An Analysis of Short-Term Fairness in Wireless Media Access Protocols

Can Emre Koksal, Hisham Kassab, and Hari Balakrishnan
Massachusetts Institute of Technology
{hari,hans,emre}@mit.edu

Abstract

We investigate the problem of unfairness over short time scales in decentralized wireless media access (MAC) protocols. Motivated by experimental results over a CSMA/CA-based WaveLAN wireless LAN that shows starvation and degraded TCP performance, we seek to derive a framework for evaluating and analyzing fairness in the context of distributed MAC protocols. In this paper, we develop two short-term fairness metrics and analyze CSMA/CA, showing quantitatively that while it reduces collision probabilities via exponential backoff, it is unfair over short time scales even for small population sizes. In contrast, ALOHA has better fairness properties but much higher collision probability. Our first fairness metric uses a sliding window scheme coupled with the Kullback Leibler distance from information theory, while the second one uses renewal reward theory based on Markov chain modeling of MAC protocols. Short-term fairness is important in several contexts, e.g., smooth acknowledgment flow for TCP connections and low jitter for real-time audio and video; we therefore hope that these measures will be used by MAC protocol designers in conjunction with traditional performance measures such as the collision probability to evaluate overall protocol performance.

Extended version of short paper from ACM SIGMETRICS 2000, Santa Clara, CA, June 2000.

1 Motivation

The increasing importance of mobile computing has led to the availability of a number of wireless LAN technologies. Most LANs use a distributed *media access (MAC)* protocol[1] to arbitrate amongst several contending transmitter stations. These protocols are designed to minimize *collisions,* which occur if two or more stations transmit at roughly the same time. Collisions often garble the data bits at the receivers. Examples of such protocols include ALOHA [1], Carrier Sense Multiple Access (CSMA) protocols [11], Time Division Multiple Access (TDMA) protocols [3], MACA [10], and MACAW [4].

Of particular interest to us is the CSMA family of protocols because of their widespread popularity, and slotted ALOHA because of its simplicity. In addition to being decentralized, these protocols are *randomized*, trading off the complexities of dynamic allocation and control messaging (required in TDMA-style protocols) for occasional collisions. A key principle used in these protocols is that of *exponential backoff*, where an unsuccessful transmission of a frame causes a waiting time (before the next attempt) that is roughly double the current backoff time. Various researchers have shown that exponential backoffs are necessary to achieve long-term stability (low collision probability) in a large (infinite) population; for example, see [2] for an exposition on this topic.

The efficiency of a MAC protocol can be measured using two independent parameters: *collision*

[1] Also called a channel access protocol.

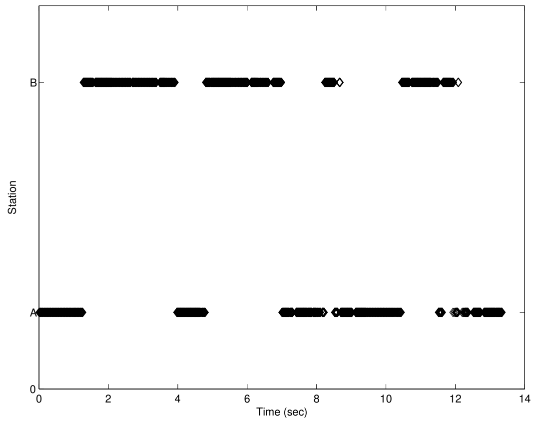

Figure 1: Typical packet trace resulting from two stations (A and B) contending for the channel using WaveLAN's CSMA/CA protocol.

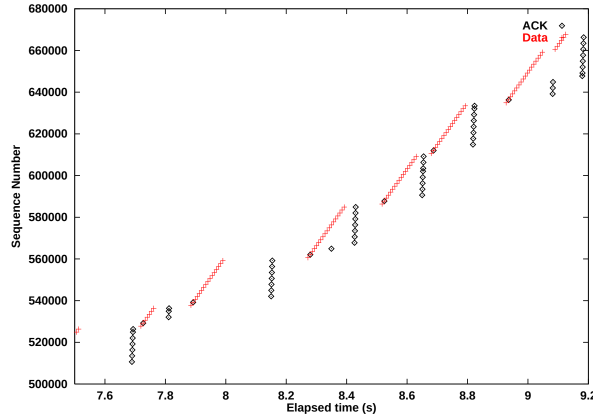

Figure 2: TCP transfer across a WaveLAN channel showing severe ACK compression and burst transmission because of short-term starvation.

probability and *fairness* across competing stations. The lower the probability of collision, the higher the resulting throughput (in terms of probability of successful transmission per attempt). Exponential backoffs have the nice property that they reduce the collision probability to a small fraction (often less than 1%) of the number of transmissions. However, it is unclear that the channel bandwidth is shared equitably by all the contending stations in these situations.

In this paper we focus on analyzing the fairness of randomized MAC protocols and introduce two metrics for measuring *short-term fairness.* Unlike past efforts (e.g., [5], [8]) that studied fairness over long time scales, we focus on the distribution of accesses of different stations to a shared channel over short time scales. To see why short-term fairness is important, consider the packet sequence traces shown in Fig. 1. It shows a typical packet trace of two stations contending for a wireless channel using Lucent WaveLAN's CSMA/CA (Carrier Sense Multiple Access/Collision Avoidance) protocol. In this experiment, two distinct pairs of stations are transmitting data at a constant rate of 500 Kbps each using 1000-byte packets; the maximum rated bandwidth of the channel is 2 Mbps. We see that each station tends to monopolize the channel for several packet transmission times before it is captured by the other station, which in turn monopolizes the channel for a long time, even though at each point in time the other station has packets waiting to be sent. This observation about WaveLAN CSMA/CA was described out in [13] and [15], and similar observations about the Ethernet channel running CSMA/CD (collision detection) were previously made in [12] and [7]. In this paper, we systematically investigate this problem and derive metrics to analyze and quantify the degree of unfairness.

