then the determination of $b(0,W_0-1)_e$. Transitions into $(0,W_0-1)_e$ from $(0,0)_e$ occur if there is an arrival, the medium is sensed idle and no collision occurs. Transitions into $(0,W_0-1)_e$ also occur from $(i,0)$ if no collision and no arrival occurs

$$b(0,W_0-1)_e = b(0,0)_e\frac{q(1-p)P_{\text{idle}}}{W_0} + \frac{(1-p)(1-q)}{W_0}\sum_{i\geq 0} b(i,0). \tag{3}$$

Combining equations (2) and (3) gives

$$b(0,W_0-1)_e = b(0,0)_e\frac{q(1-pq)P_{\text{idle}}}{W_0} + b(0,0)\frac{1-q}{W_0}.$$

We then have for $W_0 - 1 > k > 0$, $b(0,k)_e = (1-q)b(0,k+1)_e + b(0,W_0-1)_e$, with $b(0,k)_e$ on the left hand side replaced by $qb(0,0)_e$ if $k = 0$. Straight forward recursion leads to expressions for $b(0,k)_e$ in terms of $b(0,0)_e$ and $b(0,0)$, and so we find

$$\frac{b(0,0)_e}{b(0,0)} = \frac{1-q}{q}\left(\frac{1-(1-q)^{W_0}}{qW_0 - P_{\text{idle}}(1-pq)(1-(1-q)^{W_0})}\right). \tag{4}$$

Using these equations we can determine the second sum in equation (1)

$$\sum_{k=0}^{W_0-1} b(0,k)_e = b(0,0)_e\frac{qW_0}{1-(1-q)^{W_0}}.$$

The $(0,k)$ chain can then be tackled, starting with the relation

$$b(0,W_0-1) = \sum_{i\geq 0} b(i,0)\frac{(1-p)q}{W_0} + b(0,0)_e\frac{q(1-P_{\text{idle}})}{W_0}.$$

Iteration leads to

$$\begin{aligned}\sum_{k=0}^{W_0-1} b(0,k) &= b(0,0)_e\left[\frac{q}{1-q}\frac{W_0+1}{2}\right.\\ &\quad\left(\tfrac{q^2W_0}{1-(1-q)_0^W} + (1-P_{\text{idle}})(1-q) - qP_{\text{idle}}(1-p)\right)\\ &\quad\left.+\frac{qW_0(qW_0+q-2)}{2(1-(1-q)_0^W)} + 1 - q\right].\end{aligned}$$

Using equation(4) we can determine $b(1,0)$ in terms of $b(0,0)_e$:

$$b(1,0) = b(0,0)_e\frac{pq^2}{1-q}\left(\frac{W_0}{1-(1-q)_0^W} - (1-p)P_{\text{idle}}\right).$$

Finally, after algebra, the normalization (1) gives

$$\begin{aligned}1/b_{(0,0)_e} &= (1-q) + \tfrac{q^2W_0(W_0+1)}{2(1-(1-q)^{W_0})}\\ &\quad+\tfrac{q(W_0+1)}{2(1-q)}\left(\tfrac{q^2W_0}{1-(1-q)^{W_0}}+\right.\\ &\qquad\left.(1-P_{\text{idle}})(1-q) - qP_{\text{idle}}(1-p)\right)\\ &\quad+\tfrac{pq^2}{2(1-q)(1-p)}\left(\tfrac{W_0}{1-(1-q)^{W_0}} - (1-p)P_{\text{idle}}\right)\\ &\qquad\left(2W_0\tfrac{1-p-p(2p)^{m-1}}{1-2p}+1\right).\end{aligned} \tag{5}$$

The main quantity of interest is τ, the probability that the station is attempting transmission. A station attempts transmission if it is in the state $(i,0)$ (for any i) or if it is in the state $(0,0)_e$, a packet arrives and the medium is sensed idle. Thus $\tau = q(1-p)b(0,0)_e + \sum_{i\geq 0} b(i,0)$, which reduces to

$$\tau = b(0,0)_e\left(\frac{q^2W_0}{(1-p)(1-q)(1-(1-q)^{W_0})} - \frac{q^2P_{\text{idle}}}{1-q}\right), \tag{6}$$

where $b(0,0)_e$ is given in equation (5), so that τ is expressed solely in terms of p, P_{idle}, q, W_0 and m. Placing the station in saturation by taking the limit $q \to 1$, the model reduces to that of Bianchi [1]. With q, W_0 and m fixed for each station, in order to determine the collision probability, p, we must determine a relation between the stations competing for the medium; we do this in Section III-B. We discuss how to model P_{idle} in Section III-C and then show how q may be related to real-world offered load in Section III-D.

B. Heterogeneous Network Model

Consider the case where n stations are present, labeled $l = 1,\ldots,n$. We subscript the per-station quantities from the previous section with the station label. Equation (6) gives an expression for τ_l, the per-station transmission probability, in terms of a per-station arrival probabilities q_l and a per-station collision probability p_l. Note that

$$1 - p_l = \prod_{j\neq l}(1-\tau_j), \text{ for } l = 1,\ldots,n, \tag{7}$$

that is, there is no collision for station l when all other stations are not transmitting. With n stations, (6) and (7) provide $2n$ coupled non-linear equations which can be solved numerically for $p_1,\ldots p_n$ and $\tau_1,\ldots\tau_n$. The value

$(1-p_i)(1-\tau_i)$ is the same for all $i = 1, \ldots, n$ and represents the probability that the medium is idle ($1-p_i$ is the probability that other stations are silent and $1-\tau_i$ is the probability that this station is silent). These equations imply that different stations' collision probabilities are not the same unless their transmission probabilities are equal. In the case where the stations are homogenous, the equations (7) reduce to $1-p = (1-\tau)^{n-1}$.

The length of each state in the Markov chain is not a fixed period of real time. Each state may be occupied by a successful transmission, a collision or the medium being idle. To convert between states and real time, we calculate the expected time spent per state. To do this we consider the probability of an idle slot (i.e. 0 stations transmitting), of successful transmissions (i.e. exactly 1 station transmitting) or of a collision (i.e. $r \geq 2$ stations transmitting), which gives

$$\begin{aligned} E_s &= (1-P_{tr})\sigma + \textstyle\sum_{i=1}^{n} P_{s_i} T_{s_i} \\ &\quad + \sum_{r=2}^{n} \sum_{1 \leq k_1 < \cdots < k_r \leq n} P_{ck_1 \ldots k_r} T_{ck_1 \ldots k_r}, \end{aligned} \tag{8}$$

where:

$$P_{s_i} = \tau_i \prod_{j \neq i} (1-\tau_j)$$

is the probability station i successfully transmits; T_{s_i} is the expected time taken for a successful transmission from station i, (including overhead, ACK and frame spacing);

$$P_{ck_1 \ldots k_r} = \prod_{i=1}^{r} \tau_{k_r} \prod_{j \neq k_1 \ldots k_r} (1-\tau_j),$$

the probability that only the stations labeled k_1 to k_r experience a collision by attempting transmission;

$$P_{tr} = 1 - \prod_{i=1}^{n} (1-\tau_i)$$

is the probability at least one station attempts transmission; and σ is the slot-time; $T_{ck_1 \ldots k_r}$ is the expected time taken for a collision from stations labeled k_1 to k_r (i.e. the expectation of the maximum of the transmission times for stations k_1 to k_r, including overhead, ACK timeout and frame spacing).

Once the mean state time is known, we estimate the proportion of time that the medium is used by each station for successfully transferring data:

$$S_i = \frac{P_{s_i} L_i}{E_s}, \tag{9}$$

where L_i is the expected time spent transmitting payload data for source i. The normalized throughput of the system is then

$$S = \sum_{i=1}^{n} S_i. \tag{10}$$

In order to determine the throughput and collision probability for each station and the overall throughput, one first solves equations (7) using equations (5) and (6). Then one uses equations (8), (9) and (10).

