Modeling the Impact of Buffering on 802.11

Ken Duffy and Ayalvadi J. Ganesh

Abstract—A finite load, large buffer model for the WLAN medium access protocol IEEE 802.11 is developed that gives throughput and delay predictions. This enables us to investigate the impact of buffering on resource allocation. In the presence of heterogeneous loads, 802.11 does not allocate transmission opportunities equally. It is shown that increased buffering can help this inequity, but only at the expense of possibly significantly increased delays.

Index Terms—WLAN, IEEE 802.11, performance evaluation.

I. INTRODUCTION

By design, in a network of equally loaded stations the IEEE 802.11 Distributed Coordination Function (DCF) Medium Access Controller (MAC) gives, in the long run, symmetric access. Each station gets an approximately equal number of access opportunities. However, in most deployments offered loads are asymmetric. For example consider the typical usage case: an infrastructure mode network where the Access Point (AP) has a wired back-haul. Load at the AP is approximately proportional to the number of stations in the network.

Using a finite load analytic model of 802.11 it is reported by Malone et al. [1] that in the presence of heterogeneous loads there is long-term inequity, with heavily loaded stations gaining more than their “fair share” of the bandwidth. For example, a 64kbps two-way voice conversation is considered in the presence of stations that always have a packet to send. With as few as 5 heavily loaded stations, the voice conversation’s throughput is less than 32kbps. This inequity is due to the asymmetric nature of typical offered loads: data traffic, such as web and email, is typically bursty in nature while streaming traffic operates at relatively low rates and often in an on-off manner. To capture this analytically we require a finite load model. Note that short-term unfairness in 802.11 has previously reported (for example, see [2], [3], [4]), but it is fundamentally different to this long-term unfairness.

Stations with short buffers are treated in [1]. Here we extend the modeling paradigm developed in [5], [1] to give expressions for stations with large buffers and Poisson arrivals. This enables us to consider the impact of buffering on bandwidth-share inequity. With large buffers one expects that the inability to win access to the medium results in a backlog of packets awaiting transmission. This leads to an effectively higher offered load (in comparison to a short buffer where traffic arrives while a packet awaiting access to the medium is lost) and thus a return towards a more equitable bandwidth share. With a short buffer packet loss is a good quality-of-service indicator and delay is less important. With large buffers total delay (MAC plus queuing delay) is the most important performance indicator, so we provide an estimate of it. Note that these large buffer expressions for throughput and delay can be used in conjunction with the mesh network model proposed in [6].

II. PRELIMINARIES

As we extend the methodology in [5], [1] to treat an infinite buffer with Poisson arrivals, we start with a brief overview of the model. It is a mean-field Markov model of the sort introduced by Bianchi [7]. With a network of \( N \) stations, we assume that each station \( n \in \{1, \ldots, N\} \) has a fixed probability \( p_n \) of collision given it is attempting transmission, irrespective of its back-off stage. We describe the offered load of each station \( n \) by two probabilities, \( q_n \) and \( r_n \), that are internal to the model. We will relate these to real-world offered load. When a station is in post-backoff or its count-down has completed and it is awaiting a packet, \( q_n \) denotes the probability a packet arrives to the MAC during an average slot time on the medium (which can be occupied by no station transmitting, a station successfully transmitting or a collision). The parameter \( r_n \) corresponds to the probability that immediately after a successful transmission a packet is available to the MAC. This is a generalization from [5], [1] where \( q_n = r_n \), as when a station has no buffer these are the same, but in the presence of buffers they differ.

Under these assumptions, the back-off procedure forms an embedded (non real-time) Markov chain. Its stationary distribution can be calculated explicitly by the derivation described in [1] to give an expression for \( \tau_n \) where

\[
\tau := \frac{1}{\eta} \frac{1}{1-r} \left( \frac{q^2 W_0}{(1-p)(1-(1-q)W_0)} - r q (1-p) \right),
\]

where

\[
\eta := (1-q) \left( \frac{q^2 W_0}{2(1-(1-q)W_0)} + \frac{W_{0+1}}{2(1-r)} \left( \frac{q^2 W_0}{1-(1-q)W_0} + qp(1-r) - qr(1-p)^2 \right) + \frac{W_{0-1}}{2(1-r)(1-p)} \left( \frac{q^2 W_0}{1-(1-q)W_0} - r q (1-p)^2 \right) \right),
\]

\( W_0 \) is the station’s minimum contention window and \( W_0^{2m} \) is the station’s maximum window size.