Unfairness over short time scales where a small number of stations starve others has significant performance implications for applications and transport protocols. Applications like real-time audio and video are delay-sensitive and perform better when jitter is low. TCP performance degrades greatly when MAC protocols exhibit short-term unfairness, as shown in Fig. 2. Here, acknowledgments sent from the receiver do not reach the sender in a timely manner, resulting in large transmission bursts. Thus, instead of observing the smooth ACK-clocked transmissions characteristic of ideal TCP transfers where data packets are interspersed with ACKs, we observe a severe form of *ACK compression* [16]. Here, all the ACKs for a window arrive in closely spaced bursts, in response to which the sender transmits several data packets in close succession. This transmission pattern degrades performance for two reasons. First, bursts of data tend to cause losses because bottleneck routers usually do not accommodate bursts well. Second, as Fig. 2 shows, there is an idle period between each burst of data and ACKs during which the link is unused.

This paper investigates why certain MAC protocols display such behavior and how we can analyze and quantify the degree of unfairness over small time scales. Fig. 1 shows that while the channel tends to be monopolized over short horizons by one station, each station has the channel for roughly half the time over a long time horizon. Thus, even though CSMA/CA does not appear to be fair over a short horizon, it does tend towards a fair allocation over long periods of time.

We develop two metrics for measuring the short-term fairness of a MAC protocol such as CSMA/CA. The first method takes packet traces of transfers as input and sweeps through differently-sized sliding windows, evaluating the distribution of accesses in each window, and returns a consolidated short-term fairness metric. The second method is analytic; it takes a description of the MAC protocol and models it as a Markov chain where the states correspond to the backoff values of the different stations. It assigns rewards to transitions as a concave function of the states involved, and derives the expected reward as the short-term fairness metric. This method can be used to analyze other parameters of the protocol (e.g. burstiness, collision probability), by appropriately changing the reward structure. Although this method is primarily analytical, it can also be used to analyze experimental packet trace data.

Our results indicate that these short-term fairness metrics are a promising way to quantify the behavior of random access protocols over short time scales. Because short-term fairness is an important consideration for the performance of real-time and TCP applications, we hope that these metrics will be adopted by MAC designers, in addition to conventional metrics like throughput and collision probability, in the design of future systems. We also believe that our work helps explain TCP performance anomalies (such as the one shown in Fig. 2) and jitter for real-time streams over many contention-based wireless channels.

2 The Problem

This section details the problem and formally describes the CSMA/CA and slotted ALOHA protocols. We start with some important assumptions and observations about short-term fairness.

1. Short-term fairness implies long-term fairness, but not vice versa. If stations have equal share of the channel over a short horizon, then in the long run (which is a large number of short horizons put together) the stations will also have equal share of the channel.

2. We assume that all contending stations in our model always have frames to send. This is reasonable because the question of fairness arises only when there are different stations competing for the channel. We can thus think of this as modeling a snapshot in time of the system, when a subset of all stations have one or more frames to send, and that subset of stations competes for the channel. The number of stations in the model will then be equal to the number of stations in the subset.

3. With the above assumption, the closer the performance of the random access scheme to perfect round-robin TDMA (without idle slots), the better. When all stations have data to transmit, round-robin TDMA is the optimal way to share the channel in terms of both throughput and short-term fairness (assuming equal frame lengths).

   Unfortunately an efficient implementation of TDMA that does not incur idle slots is complex. Time slots have to be assigned and reassigned dynamically to stations, as stations alternate between being active and idle. Furthermore, coordination of the transmission schedule among stations requires a control channel to be present.

4. We separate measures of the probability of collision and probability of dropped frames from that of fairness. Different random access schemes handle collisions differently; for example, in slotted ALOHA, if a transmitted frame collides with another, then that frame is

repeatedly retried until it is received successfully. In CSMA/CA the frame is dropped after fifteen unsuccessful attempts. We therefore argue that any fairness measure should only consider the share of successful transmissions over the channel that each station gets. Other performance measures should be used to determine other components of the protocol performance, such as collision . A combination of the different performance measures should then be used to evaluate the overall performance of the protocol.

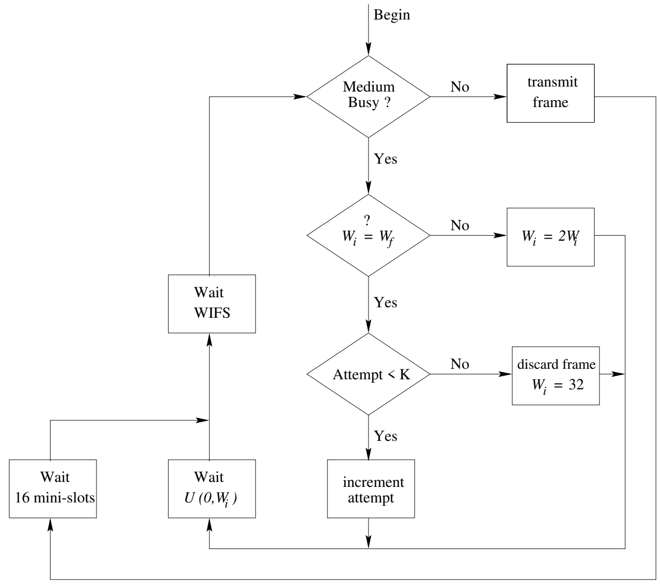

Figure 3: This flowchart summarizes the procedure Station i follows to transmit a frame in WaveLAN CSMA/CA. W_i, W_f and K represent the current backoff timer for Station i, maximum possible value of the backoff timer and the maximum possible number of backoffs respectively.