C. Channel idle probabilities

We used P_{idle} to denote that the channel was found to be idle at the time a packet arrived in the $(0,0)_e$ state. If the MAC checks for a new packet at the beginning of each slot, then the probability that the medium is sensed idle is simply the probability that the next slot is empty given that our station is not transmitting, i.e. $P_{\text{idle}l} = \prod_{i \neq l}^{n} (1-\tau_i) = 1 - p_l$. For throughput calculations, which are based on the model's stationary distribution, we use this relationship. For calculations not based on the stationary distribution, such as MAC delay, it is more appropriate to use a real-time relation. The one that we adopt is described in Section III-E.

D. Relating offered load to model parameters

The model represents offered load using q_l, the probability that a packet becomes available to the MAC in a slot. It is important to be able to relate this parameter to the the station's offered load. Taking $q_l \to 1$ models a saturated station, where a packet is always available to the MAC.

For small buffers, a crude approximation in the unsaturated setting is to assume that packet arrivals are uniformly distributed across slots and set $q_l = \min(E_S/\text{mean inter-packet time}, 1)$. If packets arrive at the MAC in a Poisson manner with rate λ_l, then a more satisfying estimate of q_l is $1 - \exp(-\lambda_l E_S)$, the probability that one or more packets arrive in a expected slot time.

It is also possible to produce an estimate of for q_l that does not use mean slot times. In the model each slot is either idle, a transmission from a particular station or a collision caused by a particular combination of stations. The type of slot is considered to be independent and identically distributed, so we can write $q_l = \sum P[\text{packet becomes available}|\text{slot type}]P[\text{slot type}]$. For example, for constant packet lengths and Poisson arrivals we can explicitly write

$$\begin{aligned} q_l &= (1-P_{tr})\left(1 - e^{-\lambda_l \sigma}\right) + \textstyle\sum_{i=1}^{n} P_{s_i} \left(1 - e^{-\lambda_l T_{s_i}}\right) \\ &\quad + \sum_{r=2}^{n} \sum_{1 \leq k_1 < \cdots < k_r \leq n} P_{ck_1 \ldots k_r} \left(1 - e^{-\lambda_l T_{ck_1 \ldots k_r}}\right). \end{aligned} \tag{11}$$

With an infinite buffer and arrivals that are Poissonian, q_l can be identified through the well-known M/G/1 relation [10] for the likelihood the station has a packet. This requires knowing the mean MAC delay, which we derive in Section III-E.

Using a state-independent value for the probability of a packet becoming available to the MAC is an approximation for most traffic types and buffering schemes. In Section IV we will see that it can be an accurate approximation in a number of situations. This point is explored further in the Appendix.

E. Delay

We are now in a position to estimate the mean MAC delay associated with a transmission by a particular source.

Consider the situation immediately after station l completes a transmission. The station begins post backoff and chooses a backoff of k, and a packet arrives after j states. Then the mean time between the packet arrival at the MAC layer and the completion of its transmission will be

$$\begin{aligned}\Delta_l &= \sum_{k=0}^{W_0} \tfrac{1}{W_0} \sum_{j=0}^{\infty} q(1-q)^j \Delta_{ljk} \\ \Delta_{ljk} &= \begin{cases} k \geq j & (k-j)E_{s'} + (1-p)T_{s_l} + p(T_{c_l} + K_1) \\ k < j & P_{\text{idle}}((1-p)T_{s_l} + p(T_{c_l} + K_1)) + (1 - P_{\text{idle}})K_0 \end{cases}\end{aligned} \tag{12}$$

where $E_{s'}$ is the mean state length if source l is silent, T_{c_l} is the mean length of a collision involving source l, K_0 is the mean time for l to transmit a frame beginning with a stage 0 backoff,

$$K_0 = \sum_{j=0}^{\infty} \frac{2^{\min(j,m)}W_0 - 1}{2} p^j E_{s'} + \sum_{j=1}^{\infty} j p^j (1-p) T_{c_l} + T_{s_l}, \tag{13}$$

and K_1 is the mean time for l to transmit beginning with a stage 1 backoff, defined similarly.

Observe that this estimate involves conditioning on starting in particular states, and so is not a simple function of the stationary distribution of our model. Thus we use an estimate of P_{idle} that is appropriate for the real-time nature of our calculation. By considering the conditional arrival probabilities for busy and idle slots to be proportional to the lengths of those slots, we find an estimate of $(E_{s'} - (1-p_l)\sigma)/E_{s'}$, which may be substituted into (12).

F. Two Class Network Model

To study fairness of the 802.11 MAC layer, we will solve the model for two groups of stations, where all stations within each group have the same station parameters including arrival rate and payload size. Suppose there are n_1 stations in the first class and n_2 stations in the second class, then we may solve for the collision probabilities p_1 and p_2 for a station in each group using (7) to produce the coupled non-linear equations:

$$\begin{aligned} 1 - p_1 &= (1-\tau_1)^{n_1 - 1}(1-\tau_2)^{n_2}, \\ 1 - p_2 &= (1-\tau_1)^{n_1}(1-\tau_2)^{n_2 - 1}. \end{aligned}$$

Letting T_s be the time for a successful transmission and T_c be the time for a collision,

$$E_s = (P_{s1} + P_{s2})T_s + (1 - P_{s1} - P_{s2})T_c + (1 - P_{tr})\sigma,$$

where P_{si} is the probability that a station in class i, $i = 1, 2$, successfully transmits. Normalized throughput for each class is $S_1 = P_{s1}L_1/E_s$ and $S_2 = P_{s2}L_2/E_s$, where L_i is the average payload duration for a station in class i.

IV. Model verification

We first consider a homogenous group of stations and then consider the heterogeneous setting where each station has one of two arrival rates. Station parameters[2] are shown in Table I.

[2]Note that the 802.11 standards do not specify a length for ACK-Timeout. Thus the length of a collision may depend on whether a station was involved in the collision (including a vendor selected ACK-Timeout) or was an onlooker (then using EIFS). We choose $T_c = T_s$, following the spirit for the 802.11 standard. For a model of what occurs when they are set differently in a saturated situation, see Robinson and Randhawa [9].

We compare predictions of the model from Section III with simulations using the ns2 based 802.11 simulator produced by TU-Berlin [11]. We compare model predictions with simulation for various numbers of stations and arrival rates. Queues are set as small as ns2 will permit and traffic arrivals are Poisson. We show the predictions of the model for each of the input rate relationships outlined in Section III-D.

For the homogeneous case, Figure 2 shows how collision probability depends on the total normalized offered load. Figure 3 shows how the normalized throughput of the link depends on the total normalized offered load. Results for all three load relationships discussed in Section III-D are shown. In all cases there is good agreement between the model and simulations. The model has captured a number of important features of the behavior, including:

- the linear relationship between the offered load and throughput when well below saturation.
- the behavior of throughput as predicted by Bianchi's model and simulation at high offered loads (corresponding to saturation).
- for larger numbers of stations the maximum throughput is achieved before saturation in both the model and simulation. The point at which this maximum occurs is relatively insensitive to the number of stations.
- a complex transition from under-loaded to saturated with a sudden increase of collision probabilities from a low level toward their saturated values.

We note that although there are numerical differences between the predictions of each input rate relationship, the results are qualitatively similar. As expected, assuming uniformly spaced arrivals results in higher throughput predictions, whereas the technique that considers the possibility of longer than average slots results in lower throughput predictions. We have observed similar results in other situations. For clarity we will use the relationship assuming Poisson arrivals over a mean slot time for the remainder of this paper.

As a function of collision probability average delays experienced by a single station are independent of the number of stations. Thus Figure 4, which shows simulated and estimated delays, includes values from all validation experiments. The estimated delays in Figure 4 are determined by equation (12). The term K_0 from equation (13), which does not account for post-backoff, is also shown. The similarity of the estimated delay and K_0 suggest that the K_0 dominates. Both are accurate for small collision probabilities but become mild underestimates for high collision rates.

For the heterogeneous setting of where stations are divided into two classes with each class having a different arrival rate, Figure 5 shows the model's normalized throughput prediction for a station in each class, with $n_1 = 12$

W_0	31	L	364us	=	500.0 bytes @ 11Mbps
m	5	T_s	944us	=	Header + L + SIFS + δ + ACK + δ + DIFS
σ	20us	T_c	944us	=	Header + L + SIFS + δ + ACKTimeout
SIFS	10us	DIFS	50us	=	2σ + SIFS
δ	2us	ACK	304us	=	192 bits @ 1Mbps + 14 bytes @ 1Mbps

TABLE I

PARAMETERS VALUES FOR MODEL AND SIMULATION.

