Abstract—Analysis of the 802.11 CSMA/CA mechanism has received considerable attention recently. Bianchi [3] presents an analytic model under a saturated traffic assumption. Bianchi’s model is accurate, but typical network conditions are non-saturated 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 also allows stations to have different traffic arrival rates, enabling us to address the question of fairness between competing flows.

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 [3] 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. Network traffic is frequently bursty in nature while streaming traffic such as sampled data operates at relatively low rates and often in an on-off manner, i.e. stations are often far from being saturated. Our aim in this paper is to derive a mathematical model of CSMA/CA that relaxes the restriction to saturated operation while retaining as much as possible of the attractive simplicity of Bianchi’s model, in particular, the ability to obtain analytic relationships.

II. MODELLING NON-SATURATED HETEROGENOUS STATIONS

Following the seminal paper of Bianchi [3], 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 [2], [8]. 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. In particular, traffic generated by measurement sensors and by actuator control signals is derived from sampling processes that naturally lead to on-off traffic characteristics. Creating an analytic model that includes fine detail of traffic arrivals and queuing 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 sampled data streams. As in [3], our fundamental assumption is that each station has a fixed probability of collision when it attempts to transmit, irrespective of its history.

Bianchi [3] 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 \(W_i = 2^i W\) 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 \(k = 0\).

![Non-saturated Markov Chain](image_url)
We introduce new states \((0, k)_c\) for \(k \in [0, W_0 - 1]\), representing a node which has transmitted a packet, but has none waiting. This is called postbackoff. 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. 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 \(\tau\), the stationary distribution’s probability that the station transmits in a slot. These relationships can be solved for \(p\) and \(\tau\), 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.

Under our assumptions, we have for \(0 < k < W_i\)

\[
0 < i \leq m, \quad P[(i, k - 1)|((i, k))] = 1, \\
P[(0, k - 1)_c|(0, k)_c] = 1 - q, \\
P[(0, k - 1)|(0, k)_c] = q.
\]

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 \(0 \leq i \leq m\) and \(k \geq 0\) we have

\[
P[(0, k)_c|(i, 0)] = \frac{(1-p)(1-q)}{W_0}, \\
P[(0, k)|(i, 0)] = \frac{(1-p)}{W_0}, \\
P[(\min(i + 1, m), k)|(i, 0)] = \frac{p}{W_{\min(i+1,m)}}.
\]

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

The final transitions are from the \((0, 0)_c\) state, where postbackoff is complete, but the station’s buffer is empty. In this case we remain in this state if the station’s buffer remains 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 \(P_{\text{idle}}\) denoting the probability that the medium is idle during a typical slot, the transitions from the \((0, 0)_c\) state are:

\[
P[(0, 0)_c|(0, 0)_c] = 1 - q + \frac{qP_{\text{idle}}(1-p)}{W_0}, \\
k > 0, \quad P[(0, k)_c|(0, 0)_c] = \frac{qP_{\text{idle}}(1-p)}{W_0}, \\
k \geq 0, \quad P[(1, k)|(0, 0)_c] = \frac{q(1-P_{\text{idle}})}{W_0}, \\
k \geq 0, \quad P[(0, k)|(0, 0)_c] = \frac{q(1-P_{\text{idle}})}{W_0}.
\]

Observe that \(p\), the probability of a collision given that we are about to transmit, is the probability that at least one other station is transmitting. This is also the probability that the medium is busy if we know the station under consideration has been silent. Hence we substitute \(P_{\text{idle}} = 1 - p\).