2.1 CSMA/CA

CSMA/CA is a distributed multiple access protocol, similar to CSMA/CD used over wired Ethernets. Because collision detection cannot be done over RF-based wireless LANs, these networks implement *collision avoidance* heuristics. Our work is based on CSMA/CA as implemented in Lucent's popular WaveLAN system [14]. Here, each station follows the procedure shown in Fig. 3 and summarized below.

The medium is divided into mini-slots in time. If a station has a frame to transmit, it senses the medium to see if it is busy. If not, then it can transmit right away. If it senses a busy channel, then it goes into backoff mode and implements collision avoidance scheme as follows.

Upon initially sensing a carrier, the backed-off station continues to sense the medium until the channel becomes idle. At this point both the currently transmitting station and the backed-off station wait for a period of time, called the WaveLAN Inter-Frame Space (WIFS). At the end of WIFS, the (just-finished) transmitting station waits for another mandatory 16 mini-slots, after which it senses the channel. If the channel is idle, then the station transmits the frame.

On the other hand, at the end of WIFS, the backed-off station selects a *backoff time* before retrying. The backoff time is uniformly distributed over the range $[1, W_i]$, where W_i is called the *backoff window* for Station i. W_i is initially set to 32 mini-slots as shown in Fig. 4 (e.g., if Station i randomly selects 2 as its backoff time, then the transmission will start after a delay of one mini-slot). After the backoff delay, if no carrier is sensed, the backed-off station starts transmitting its frame. Notice that the choice of mini-slots as units for the backoff mechanism has the property that a station starting one mini-slot after another will be able to detect that station's energy and conclude that the medium is busy.

If the carrier continues to be sensed (by Station i) after the backoff delay, then the station updates its backoff window, W_i. W_i doubles each time,

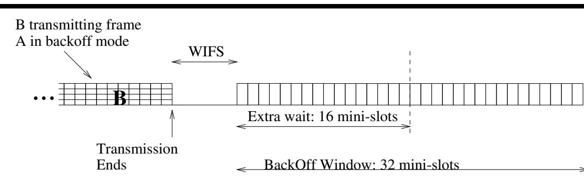

Figure 4: Illustration of WaveLAN CSMA/CA with two stations.

thus increasing exponentially; hence the term *exponential backoff.* This continues until it reaches a fixed maximum W_f. In WaveLAN CSMA/CA (as in many others), W_f is 256 mini-slots.

The total number of retries for a given frame is limited to a maximum, K, after which the frame is dropped. WaveLAN CSMA/CA sets K to 15. In our model, we assume that the number of stations in the model, N, is such that $K \geq N-1$. If this does not hold, then even if CSMA/CA became completely fair and emulated TDMA, some stations would necessarily back off and drop frames. To prove this, suppose the order of channel stations is Station 1, Station 2, $\cdots$, Station N, and $K < N-1$. When Station 1 is transmitting, Station N in Backoff Stage 1. When Station $N-1$ is transmitting, Station N is in backoff Stage $N-1 > K$, implying that Station N has to drop its frame. We note that in practice the number of stations in the entire system will usually be greater than the number of backoff stages. However, recall that the number of stations in our model represents the number of stations in the subset of stations competing for the channel. Therefore, our assumption implies that the typical number of stations competing for the channel at any one time is less than the number of backoff stages.

A Markov chain representing the dynamics of a two-station CSMA/CA system is shown in Fig. 5. Each state in the chain represents a tuple corresponding to the backoff stages of each station, with T corresponding to the current transmitter. The fraction on each arc is the probability of transition between the states at the ends of the arc. Because the backoff stages are not updated if a collision occurs, collisions are represented by self-transitions. Recall that after the WIFS period, the station that sent the last frame will delay for another 16 mini-slots and the other station will wait for a backoff period selected uniformly over $[1, W]$, where W is the backoff timer of the backed-off station. If this number is less than 16, then the backed-off station will capture the channel. If it is exactly 16, then a collision occurs, while if it is greater than 16, the same station keeps the channel and the other backs off further. Thus, the probability that the same station transmits the next frame is $\frac{15}{W}$, the probability of collision is $\frac{1}{W}$, and the probability that the backed-off station captures the channel is $\frac{W-16}{W}$, as shown in Fig. 5.

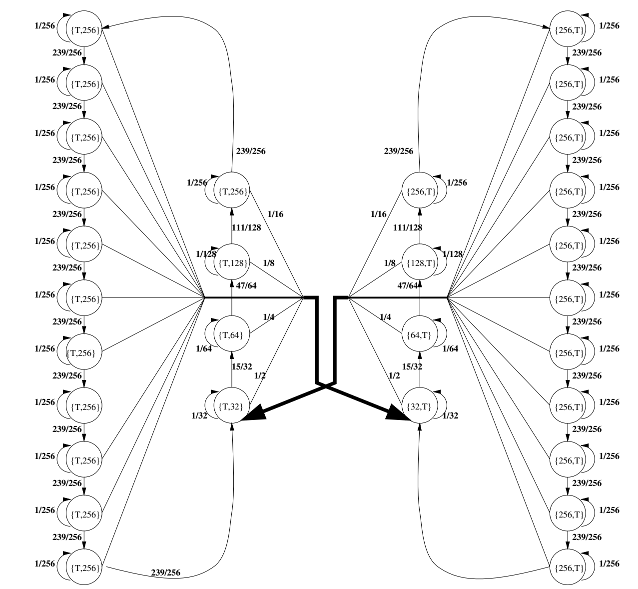

Figure 5: Markov chain representation for a two-station CSMA/CA with fifteen backoff stages.