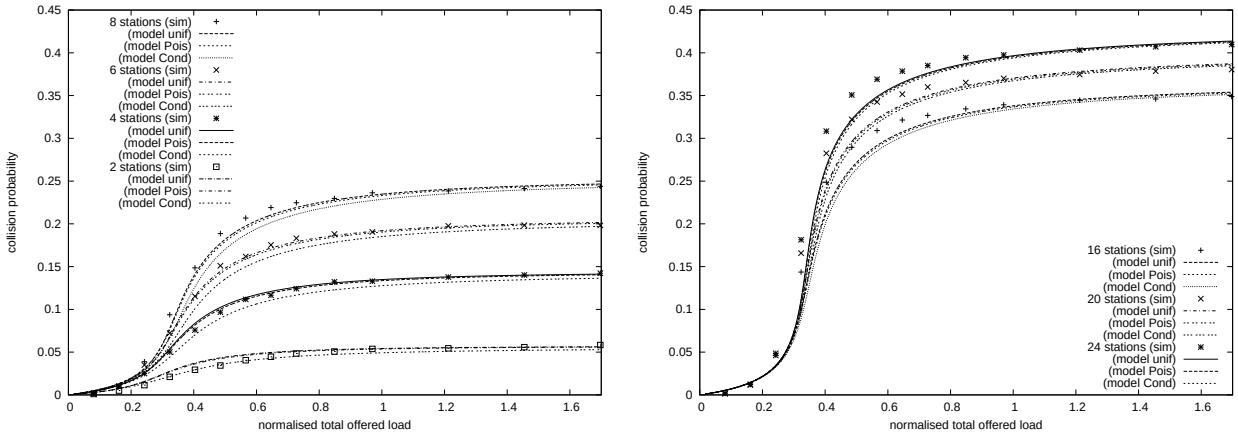


Fig. 2. Collision probability as the traffic arrival rate is varied. Results for the three load relationships (uniform, Poisson and conditional) presented in Section III-D are shown.

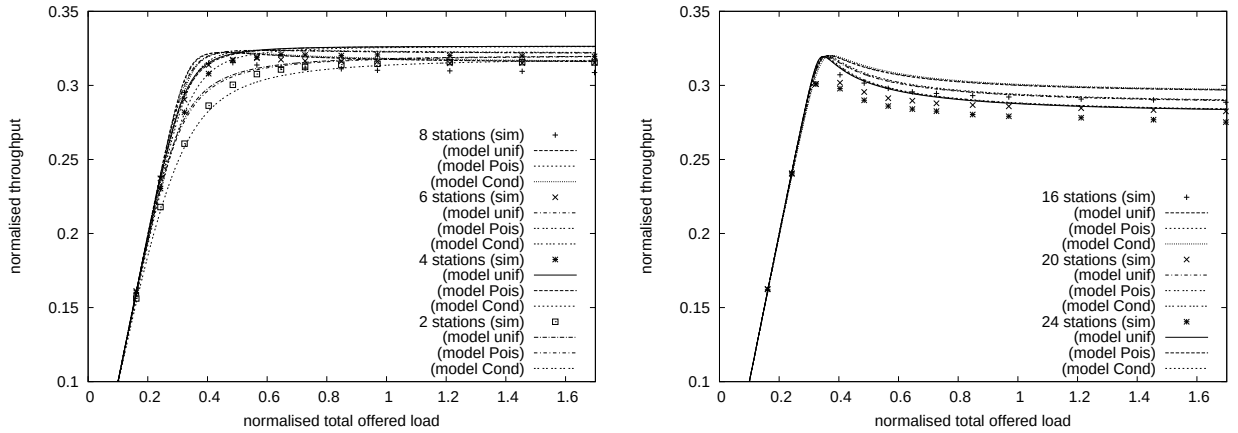


Fig. 3. Throughput as the traffic arrival rate is varied. At rates below those shown there is agreement between the model and simulation. Results for the three load relationships (uniform, Poisson and conditional) presented in Section III-D are shown.

and $n_2 = 24$. The throughput is plotted against normalized arrival rate for a station in each class. We take a representative slice through this surface along the line where the arrival rate to the second group is 1/4 of that of the first group. Figure 6 shows predicted and simulated throughputs and collision probabilities against overall normalized offered load. There is good match between predicted and observed throughputs, although the simulated collision probabilities are slightly lower than the model predicts. The collision probabilities of a station in each class are always close, but not the same. As commented after equation (7), this is expected because of an asymmetry in the system: a station in class 1 sees 11 other class 1 stations and 24 class 2 stations; a station in class 2 sees 12 class 1 stations and 23 class 2 stations.

We have taken a large number of slices for ranges of values of n_1 and n_2. For smaller numbers of users, we have found that while the predicted throughputs are accurate, the predicted collision probabilities are typically underestimates. For larger number of stations, the estimates' accuracy increases.

As a case-study we consider the predictions of the model in a situation that represents VoIP traffic in an ad-hoc network. Parameters for the voice calls are taken from [12]:

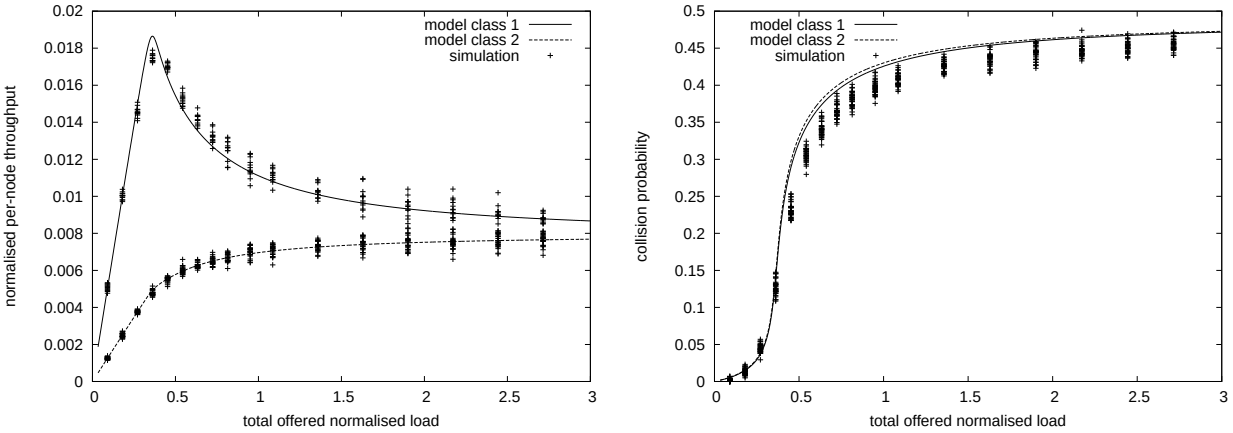


Fig. 6. Normalized per-station throughput and collision probability, where $n_1 = 12$, $n_2 = 24$ and the offered load of a class 2 station is 1/4 of a class 1 station.

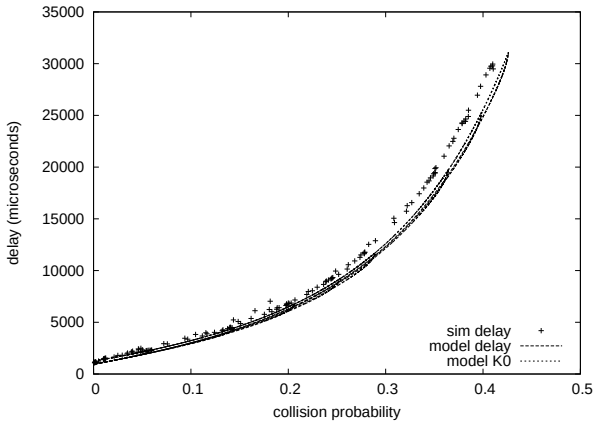


Fig. 4. Delay in the MAC as a function of collision probability.