Given the collision probability \(p\) for this station in the system and per-station parameters \(q, W_i, m\) we may solve for a stationary distribution of this Markov chain. This will enable us to determine the probability, \(\tau\), 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 \(b(0, k)_c\) denoting the stationary probability of being in states \((i, k)\) and \((0, k)_c\), as \(b\) is a probability distribution we have

\[
\sum_{i=0}^{m} \sum_{k=0}^{W_i-1} b(i, k) + \sum_{k=0}^{W_0-1} b(0, k) = 1. \quad (1)
\]

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

\[
\sum_{i \geq 1} b(i, 0) \frac{b(0, 0)_c}{1 - p} = \frac{b(0, 0)_c + b(0, 0)_c q(1-p)p}{1 - p}. \quad (2)
\]

The keystone in the calculation is then the determination of \(b(0, W_0 - 1)_c\). Probability flows into \((0, W_0 - 1)_c\) from \((0, 0)_c\) if there is an arrival, the medium is sensed busy and no collision occurs. It also flows in from \((i, 0)_c\) if no collision or arrival occurs

\[
b(0, W_0 - 1)_c = b(0, 0)_c \frac{(1-p)(1-q)}{W_0} + b(0, 0)_c q(1-p), \quad (3)
\]

Combining equations (2) and (3) gives

\[
b(0, W_0 - 1)_c = \frac{b(0, 0)_c}{1 - q} \left(\frac{1-q}{W_0}\right) \left(\frac{1-q}{W_0}\right) \left(1-q\right)\frac{1-q}{W_0}. \quad (4)
\]

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

\[
\sum_{k=0}^{W_0-1} b(0, k)_c = b(0, 0)_c \frac{q W_0}{1 - (1-q) W_0}. \quad (5)
\]
The $(0, k)$ chain can then be tackled, starting with the relation

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

Recursion leads to

$$W_0 \sum_{k=0}^{W_0-1} b(0, k) = b(0, 0) e \left[ q \frac{W_0 + 1}{1 - q} + \frac{q^2 W_0}{2(1 - q) W_0^2} + p(1 - q) - q(1 - p)^2 \right] + \frac{q W_0^2 q + q - 2}{2(1 - q) W_0} + 1.$$ 

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

$$b(1, 0) = b(0, 0) e \frac{p q^2}{1 - q} \left[ \frac{W_0}{1 - (1 - q) W_0} - (1 - p)^2 \right].$$

Finally, after algebra, the normalization (1) gives

$$\frac{1}{b(0, 0)} = (1 - q) + q^2 W_0 (1 - q) + \frac{q W_0^2}{2(1 - q) W_0^2} + p(1 - q) - q(1 - p)^2$$

$$+ \frac{2 p q^2}{1 - (1 - q) W_0 - (1 - p)^2} \left[ \frac{W_0}{2(1 - q) W_0^2} + 1 \right].$$

The main quantity of interest is $\tau$, 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 $(0, 0)$, 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

$$\tau = b(0, 0) e \left[ \frac{q^2 W_0}{1 - q} - q^2 \frac{W_0^2}{1 - q} \right] \frac{(1 - q)^2}{1 - (1 - q) W_0 - (1 - p)^2},$$

where $b(0, 0)$ is given in equation (5), so that $\tau$ is expressed solely in terms of $p, q, W_0$ and $m$. While $q, W_0$ and $m$ are fixed for each station, in order to determine the collision probability, $p$, we must give a relation between the stations competing for the medium.

Consider the case where $n$ stations are present, labeled $l = 1, \ldots, n$. Equation (6) gives an expression for $\tau_l$, the per-station transmission probability, in terms of a per-station arrival process $q_l$ and a per-station collision probability $p_l$. Observe that

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

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 $p_l$ and $\tau_l$. Observe that $(1 - p_l) (1 - \tau_l)$ is the same for all $i = 1, \ldots, n$ and represents the probability that the medium is idle (as we observed before $1 - p_l$ is the probability that other stations are silent and $1 - \tau_l$ is the probability that this station is silent). Note that these equations imply that different stations’ collision probabilities are not the same unless their transmission probabilities are equal. We remark that in the case where the stations are homogenous, the equations (7) reduce to $1 - p = (1 - \tau)^{n-1}$. Placing the system in saturation by setting $q = 1$, the model reduces to that of Bianchi [3], as expected.