2.2 Slotted ALOHA

Slotted ALOHA, a significantly simpler protocol than CSMA/CA. Time is divided into slots, and stations transmit frames over the channel at the beginning of a time slot. If a collision occurs (and we assume that stations can detect collisions), then each transmitting station waits a random number of slots, and then retries its frame. The random wait time has a geometric distribution, and is independent of the random durations picked by the other stations. This process of waiting a random number of slots followed by retry is repeated until the frame is transmitted successfully. There are no exponential backoffs.

If nodes always have frames to transmit, collisions will be frequent. However if, as we indicated earlier, we ignore collisions and focus only on successful transmissions, then the dynamics of Slotted ALOHA can be modeled using straightforward Markov chains. Each state in the Markov chain indicates the last station to make a success-

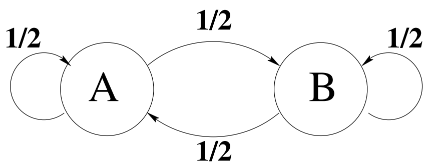

Figure 6: Underlying Markov chain for Slotted ALOHA with two stations.

ful transmission. Fig. 6 shows the Markov chain corresponding to having two stations.

We now turn to the evaluation of short-term fairness in these protocols.

3 Sliding Window Method

The first of the two methods we consider for measuring fairness in contention-based channels is the *sliding window method (SWM)*. The method is appealing because it can readily be applied to experimental or simulated traces without any detailed analysis of the protocol. Unfortunately, although applying SWM analytically (i.e., by using the underlying stochastic process, without running experiments or simulations) is possible, it seems impractical due to its high computational complexity.

SWM is motivated by the observation that short-term fairness depends on the length of the short-term horizon one cares about. To illustrate this, consider the following two traces:

$$\cdots\text{ABABABABABAB}\cdots$$

and

$$\cdots\text{AAAABBBBAAAA}\cdots$$

The first trace is an ideal TDMA-style trace, and we can easily argue that it exhibits perfect short-term fairness independent of the time horizon one considers. In the second trace, if the short-term horizon is four transmission times, then the the trace lacks short-term fairness in some parts of it. For example, the sequence of the first four elements is all A's. However, if the short-term horizon is eight, then each contiguous sequence of 8 elements does indeed have an equal number of A's and B's (assuming that the pattern of 4 A's followed by 4 B's continues).

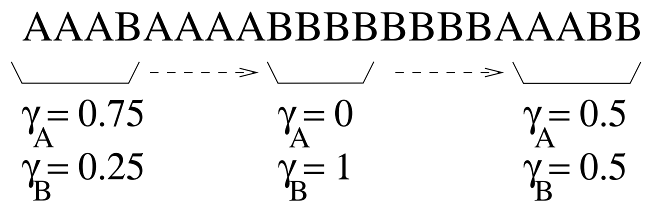

Figure 7: Illustration of the Sliding Window Method.

SWM starts with a packet trace of channel accesses and slides a window of size w across it, as shown in Fig. 7 for a window of size 4. The elements in the first window, in Fig. 7, are the first 4 elements of the sequence, AAAB. We refer to the elements within a window as a snapshot. So as we slide the window, one element at a time, we obtain a series of snapshots, where consecutive snapshots have $(w-1)$ elements in common. For each snapshot we compute the fractions of A's and B's. Let those fractions be γ_A and γ_B respectively. So in Fig. 7 the first snapshot has the following fractions:

$$\gamma_A = 0.75$$

$$\gamma_B = 0.25$$

Now, for each snapshot, we measure the fairness within it, using a per-snapshot *fairness index*. One possibility is to use Jain's fairness index [5], [9], defined as follows:

$$F_{\mathrm{J}} = \frac{\left(\sum_{i=1}^{N} \gamma_i\right)^2}{N \sum_{i=1}^{N} \gamma_i^2}$$

where N is the number of stations.

Another possibility is to derive a new index based on the Kullback Leibler distance from information theory [6] as explained in Section 3.1. After sliding the window through the entire sequence we end up with a sequence of fairness values; we calculate its average. This average corresponds to the fairness metric associated with window size w. We repeat the process with increasing window sizes, and plot the average fairness value versus the window size. In general, we expect the fairness to improve with increasing w.

The key question is to identify the window size above which the protocol exhibits adequate fairness. We show how to interpret SWM results in the Section 3.2 after introducing a new fairness index based on the Kullback Leibler distance.

3.1 The Kullback Leibler Fairness Index

We introduce the term *fraction distribution*, and denote it by Γ. If the number of stations in the model is N then Γ is an N-tuple containing the individual fractions of the stations.