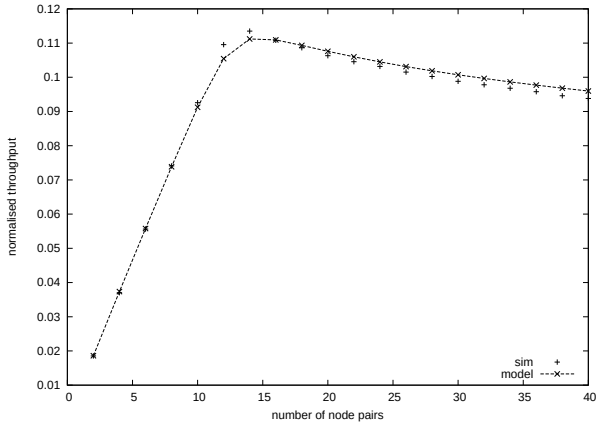


Fig. 7. Throughput for station-pairs sending 64kbps on-off traffic streams.

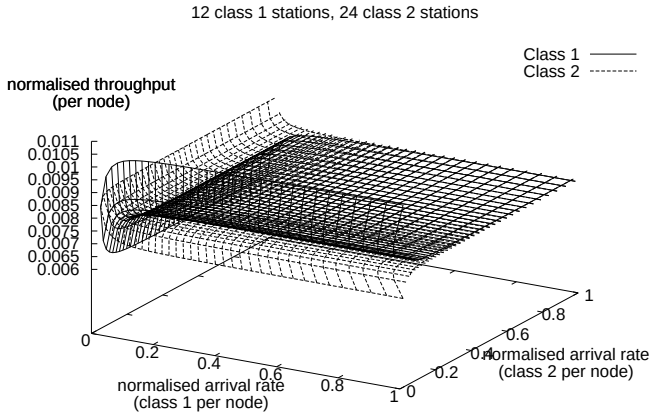


Fig. 5. Per-station throughput for two classes of stations offering different loads, $n_1 = 12$, $n_2 = 24$.

64kbps on-off traffic streams where the on and off periods are distributed with mean 1.5 seconds. Periods of less than 240ms are increased to 240ms in length, to reproduce the minimum talk-spurt period. Traffic is between pairs of stations; the on period of one station corresponds to the off period of another. When modeled, we treat each pair of stations as a single transmitter. Figure 7 shows the predicted and simulated throughput, as the number of station-pairs is increased. It can be seen that the model makes remarkably accurate throughput predictions.

V. Throughput efficiency

The value of the CWmin parameter, W_0, plays a key role in the performance of the 802.11 MAC. In saturated networks, where every station always has a packet, intuitively it is clear that a CWmin that is too large results in the medium being idle when it could be used for transmission and thus reduced throughput efficiency. Conversely, if CWmin is too small, then competing stations are more likely to attempt transmission at the same time, resulting in increased collision rates and this again leads to a reduction in throughput efficiency. Hence, there exists a value of

CWmin (dependent on the number of stations) that maximizes throughput efficiency[3].

In a network with saturated stations, it is known that the default 802.11b value of CWmin, $W_0 = 32$, does not optimize network throughput. In [1], Bianchi determines an approximate value of CWmin that optimizes throughput. Throughput efficiency in unsaturated conditions is more complex and less well understood. For example, it is known that efficiency can be significantly higher in the unsaturated setting than when saturated, see Figure 3. As we know that peak throughput occurs below saturation, we investigate what gains are potentially available by optimizing CWmin for a range of offered loads. Consider a homogenous group of stations with parameters given in Table I and three different payload sizes, 100 bytes, 500 bytes and 1000 bytes. Using the model we search for the value of CWmin predicted to produce optimal throughput. We compare this with the fixed value of CWmin, 32, from 802.11b.

Figure 8 shows the throughput and optimal CWmin value for 2 stations. We can see that the default value of CWmin is too large and that for moderate loads by reducing CWmin throughput is increased. The optimized throughput increases linearly with offered load until levelling off. The unoptimized throughput is always less that optimized throughput, even when both stations are heavily loaded. With a normalized offered load of 2, the gain in throughput is 9% for 100 byte payloads, 5% for 500 byte payloads and 3% for 1000 byte payloads.

Figures 9, 10 and 11 show the results for 10, 20 and 40 stations respectively. For light loads prior to the peak throughput, tuning CWmin does not result in a significant increase in throughput, but does create a linear relationship between offered load and throughput. Once the offered load is greater than peak throughput for CWmin=32, however, the default value of CWmin is too low, resulting in loss of throughput through collisions.

Observe that the optimal throughput plateaus at the peak throughput, implying that the optimum unsaturated throughput is no better than the optimum saturated throughput achieved by tuning CWmin. We have seen the same effects using the standard parameters from 802.11 and 802.11g, as well as 802.11b shown here. Using the sort of reasoning that is employed in [13], we consider that in a multi-access network of n homogenous independent stations there will be some transmission probability that will produce optimum throughput. In the case of 802.11, this transmission probability can be controlled by adjusting the load or adjusting CWmin. As long as the optimal transmission probability can be reached, the optimal throughput

[3]While we focus on throughput efficiency, we note that the average MAC-delay is closely related to throughput in the saturated case. Time on the medium can be used to counting down, for collisions or transmissions. Maximum throughput corresponds to minimizing the time spent during collisions and counting down. This, in turn, minimizes the time between successful transmissions. In particular, the least average MAC-delay is achieved by tuning CWmin for highest throughput.

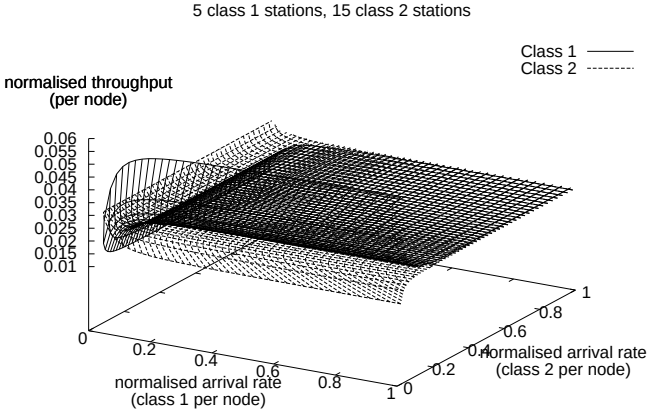


Fig. 13. Per-station throughput for two classes of stations offering different loads, $n_1 = 5$, $n_2 = 15$.

will be the same regardless of how it is achieved[4].

As a case-study of the efficiencies available through tuning CWmin, we return to the scenario introduced at the end of Section IV of VoIP traffic between stations in a peer-to-peer network. Voice call parameters are taken from [12]. Using our model, we calculated values of CWmin that optimize throughput. Simulations were then conducted using these values of CWmin and the resulting throughput is shown in Figure 12. It can be seen from Figure 12 that while tuning CWmin increases throughput by up to 10% for larger numbers of voice calls, the benefits are much less for smaller numbers of calls.

In the context of voice traffic it is important to consider the delays experienced by a frame in the MAC layer as well as throughput. Figure 12 also shows the delays for these simulations and mean plus 1.96 times the variance of the MAC-delay, corresponding to a 95% confidence interval for normally distributed data. From Figure 12 we see that the MAC delay (associated with channel contention and collisions) quickly increases when the number of voice calls rises above 10. The horizontal line marked in this figure indicates the inter-packet spacing of a single voice call; hence queueing delays quickly become unacceptable for QoS as the MAC delay approaches this value. While tuning CWmin reduces the MAC-delay's mean and variance, it has only a marginal effect for numbers of voice calls for which the delay lies below the packet duration and hence appears to offer limited practical benefit.

We conclude that while the optimal CWmin is a complex function of the traffic and the network, performance is relatively insensitive to adjustments in CWmin and the default value of 32 for 802.11b is not far from optimal in a variety of situations.

VI. Fairness

Having validated the 2-class model in Section IV, we consider the model's predictions regarding protocol fairness.

[4]This explanation was suggested to us by an anonymous reviewer.