For given probabilities \( \{q_n, r_n\} \), the conditional collision probabilities \( \{p_n\} \) and transmission probabilities \( \{r_n\} \) are completely determined by the solution of the fixed point equations that say that the probability a station doesn’t experience...
a collision given it is attempting transmission is the probability that no-one else is attempting transmission:

\[ 1 - p_n = \prod_{i \neq n} (1 - \tau_i(p_i, q_i, r_i)) \quad \text{for } n \in \{1, \ldots, N\}. \tag{2} \]

As the Markov chain doesn’t evolve in real-time, to make real-time predictions we must determine the expected time between counter decrements (as given in [7]):

\[ T := (1 - P_{tr}) \sigma + P_{tr} P_s T_s + P_{tr} (1 - P_s) T_c, \tag{3} \]

where \( P_{tr} = 1 - \prod_{i=1}^{n} (1-\tau_i) \), \( P_s = \sum_{i=1}^{n} \tau_i \prod_{j \neq i} (1-\tau_j) / P_{tr} \), \( E \) is the time spent transmitting payload data (which for simplicity we assume is the same for packets from all stations; general expressions can be found in [1]), \( \sigma \) is the time for the counter to decrement, \( T_s \) is the time for a successful transmission and \( T_c \) is the time for a collision. For example, the throughput of station \( n \) is then \( S_n = \tau_n (1 - p_n) E / T \).

III. RELATING OFFERED LOAD TO \((q, r)\)

To make the model predictive we related the internal load parameters \((q_n, r_n)\) to real-world offered load. In [5], [1], a relation is given in the absence of buffers, so that \( q_n = r_n \). With i.i.d. inter-arrivals with exponential distribution \( t_n \) of rate \( \lambda_n \), the probability that no packet arrives during an average transition time in the Markov chain is \( 1 - q_n = P(t_n > T) = \exp(-\lambda_n T) \), where \( T \) is given in equation (3). Thus for a given collection of arrival rates \( \{\lambda_n\} \), one solves (2) for a range of \( \{q_n\} \), identifying a collection such that \( q_n = 1 - \exp(-\lambda_n T) \) for all \( n \).

Here we give a new relation based on an infinite buffer with Poisson arrivals. We relate the probability \( q_n \) to \( \lambda_n \) as above, but \( r_n \) no longer equals \( q_n \). We treat each station as an M/G/1 queue, where the service time distribution \( G \) is the throughput of station \( n \) is then \( S_n = \tau_n (1 - p_n) E / T \)

1. \[ T := (1 - P_{tr}) \sigma + P_{tr} P_s T_s + P_{tr} (1 - P_s) T_c, \]

2. \[ \lambda_n, E(B(p_n^*)^2)T^2 \]

IV. MODEL VALIDATION

Although it is not possible to present extensive validation due to space constraints, Figure 1 gives a good indication of the model’s throughput accuracy. All stations have 560 byte
Note the sudden, sharp climb in delay as a function of offered load for a network with 10 stations. Figure 2 plots the pre-saturation peak reported in [5] for short buffers, although present, is less pronounced for large number of stations. For a network with 10 stations, Figure 2 plots throughput and delay predictions versus simulation results. Note the sudden, sharp climb in delay as a function of offered load, which takes place near peak throughput.

**VI. CONCLUSIONS**

Increasing buffering to enable a station grab its “fair” share of the bandwidth in an asymmetrically loaded network is a double-edged sword. It aids with bandwidth share, but does so at the penalty of significantly increased delays. Thus increased buffering is probably not a suitable solution for time-constrained traffic such as VoIP and it is necessary to use the feature set of 802.11e.

**ACKNOWLEDGMENT**

The authors thank D. Malone for the simulation results.

**REFERENCES**