For fairness to be achieved within that snapshot, the ideal fraction distribution with N stations, denoted by $\tilde{\Gamma}$ is

$$\tilde{\Gamma} = \left[\frac{1}{N}, \cdots, \frac{1}{N}\right]$$

We are interested in measuring the *deviation* of the measured fractions, γ_i from the ideal value of $1/N$. We turn to the Kullback Leibler distance (or *relative entropy*), used in information theory to measure the distance between two probability distributions. If p and q are two probability distributions, then the Kullback Leibler distance between them is denoted by $D(p\|q)$. It is a measure of "the inefficiency of assuming that the probability distribution is q when the true distribution is p" [6] (page 18). We note that Kullback Leibler distance is not a distance measure in the true sense, because it does not exhibit commutativity ($D(p\|q) \neq D(q\|p)$). The Kullback Leibler distance is appropriate for our construction because the fraction distribution Γ behaves like a probability mass function. Moreover Γ represents the true fraction distribution, whereas $\tilde{\Gamma}$ is the target distribution, and we are trying to determine how close Γ is to $\tilde{\Gamma}$. For each snapshot, D is calculated as follows:

$$\begin{aligned} D\left(\Gamma\|\tilde{\Gamma}\right) &= D([\gamma_1, \gamma_2, \ldots, \gamma_n]\|[1/N, 1/N, \ldots, 1/N]) \\ &= \left(\sum_{i=1}^{N} \gamma_i \log_2 \gamma_i\right) - \log_2 N \end{aligned}$$

3.2 Results

We applied the SWM method to constant-rate WaveLAN UDP traces involving two, three and four competing stations. Fig. 10 plots the average Jain's fairness index versus the window size, and Fig. 9 plots the average Kullback Leibler fairness index. Note that for Jain's index, 1 indicates absolute fairness, whereas for the Kullback Leibler index 0 indicates absolute fairness. From both figures, we see that as expected small window sizes exhibit high unfairness, but as the window size increases it starts to exhibit more and more fairness. The horizontal lines in the two figures indicate the thresholds beyond which we deem fairness to be achieved. We set the threshold at 0.95 for Jain's index, and at 0.05 for the Kullback Leibler index. In both graphs, fairness is achieved for window sizes larger than roughly 950 in the case of two stations, roughly 250 in the case of three stations , and roughly 450 in the case of four stations.

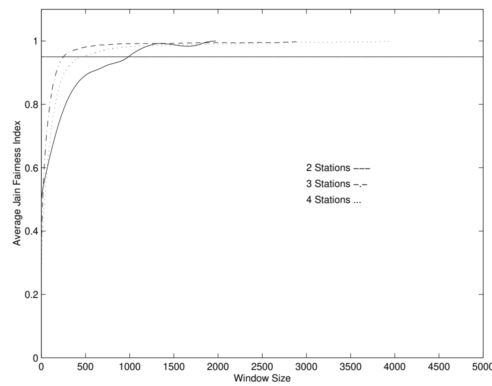

Figure 8: Applying the SWM method to CSMA/CA traces using Jain's Index as the fairness measure.

It is interesting to see that in these traces, short-term fairness for the three-station case is slightly better than the two- and four-station cases. However, the plots in Figures 10 and 9 clearly show the lack of short-term fairness for any of the three cases.

The SWM method allows the protocol designer to determine how short the short-term horizon can get before the protocol no longer exhibits short-term fairness. This critical horizon occurs at the intersection of the curve and the horizontal line in Figures 8 and 9.

Often, we can describe the algorithms used in the MAC protocol but not have access to long traces

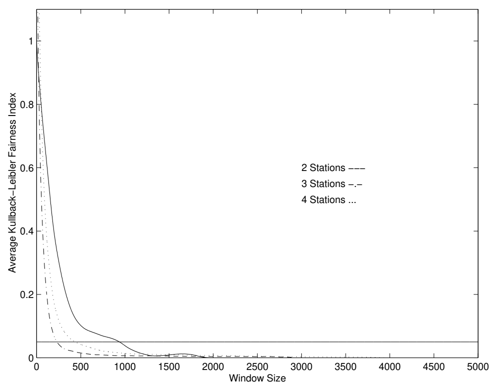

Figure 9: Applying the SWM method to CSMA/CA traces using Kullback Leibler Index as the fairness measure.

(e.g., during the design stage of a protocol). As previously mentioned, applying SWM analytically is impractical. It is therefore important to derive a practical analytical model and technique for evaluating short-term fairness, which we do using the *renewal rewards method*.

4 Renewal Rewards Method

To achieve good short-term fairness, a MAC protocol should preferentially try to allow a station to transmit proportional to how backed-off it is. Therefore, in any short-term fairness metric, a successful access to the channel by a backed-off station should be viewed more favorably than that of a less backed-off station. Given that many MAC protocols can be completely represented by Markov chains (e.g., Figures 5 and 6), it seems natural to use the theory of Markov chains with rewards to evaluate the fairness of MAC protocols. Here, each transition from State i to j in the Markov chain is associated with some reward, r_{ij}. As the chain proceeds from state to state, there is an associated sequence of rewards that are not independent, but are related by the statistics of the Markov chain. The average reward associated with a state, i, can then be calculated as:

$$r_i = \sum_{j \in \mathcal{S}} P_{ij} r_{ij}$$

where $\mathcal{S}$ is the state space and P_{ij} is the transition probability from State i to State j. The steady-state expected reward-per-stage can then be calculated as,

$$g = \sum_{i \in \mathcal{S}} \pi_i r_i$$

where π_i is the steady-state probability of being in State i.

The expected reward-per-stage, g, clearly depends on the reward structure (i.e. the r_{ij}'s) we use. By using Markov chains with different reward structures, we can calculate different parameters. The next section illustrates the calculation of two different parameters using different reward structures.