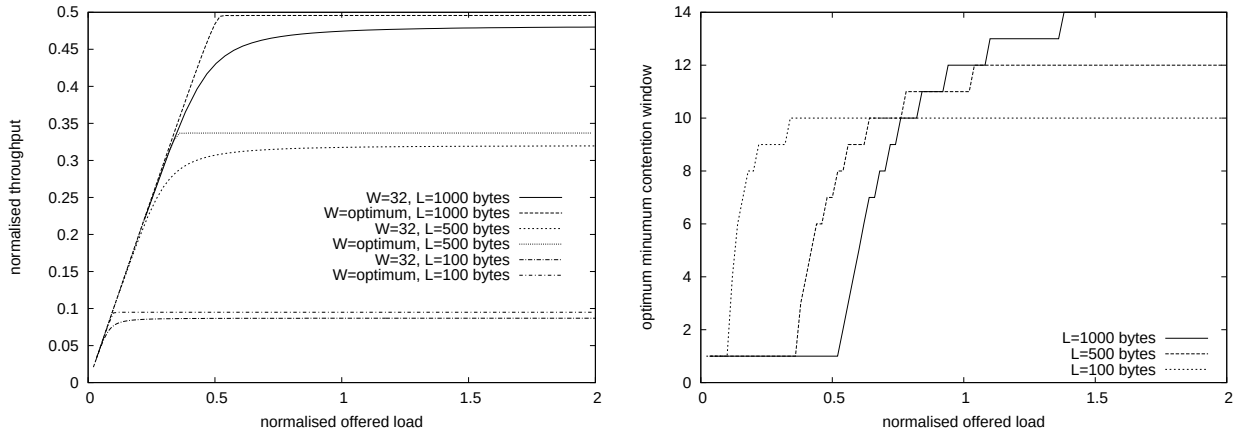


Fig. 8. Throughput for 2 stations as the offered load is varied for CWmin= 32 and with CWmin optimized. Results for various payload sizes, L, are also shown.

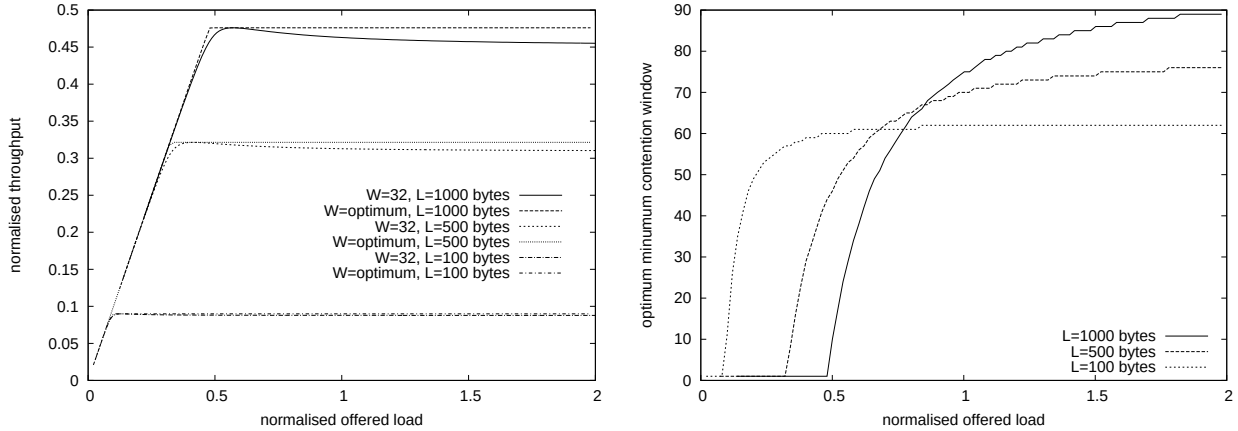


Fig. 9. Throughput for 10 stations as the offered load is varied for CWmin= 32 and with CWmin optimized. Results for various payload sizes, L, are also shown.

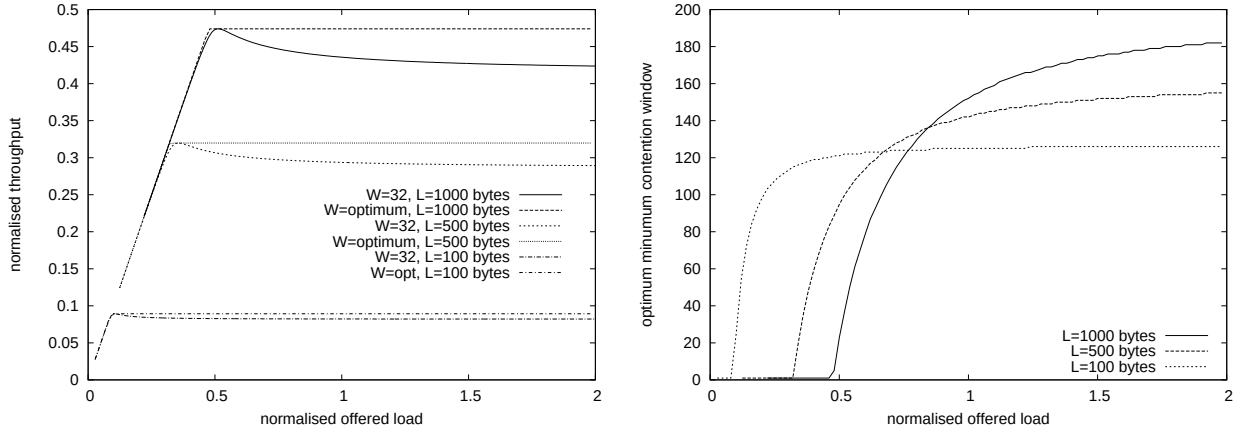


Fig. 10. Throughput for 20 stations as the offered load is varied for CWmin= 32 and with CWmin optimized. Results for various payload sizes, L, are also shown.

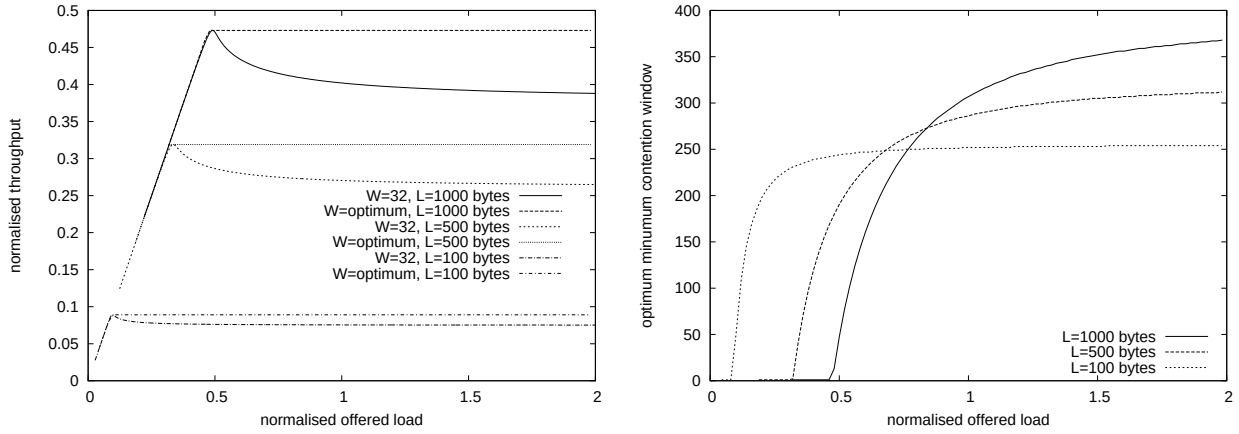


Fig. 11. Throughput for 40 stations as the offered load is varied for CWmin= 32 and with CWmin optimized. Results for various payload sizes, L, are also shown.