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 must calculate the expected time spent per state, which is given by

$$E_s = (1 - P_{\text{tr}}) \sigma + \sum_{i=1}^{n} P_{s_i} T_{s_i} + \sum_{r=2}^{m} \sum_{k_1 < \ldots < k_r \leq n} P_{c_{k_1}, \ldots, k_r} T_{c_{k_1}, \ldots, k_r} \quad (8)$$

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 $i$;

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

the probability that only the stations labeled $k_1$ to $k_r$ experience a collision by attempting transmission; $T_{c_{k_1}, \ldots, k_r}$ is the expected time taken for a collision from stations labeled $k_1$ to $k_r$;

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

is the probability at least one station attempts transmission; and $\sigma$ is the slot-time.

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

$$S_i = \frac{P_{s_i} L_i}{E_s}, \quad (9)$$

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

$$S = \sum_{i=1}^{n} S_i. \quad (10)$$

Thus 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).

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 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
\[
\Delta = \sum_{i=0}^{\infty} \frac{1}{\lambda_i} \sum_{k=0}^{\infty} q(1-q)^k \Delta_{ik}
\]
\[
\Delta_{ik} = \begin{cases} 
  k \geq j & (k-j)E_p + (1-p)T_s + p(T_c + K_1) \\
  k < j & (1-p)T_s + (1-p)p(T_c + K_1) + pK_0
\end{cases}
\]
where \(E_p\) is the mean state length if this source is silent, \(T_s\) is the mean length of a collision involving this source, \(K_0\) is the mean time to transmit a frame beginning with a stage 0 backoff, defined similarly.

Delay, as estimated directly by equation (11), is an underestimate of the simulated delay. Investigation indicates that this is because the mean state length approximation is too coarse grained for accurate delay estimation. For example, the probability of there being a packet available for transmission immediately after a successful transmission is too coarse grained for accurate delay estimation. For a specific traffic model this can be corrected. Similarly, after post-backoff a better estimate of the probability that a packet is available while the medium is idle is \((E_p - (1-p)\sigma)/E_p\).

III. MODEL VERIFICATION

We first consider a homogeneous group of stations and then consider the heterogeneous setting where each station has one of two arrival rates. Station parameters\(^1\) are shown in Table I.

We compare predictions of the model from Section II with simulations using the ns2 based 802.11 simulator produced by TU-Berlin [10]. We compare model predictions with simulation for various numbers of stations and arrival rates.

In order to move between model and simulation arrival rates, we use the following logic. Queues are set as small buffers, the parameter \(q_i\) is the probability that at least one packet arrives in the expected time spent per state, \(E_s\) defined in equation (8). In simulation, the probability that at least one packet arrives during \(E_s\) is one minus the probability that the first inter-packet time is greater than \(E_s\). Hence, when inter-packet arrival times are exponentially distributed the exponential rate \(\lambda_i\) should be set so that \(q_i = 1 - \exp(-\lambda_iE_s)\), i.e. \(\lambda_i = -\log(1-q_i)/E_s\). With \(\lambda_i\) so chosen, the arrival rate in the model and in simulation agree.

For the homogeneous case, Figure 2 shows how collision probability depends on the total normalized offered load. 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.

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 (11), post-transmission q corrected using \(1 - (1-q)T_s/E_p\). The term \(K_0\) from equation (12), 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 \(n_1 = 12\) 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

\(^1\) Note that the 802.11 standards do not specify a length for ACKTimeout. The length of a collision may depend on whether a station was involved in the collision (including a vendor selected ACKTimeout) 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 [8].
\[ W_0 = 31 \quad L = 364 \text{us} = 500.0 \text{ bytes @ 11Mbps} \]
\[ m = 5 \quad T_s = 944 \text{us} = \text{Header} + L + \text{SIFS} + \delta + \text{ACK} + \delta + \text{DIFS} \]
\[ \sigma = 20 \text{us} \quad T_c = 944 \text{us} = \text{Header} + L + \text{SIFS} + \delta + \text{ACKTimeout} \]
\[ \text{SIFS = 10us} \quad \text{DIFS = 50us} = 2\delta + \text{SIFS} + \delta + \text{DIFS} \]
\[ \delta = 2\text{us} \quad \text{ACK = 304us} = 192 \text{ bits @ 1Mbps} + 14 \text{ bytes @ 1Mbps} \]

**TABLE I**

Parameters values for model and simulation.