4.1 Examples

Recall that the self-transitions in the chain of a CSMA/CA system correspond to collisions. Thus, if we consider the reward structure where 0 units of reward are assigned to all the out-of-state transitions and 1 unit is assigned to self-transitions, then the resulting expected reward-per-stage gives the collision probability of the protocol.

Consider another reward structure, where a unit reward is assigned to transitions corresponding to a different station capturing the channel, and zero reward to transitions corresponding to the same station keeping the channel. The reciprocal of the expected reward-per-stage with this structure of rewards gives the average time that a station keeps the channel. We can think of this as a measure for burstiness of the service, i.e., the expected number of frames transmitted by a station upon capturing the channel, before releasing it. This measures the degree of monopoly by a station.

This idea of Markov chains with rewards may be viewed as a generalization of the fairness metric developed in [13], which investigated the use of the entropy rate of the Markov chain of the protocol to measure fairness. Notice that the entropy rate of a Markov chain can be computed using the reward structure where $r_{ij} = \log \frac{1}{P_{ij}}$. The expected reward per stage can be found as,

$$g = H(\{X_i\}) = \sum_{i \in \mathcal{S}} \pi_i \sum_{j \in \mathcal{S}} P_{ij} r_{ij}.$$

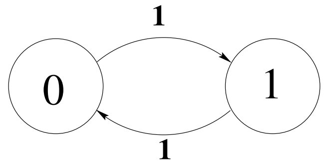

Figure 10: Markov chain of a two station TDMA system.

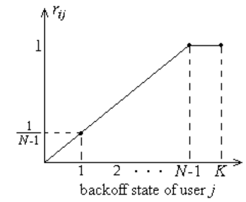

Figure 11: Linear reward function.

Unfortunately, the entropy rate is not the right fairness measure. One of the main problems with it is that it is a function of only the transition probabilities of the Markov chain of the system, and does not take important factors like the backoff states into account. To demonstrate this, consider the Markov chain corresponding to a two-station TDMA protocol (Fig. 10). The entropy rate of this chain, $H(\{X_i\}) = 0$, whereas ideal TDMA should have an index of 1.

These observations motivate us to search for a more appropriate reward structure whose resulting expected reward-per-stage gives a good measure of short-term fairness for contention-based protocols.

4.2 Assigning Rewards

4.2.1 Assumptions

We start with some assumptions and intuition about the desired reward structure.

- All the stations are assumed equal (no priorities).
- Lost frames will not be penalized by negative rewards, because fairness and throughput are treated separately.
- A reward of 0 is assigned to transitions after which the same station keeps the channel. Note that a collision leads to a self-transition, and hence 0 reward.
- The maximum expected reward is 1, achieved by an ideal TDMA protocol.

4.2.2 Reward Structure

We want the reward r_{ij} to be a function of the backoff stage of the station which captures the channel (i.e., j). This backoff stage will be represented by the number of times that the station gets backed-off, b, before capturing the channel. Note that $0 \leq b \leq K$. There are three important points to note about $r_{ij}(b)$:

1. $r_{ij}(b)$ should be non-decreasing. The more a station is backed-off, the higher reward should be assigned for a transition that causes the station to capture the channel. Furthermore, from our assumptions, $r_{ij}(0) = 0$ and $r_{ij}(K) = 1$.
2. $r_{ij}(b)$ should reach its maximum (of 1) at $b = N - 1$, which is the number of backed-off stations. To see this, consider the following two examples. If the maximum is achieved at a higher b, say N, then even with a perfectly fair protocol, maximum fairness will not be achieved. On the other hand, if the maximum is reached at a lower b, say $N-2$, then an unfair scheme like TDMA among only $N-1$ stations (with the N^{th} station never capturing the channel), will achieve the maximum expected reward of 1. Therefore, $r_{ij}(b) = 1$ for $b \geq N - 1$.
3. $r_{ij}(b)$ should be increasing for $b \leq N - 1$. In addition, we pay close attention to the second derivative of the function in this region. There are three possibilities: linear, convex or concave, as shown in Figures 11, 12, and 13 respectively.

   First consider *linear reward*. It fails to differentiate between different sequences in many cases. For instance, for a 3 station (A, B and C) system, consider the following two sequences, "AABBAAC" and "ABABABC."

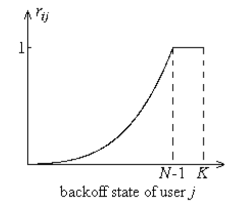

Figure 12: Convex reward function.

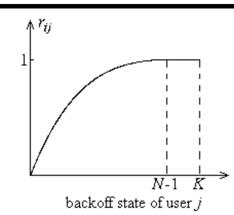

Figure 13: Concave reward function.

With linear reward structure, the rewards generated by these two sequences will be equal ($0+1+0+1+1 = \frac{1}{2}+\frac{1}{2}+\frac{1}{2}+\frac{1}{2}+1$). However, the metric for the second sequence should be greater than that of the first. Therefore, choice of linear rewards is inappropriate.

A convex reward structure makes the situation worse. In the above example, the reward corresponding to the first sequence is greater with a convex reward function because such a reward structure provides an incentive for longer backoffs. The only alternative left is a concave reward function.

Concave reward functions provide a disincentive to continual backoffs. Of course, there are a number of choices for this function. Our particular choice, chosen somewhat arbitrarily after considering many choices, is:

$$r_{ij}(b) = \begin{cases} \sqrt{\frac{b}{N-1}}\,, & 0 \le b \le N-1 \\ 1\,, & N-1 < b \le K \end{cases}$$

This function is illustrated in Fig. 14 for a system of 5 stations and a maximum of 5 backoff stages. Although other possible functions are certainly reasonable, we believe that our choice captures the essential properties of an apt function.