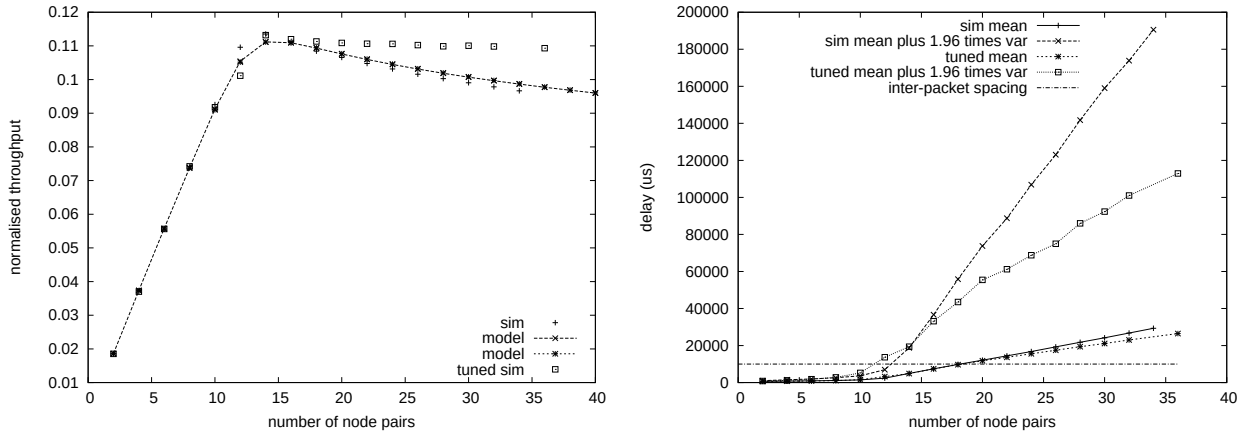


Fig. 12. Throughput and average MAC delays for station-pairs sending 64kbps on-off traffic streams.

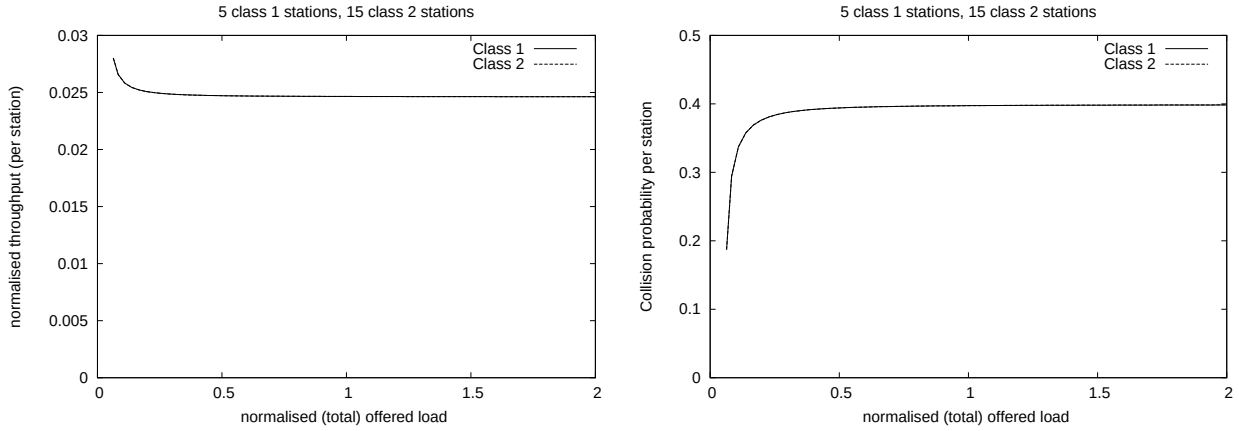


Fig. 14. Per-station throughput and collision probabilities for two classes of stations equal offered load, $n_1 = 5$, $n_2 = 15$. Class 1 and 2 throughput and collision probability are the same.

As a working definition of fairness, we consider the network to be fair if each station achieves a long-term throughput that is either at least (a) its demand or (b) a $1/n$ share of the total achieved throughput. With $n_1 = 5$, $n_2 = 15$, Figure 13 shows the normalized throughput of a station in each class against the normalized offered load of a station in each class. Station parameters are those in Table I, but with 1500byte payloads. Taking a slice along the line where the offered load from stations in both classes are equal, shown in Figure 14, demonstrates fairness in this case. The collision probabilities and throughputs of all stations are equal.

Taking slices through Figure 13 when the offered loads of stations in each class differ, however, reveals long term unfairness that is different to the well-studied short-term issue [14], [15], [16]. We fix the normalized arrival rate in class 1 per-station to be each of the four values 0.01, 0.02, 0.05 and 0.1 and vary the arrival rate per-station in class 2. Note that when class 1 stations offer 0.1 normalized load, although they are not saturated the offered load exceeds the network's capacity, even when no class 2 stations are present.

Overall normalized throughput and per-station collision probabilities are shown in Figure 15. Collision probabilities of stations in each class are approximately equal, with a maximum difference of 5% for the lowest class 1 offered load (0.01) and heavily loaded class 2 stations. At higher loads the overall channel throughput is insensitive to the class 1 arrival rate, but the bandwidth share does depend on the class 1 arrival rate; this is shown in Figure 16 where normalized throughput for a source in each class is shown against normalized offered load per source for a station in class 2.

In Figures 16 (a), (b) and (c), the network is underloaded for small class 2 offered load, so that the class 1 stations are not adversely affected by class 2. When the class 2 stations offer the same load as class 1 stations, the system is homogeneous and each station gets the same share of bandwidth. However, when the class 2 load ramps up beyond this level, class 1 stations lose their bandwidth share. The biggest drop from bandwidth fairness occurs when class 2 station are saturated, i.e. always have a packet ($q_2 = 1$). The percentage drop in throughput from fair share for these four class 1 offered loads are 16%, 32%, 22% and 8% for Figures 16 (a), (b), (c) and (d) respectively. The network is far from being fair, with greedy stations being able to steal bandwidth.

This unfairness has Quality of Service (QoS) implications. To demonstrate this we consider a scenario representing a single voice-call between two stations competing with stations carrying TCP connections. The voice-call pair is modeled as in Section IV. The stations with TCP connections have 1500 byte payloads and are saturated. Figure 17 shows that collision probabilities are approximately equal for the VoIP and TCP stations, but the TCP sources steal bandwidth from the VoIP calls, with 5 TCP flows sufficient to reduce the VoIP throughput by 50%. Note that this is despite the fair-share of the channel for the VoIP station being roughly an order of magnitude above the throughput of the VoIP station (this share is not accessible due to the non-saturated nature of the VoIP traffic).

VII. Model Scope

We assume a perfect PHY, so transmission errors are caused only by collisions and do not occur due to noise on the medium. As collisions and transmission failure because of a noisy medium are treated by the MAC in the same way, it is possible as a first approximation to add an extra, independent component to the collision probabilities to model this effect. For saturated 802.11 networks such a procedure has been carried out, see [17].

We have presented this particular model because of its accuracy, while it still remains attractively simple. Minor model variations, such as discounting carrier sense in state $(0, 0)_e$ or disallowing packet arrival immediately after transmission, are easy to consider. We have also considered a model with queue-empty probabilities conditioned on being in a transmit state or a post-backoff state, described in the Appendix. Such variations perturb the numerical results, but do not result in qualitative changes in the model's predictions. It is also straight forward to consider variations which have been studied for saturated models, such as finite retry limits and per-station backoff factors [18].

Except for saturated stations, we match mean simulation offered loads to q as described in Section III-D, even for non-Poisson traffic. As demonstrated by the examples in this paper, this approximation works well if interface buffers are short, which is a reasonable assumption for delay sensitive traffic. If interface buffers are large, but the station is not saturated, the effective offered load at the MAC is increased. This can be captured by a more elaborate queueing model, or by allowing q after a transmission to depend on the backoff stage. Alternatively the Markov chain may be extended to include buffering beyond the MAC, but not without considerable effort.

VIII. Conclusions

We have presented a model and analysis of the 802.11 MAC under non-saturated and heterogeneous conditions. The model's predictions were validated against simulation and seen to accurately capture many interesting features of non-saturated operation, including predicting that peak throughput occurs prior to saturation. We have shown that a node can approach its saturation throughput from above or below depending on factors such as the number of nodes in the system and their relative loads. We address the question of fairness between competing flows showing, for example, that saturated data flows may significantly reduce the bandwidth available to low-rate VoIP flows.