---

Fig. 2. Collision probability as the traffic arrival rate is varied.

Fig. 3. Throughput as the traffic arrival rate is varied. For throughput rates below those shown there is agreement between the model and simulation.

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

---

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 a sampled data stream in an ad-hoc network. A 64Kbs sampled data stream is transmitted between pairs of stations. 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 predictions.
probability of not transmitting is introduced that represents a\inge. In [1] a modification of [3] is considered where a
work see [4] and the references therein.

802.11 MAC has also been studied extensively, for recent
for competing traffic types. The
p

setting and so have not been able to consider fairness issues
of these previous models have gone beyond the homogeneous
quantities. It appears to produce inaccurate predictions. None
of these previous models have gone beyond the homogeneous
setting provides good approximation to certain unsaturated
but is based on an unjustified assumption that the saturated
limited validation. In [9] a non-Markov model is developed,
is presented, but not solved analytically and is subject to
virtue of its design it cannot predict the pre-saturation peak in
saving feature of the 802.11 MAC. In [11] a Markov model
be estimated from simulation. In [6] idle states are added
after packet transmission to represent bursty arrivals in a
way that does not account for postbackoff, a key bandwidth
saving feature of the 802.11 MAC. In [11] a Markov model
where states are of fixed real-time length is introduced, but by
virtue of its design it cannot predict the pre-saturation peak in
is presented, but not solved analytically and is subject to
limited validation. In [9] 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 gone beyond the homogeneous
setting and so have not been able to consider fairness issues
for competing traffic types. The \(p\)-persistent approach of the
802.11 MAC has also been studied extensively, for recent
work see [4] and the references therein.

IV. RELATED WORK

There are alternative approaches to non-saturated model-
ing. In [1] a modification of [3] 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 [6] idle states are added
after packet transmission to represent bursty arrivals in a
way that does not account for postbackoff, a key bandwidth
saving feature of the 802.11 MAC. In [11] a Markov model
where states are of fixed real-time length is introduced, but by
virtue of its design it cannot predict the pre-saturation peak in
is presented, but not solved analytically and is subject to
limited validation. In [9] 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 gone beyond the homogeneous
setting and so have not been able to consider fairness issues
for competing traffic types. The \(p\)-persistent approach of the
802.11 MAC has also been studied extensively, for recent
work see [4] and the references therein.

V. 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 through-
put occurs prior to saturation.

VI. ACKNOWLEDGEMENTS

This work was supported by Science Foundation Ireland
grant IN3/03/I346.

REFERENCES

differentiation for real-time and best-effort traffic in stateless wireless
ad hoc networks (SWAN). IEEE Transactions on Mobile Computing,
[2] Roberto Battiti and Bo Li. Supporting service differentiation with
enhancements of the IEEE 802.11 MAC protocol: models and analysis.
nation function. IEEE Journal on Selected Areas in Communications,
802.11 wireless LANs performance. IEEE Transactions on Parallel and
under finite load and improvements for multirate 802.11. to appear
in IEEE 802.11a. ACM-Kluwer MONET Special Issue on WLAN
of voice communications over internet backbones. IEEE Transactions
analysis of IEEE 802.11e enhanced distributed coordination function.
IEEE Journal on selected areas in communications, 22(5):917–928,
802.11 random access. In IEEE International Conference on Commu-
802.11e EDCF simulation model in ns-2. Technical Report TKN-
03-019, Technische Universität Berlin, November 2003.
dCF under finite load traffic. In First International Symposium on
Control, Communications and Signal Processing, pages 535–538,
2004.