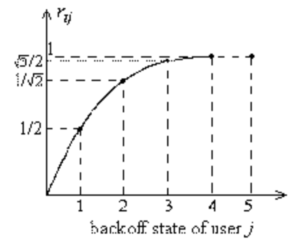

Figure 14: Concave reward function for 5 stations and 5 backoffs.

4.3 Results

In this section, we present our analytical and experimental results.

4.3.1 Analytical Results

We wrote a MATLAB program to generate Markov chains with rewards for the CSMA/CA and ALOHA[2] protocols, varying the number of stations and the number of backoff stages, and calculated the expected reward-per-stage. Fairness as a function of the number of stations is shown in Fig. 15 for CSMA/CA and Fig. 16 for ALOHA for various values of maximum number of backoff stages. For the same systems, the collision probability versus the number of stations is shown in Fig. 17 and the burstiness (the average number of frames that a station transmits upon capturing the channel) versus the number of stations is illustrated in Fig. 18. Note that the burstiness and collision probability depend only on the number of stations for the ALOHA system[3].

Short-term fairness tends to increase with the number of stations[4]. This is due to the fact that, as the number of stations increase, the number contending for the channel increases. This decreases the probability that the same station keeps the channel after a contention period. ALOHA exhibits higher fairness than CSMA/CA because in CSMA/CA, as stations get backed-off deeper,

[2] Note that in CSMA/CA frames are dropped if they cannot be transmitted after K times. For the comparison of the two schemes to be fair, we assumed that also in ALOHA, the frames which cannot be sent in K attempts are dropped.

[3] Each station is assumed to be attempting to transmit a frame with probability $\frac{1}{N}$ in each slot.

[4] Note that this statement is valid for a constant maximum number of backoffs.

their backoff window sizes increase, decreasing their probabilities of capturing the channel.

Throughput decreases as the number of stations increases. As the number of stations increases, the number contending for the channel increases which causes an increase in the number of collisions. Throughput achieved by CSMA/CA is much higher than that achieved by ALOHA. Thus, we can conclude that there is a trade-off between throughput and fairness. With increasing number of backoffs, the throughput improves at the expense of fairness.

Our burstiness measure is the average number of frames that a user transmits once it captures the channel. This measure is similar to the fairness measure because the reward structure to calculate the burstiness is set as follows.

$$r_{ij}^{\mathcal{B}}(b) = \begin{cases} 1\ , & 0 < b \leq K \\ 0\ , & b = 0 \end{cases}$$

With this reward structure, the reciprocal of the expected reward-per-stage gives us the burstiness. Thus, if a step function is assigned instead of a concave function, for $r(b)$, we get an inverse burstiness measure. Therefore, it is no surprise that fairness decreases with increasing burstiness.

To summarize these results, the short-term fairness of CSMA/CA is poor (a maximum of 0.42). It is closer to the minimum point, 0 (fairness point of the system with pure dominance of one station) than it is to the maximum point, 1 (fairness point of TDMA). ALOHA is much fairer than CSMA/CA (it achieves a fairness metric of at least 0.5). On the other hand, the collision probability of CSMA/CA is very low (a maximum of 0.04) compared to ALOHA (a minimum of 0.33). A station that captures the channel will keep it for an average of 3-4 times in CSMA/CA, whereas stations in ALOHA keep the channel for an average of at most 2 frames.

4.3.2 Experimental Results

We also evaluated the indices of fairness and burstiness for the UDP traces used to demonstrate the sliding window method in Figures 8 and 9. The results are summarized in the following table.

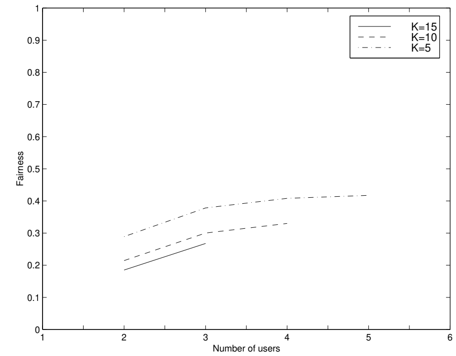

Figure 15: CSMA/CA; fairness vs. N for $K = 15, 10$, and 5.

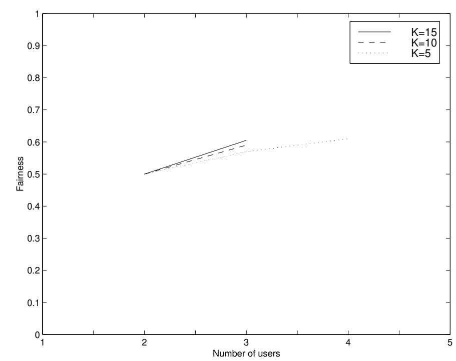

Figure 16: ALOHA; fairness vs. N for $K = 15, 10$ and, 5.

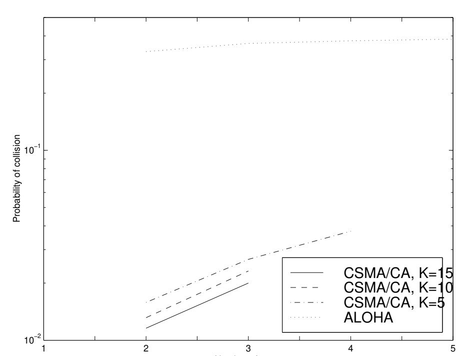

Figure 17: Collision probability versus N for CSMA/CA ($K = 15, 10$, and 5) and ALOHA.