IX. Acknowledgments

We thank the anonymous reviewers for their detailed comments and suggestions that helped to clarify this work's

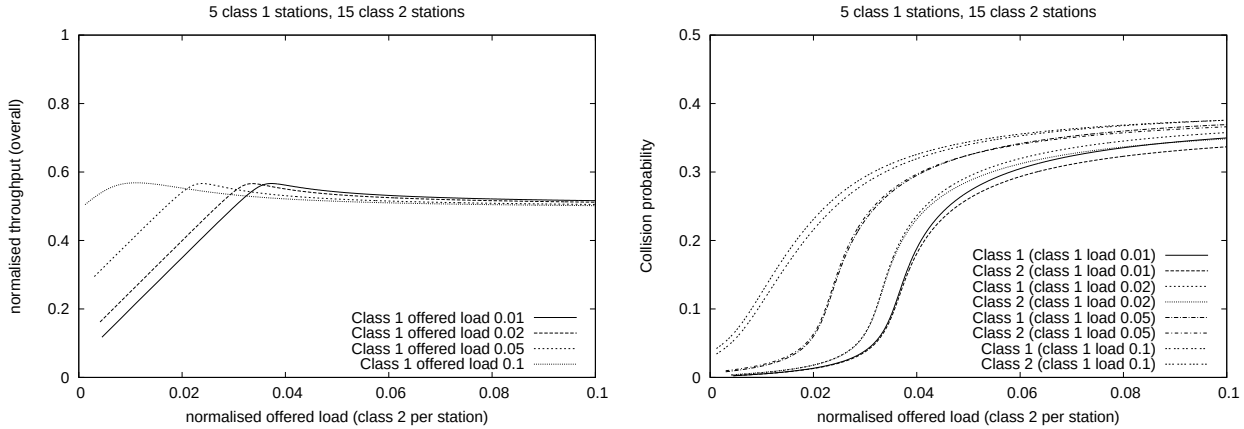
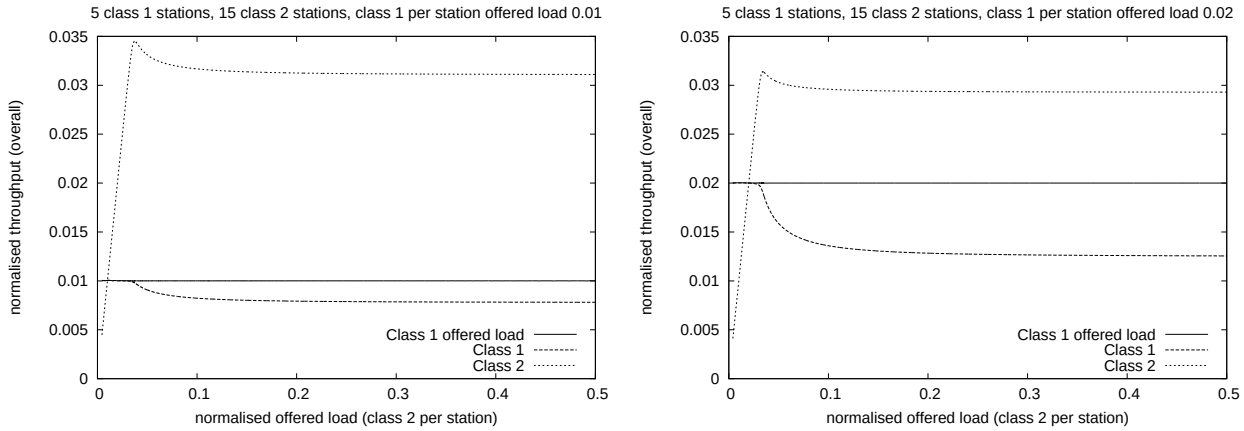
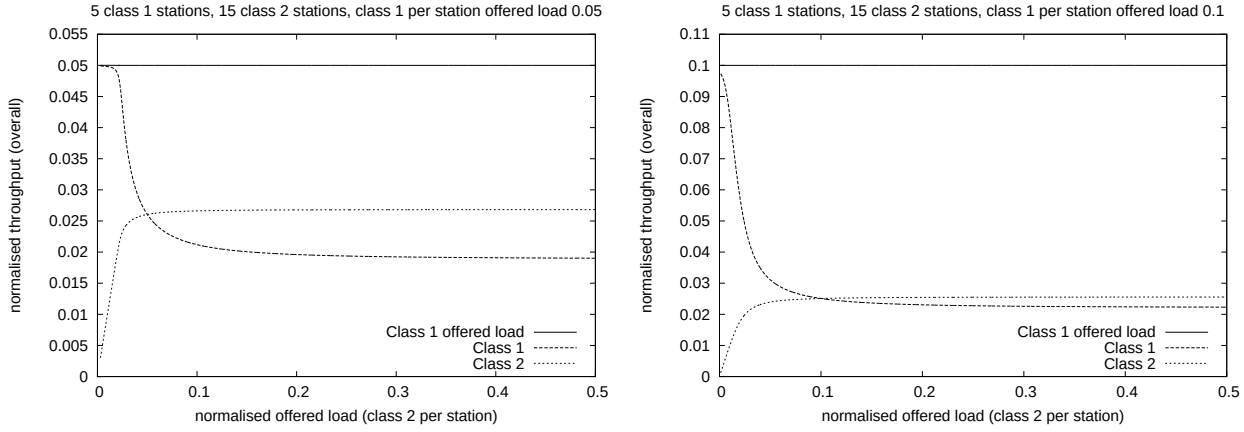


Fig. 15. Overall throughput and per station collision probabilities for two classes of stations with class 1 offering fixed per station load, $n_1 = 5$, $n_2 = 15$.



(a) Class 1 per station offered load 0.01

(b) Class 1 per station offered load 0.02



(c) Class 1 per station offered load 0.05

(d) Class 1 per station offered load 0.1

Fig. 16. Per-station throughput for two classes of stations with class 1 offering fixed per-station load, $n_1 = 5$, $n_2 = 15$.

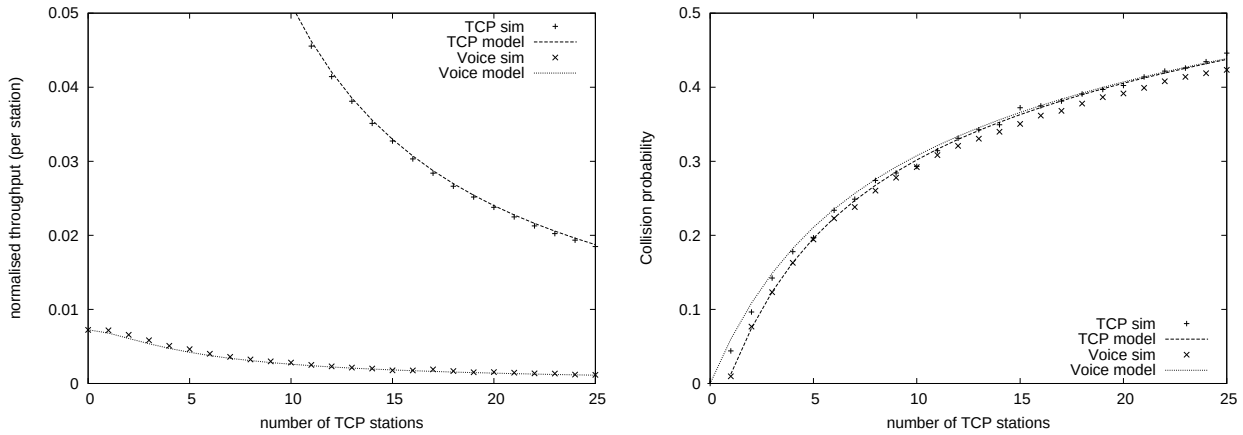


Fig. 17. VoIP and TCP.

presentation. This work was supported by Science Foundation Ireland grant 03/IN3/I396.

Appendix

A. Model with state-dependent q

As an illustration of the breadth of models considered before before settling on the one in Section III, here we describe the transition matrix and resultant equations for a model that uses conditional information in arrival probabilities. This model was not selected for two primary reasons: its predictions are similar to the selected one; and there are added computational complexities.