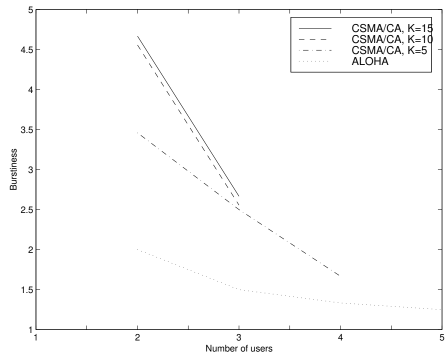

Figure 18: Burstiness versus N for CSMA/CA ($K = 15, 10$, and 5) and ALOHA.

No. of users	Fairness	Burstiness
2	0.0081	108.889
3	0.099	7.8892
4	0.1063	6.4153

The experimental results do not match those derived analytically in the previous section. We believe that there are significant implementation glitches and other bottlenecks in the measured system that account for this discrepancy. For example, errors in the sensitive timer implementations, bus bottlenecks, etc. would exacerbate the deviation of analysis from experiment. Nevertheless, the concave reward structure does detect the fact that the short-term fairness does increase with increasing number of users, using the data taken from a real system.

5 Conclusions

In this paper, we investigated the problem of short-term fairness in decentralized media-access protocols. We devised two methods to evaluate this—the sliding window method and the renewal rewards method. Both methods can be used to analyze experimental packet traces that show channel accesses by different stations for a known source workload. The rewards method, which uses a Markov chain with an appropriate reward function for transitions, is analytically tractable. Because it analyzes the protocol itself, as opposed to deductions of the protocol made from traces, it provides insight into why protocols exhibit bad short-term fairness.

We performed the Markov analysis for CSMA/CA and ALOHA protocols, finding that there is a fundamental trade-off between short term fairness and system throughput. CSMA/CA has significantly better throughput and somewhat worse fairness compared to ALOHA. Short-term fairness does increase, whereas throughput decreases with increasing number of stations in CSMA/CA.

It will be interesting to apply the fairness metrics developed here to promising MAC protocols such MACAW [4]. MACAW aims to achieve fairness by having all the contending stations backoff to identical stages when collisions occur, based on the backoff stage of the "winning" transmitter. Intuitively, it seems like MACAW ought to perform as well as ALOHA in terms of short-term fairness (and have better throughput), but a complete validation of this is needed.

Short-term fairness is important in several contexts, e.g., smooth acknowledgment flow for TCP connections and low jitter for real-time audio and video; we therefore hope that these measures will be used by MAC protocol designers in conjunction with traditional performance measures such as the collision probability to evaluate overall protocol performance. Finally, we believe that the notion of short-term fairness is important in other networking contexts in addition to MAC protocols. While we do not claim that the same measures as in this paper will be generally applicable, we believe that system and protocol designers should include short-term behavior in performance characterizations.

References

[1] Abramson, N. The ALOHA system – Another Alternative for Computer Communications. In *Fall Joint Computer Conference, AFIPS Conference Proceedings* (1970), vol. 37, pp. 281–285.

[2] Aldous, D. J. Ultimate Instability of Exponential Back-Off Protocol for Acknowledgment-Based Transmission Control of Random Access Communication Channels. *IEEE Trans. on Information Theory IT-33* (1987), 219–223.

[3] Bertsekas, D. and Gallager, R. *Data Networks*. Prentice Hall, 1992.

[4] Bharghavan, V., Demers, A., Shenker, S., and Zhang, L. MACAW: A Media-Access Protocol for Packet Radio. In *Proc. ACM SIGCOMM* (1994).

[5] Chiu, D.-M., and Jain, R. Analysis of the Increase and Decrease Algorithms for Congestion Avoidance in Computer Networks. *Computer Networks and ISDN Systems 17* (1989), 1–14.

[6] Cover, T. and Thomas, J. *Elements of Information Theory.* John Wiley & Sons, 1991.

[7] Deng, S. Capture Effect in Residential Ethernet LAN. In *Proc. IEEE GLOBECOM* (Nov. 1995).

[8] Gafni, E., and Bertsekas, D. Dynamic Control of Seesion Input Rates in Communication Networks. *IEEE Trans. on Automatic Control AC-29* (1984), 1009–1016.

[9] Jain, R. *The Art of Computer Systems Performance Analysis.* John Wiley and Sons, 1991.

[10] Karn, P. MACA – A New Channel Access Method for Packet Radio. In *Proc. 9th ARRL Computer Networking Conference* (1990).

[11] Kleinrock, L., and Tobagi, F. A. Packet Switching in Radio Channels: Part 1: CSMA Modes and Their Throughput-Delay Characteristics. *IEEE Trans. on Comm. COM-23* (1975), 1400–1416.

[12] Ramakrishnan, K. The Ethernet Capture Effect: Analysis and Solution. In *Proc. 19th Conference on Local Computer Networks* (Oct. 1994).

[13] Raman, S., and Swan, A. A Study of Short-Term Fairness in Wireless MAC Protocols. CS 294-7 Class Project Report, University of California, Berkeley, May 1998.

[14] The WaveLAN Home Page. http://www.wavelan.com, 1998.

[15] Xylomenos, G. and Polyzos, G. TCP and UDP Performance over a Wireless LAN. In *Proc. IEEE INFOCOM* (1999).

[16] Zhang, L., Shenker, S., and Clark, D. D. Observations and Dynamics of a Congestion Control Algorithm: The Effects of Two-Way Traffic. In *Proc. ACM SIGCOMM '91* (1991), pp. 133–147.