The variable q_{idle} is the probability of arrival during a state transition known to consist of an idle slot, q_{busy} is the probability of arrival during a state transition known to consist of a busy slot and q_{ave} is the probability of an arrival during a state transition without conditional knowledge. Thus $q_{ave} = q_{idle}P_{\text{idle}} + q_{busy}(1-P_{\text{idle}})$. The transition probabilities are as follows. A typical, $k > 0$, $(0,k)_e$ transition can consist of any sort of medium state. Thus q_{ave} is used and

$$\begin{aligned} 0 < i \le m, \quad P[(i,k-1)|(i,k)] &= 1, \\ P[(0,k-1)_e|(0,k)_e] &= 1-q_{ave}, \\ P[(0,k-1)|(0,k)_e] &= q_{ave}. \end{aligned}$$

The state after a station attempts transmission is always a long slot so that, for $0 \le i \le m$ and $k \ge 0$, we have

$$\begin{aligned} P[(0,k)_e|(i,0)] &= \tfrac{(1-p)(1-q_{busy})}{W_0}, \\ P[(0,k)|(i,0)] &= \tfrac{(1-p)q_{busy}}{W_0}, \\ P[(\min(i+1,m),k)|(i,0)] &= \tfrac{p}{W_{\min(i+1,m)}}. \end{aligned}$$

For the remaining transitions from $(0,0)_e$, a mixture of conditional information gives:

$$\begin{aligned} P[(0,0)_e|(0,0)_e] &= 1-q_{ave}, \\ P[(0,0)|(0,0)_e] &= q_{idle}P_{\text{idle}}, \\ P[(0,k)|(0,0)_e] &= \tfrac{q_{busy}(1-P_{\text{idle}})}{W_0}. \end{aligned}$$

Solving for the stationary distribution we get a normalization in terms of $b_{(0,0)}$: $1/b_{(0,0)} =$

$$\begin{aligned} &\tfrac{(1-q_{busy})}{q_{ave}} + 1 \\ &+\tfrac{1}{W_0}\left(\tfrac{(W_0-1)W_0}{2} + \tfrac{q_{busy}(1-P_{\text{idle}})(1-q_{busy})}{W_0 q_{ave}^2}\right. \\ &\left(\tfrac{(W_0-1)W_0}{2} - \tfrac{(1-q_{ave})\left(1+(W_0-1)(1-q_{ave})^{W_0} - W_0(1-q_{ave})^{W_0-1}\right)}{q_{ave}^2}\right) \\ &\quad\left.+\tfrac{1-q_{busy}}{q_{ave}}\left(-W_0 + \tfrac{1-(1-q_{ave})^{W_0}}{q_{ave}}\right)\right) \\ &+\tfrac{W_0 p(1-(2p)^m)}{1-2p} + \tfrac{p(1+W_0(2p)^m)}{2(1-p)}, \end{aligned}$$

and finally we solve for the transmission probability, $\tau = b_{(0,0)}/(1-p)$. Figure 18 illustrates the minor differences between this model's predictions and that from Section III. Thus, as this model is more computational involved, there seems little advantage in employing it instead of the model presented in the main body of this paper.

References

[1] G. Bianchi, "Performance analysis of IEEE 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, March 2000.

[2] G-S. Ahn, A. T. 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*, vol. 1, no. 3, pp. 192–207, 2002.

[3] M. Ergen and P. Varaiya, "Throughput analysis and admission control in IEEE 802.11a," *ACM-Kluwer Mobile Networks and Applications, Special Issue on WLAN Optimization at the MAC and Network Levels*, to appear.

[4] A. N. Zaki and M. T. El-Hadidi, "Throughput analysis of IEEE 802.11 DCF under finite load traffic," in *First International Symposium on Control, Communications and Signal Processing*, 2004, pp. 535–538.

[5] G. R. Cantieni, Q. Ni, C. Barakat, and T. Turletti, "Performance analysis under finite load and improvements for multirate 802.11," *Elsivier Computer Communications*, vol. 28, no. 10, pp. 1095–1109, June 2005.

[6] O. Tickoo and B. Sikdar, "A queueing model for finite load IEEE 802.11 random access," in *IEEE International Conference on Communications*, June 2004, vol. 1, pp. 175 – 179.

[7] L. Bononi, M. Conti, and E. Gregori, "Runtime optimization of IEEE 802.11 wireless lans performance," *IEEE Transactions on Parallel and Distributed Systems*, vol. 15, no. 1, pp. 66–80, 2004.

[8] R. Battiti and B. Li, "Supporting service differentiation with enhancements of the IEEE 802.11 MAC protocol: models and

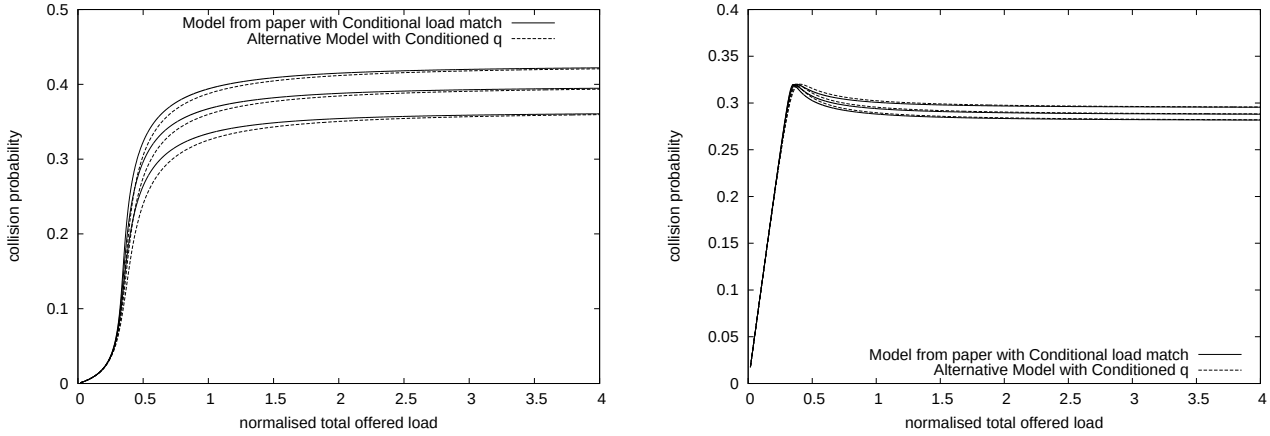


Fig. 18. Collision probability and throughput for paper and conditioned q model.

analysis," Tech. Rep. DIT-03-024, University of Trento, May 2003.

[9] J. W. Robinson and T. S. Randhawa, "Saturation throughput analysis of IEEE 802.11e enhanced distributed coordination function," *IEEE Journal on selected areas in communications*, vol. 22, no. 5, pp. 917–928, June 2004.

[10] S. Asmussen, *Applied Probability and Queues*, Springer, second edition, 2003.

[11] S. Wiethölter and C. Hoene, "Design and verification of an IEEE 802.11e EDCF simulation model in ns-2.26," Tech. Rep. TKN-03-019, Technische Universität Berlin, November 2003.

[12] A. P. Markopoulou, F. A. Tobagi, and M. J. Karam, "Assessing the quality of voice communications over internet backbones," *IEEE Transactions on Networking*, vol. 11, no. 5, pp. 747–760, October 2003.

[13] D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, 1987.

[14] C. E. Koksal, H. Kassab, and H. Balakrishnan, "An analysis of short-term fairness in wireless media access protocols," in *Proceedings of ACM SIGMETRICS*, June 2000.

[15] G. Berger-Sabbatel, A. Duda, M Heusse, and F. Rousseau, "Short-term fairness of 802.11 networks with several hosts," in *Proceedings of Sixth IFIP IEEE International Conference on Mobile and Wireless Communications (MWCN)*, October 2004.

[16] A. Kumary, "Analysis and optimisation of IEEE 802.11 wireless local area networks," in *Proceedings of the third International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WIOPT)*, April 2005.

[17] Q. Ni, T. Li, T. Turletti, and Y. Xiao, "Saturation throughput analysis of error-prone 802.11 wireless networks," *Wiley Journal of Wireless Communications and Mobile Computing*, vol. 5, no. 8, pp. 945–956, 2005.

[18] Yang Xiao, "An analysis for differentiated services in IEEE 802.11 and IEEE 802.11e wireless LANs," in *Proceedings of IEEE International Conference on Distributed Computing Systems*, 2004.