

Overload Protection for IEEE 802.11 Cells

Hector Velayos, Ignacio Más, *Members, IEEE*, and Gunnar Karlsson, *Senior Member, IEEE*

Abstract— This paper presents a distributed admission control for the distributed coordination function of IEEE 802.11 wireless LANs that limits the risk of congestion collapse due to a high arrival rate of flows. This control scheme requires no modification to the current distributed coordination function; it works by performing a short, non-disturbing probe that estimates the MAC service time. The flow is admitted if the estimate is below a threshold. We show how the threshold may be adjusted dynamically to maintain an average packet loss rate below a configurable limit. We show via extensive simulations that the admission control avoids congestion due to flow arrivals and that it maintains the loss probability below the given threshold regardless of the offered load or number of stations. Our simulations also analyze the main drawback of our scheme: a reduction in the link utilization. The admission control efficiently protects cells from overload and it may offer soft QoS guarantees to multimedia flows without the need for scheduling or polling mechanisms in IEEE 802.11.

Index Terms—Wireless LAN, admission control, probing, overload protection, 802.11 MAC service time

I. INTRODUCTION

WIRELESS LANs based on the IEEE Std. 802.11 suit the predominant choice for high-speed wireless access to the Internet. Much of the deployment of wireless LANs is rather spontaneous, driven by many different parties' incentives to provide wireless access at specific locations. A general problem for an operator of a WLAN is to plan where to place access points to provide both sufficient coverage and capacity for the users of the network. It is often not feasible (or within the operator's competence) to forecast traffic demands and user-mobility patterns in order to dimension the WLAN appropriately to provide a desired grade of service. The cells in the network are therefore susceptible to either under or overload. We have in past work addressed this problem by suggesting an iterative deployment of cells in order to build and extend WLANs and to cope with mismatches between offered load and available capacity [1]. For instance, load-balancing increases the network throughput and reduces the cell delay for overlapping access points [2]. We have also presented how to reduce the handoff time by early detection and fast active scanning, which is essential to make the load balancing non-disruptive to ongoing sessions [3]. The current work adds another component to this

architecture, namely overload protection of cells to avoid congestion collapse when the infrastructure is under-dimensioned with respect to the offered traffic. The main reason is due to the unpredictable mobility; for example, the arrival of a flash crowd could cause a cell to collapse under congestion.

The IEEE 802.11 medium access control protocol was designed as an asynchronous best-effort service, and as such it grants each station the same number of transmission opportunities in the long term. This means that the throughput per station goes down with an increasing number of stations; it could result in queue buildup and eventually loss of frames in the stations and the access point. Flow admission control has long been advocated as a means to reduce overload situations in networks instead of relying on users' impatience [4]. We have therefore developed a distributed admission control to block stations from initiating new sessions if that would cause overload. Stations that are blocked will back off before re-attempting to start the session, or they might disassociate themselves from the cell and search for access to an alternative, less loaded cell. Persistent overload that leads to a constantly high blocking probability is only resolved by deploying more access points.

Two access schemes were specified in the IEEE Std. 802.11 [5]. The point coordination function (PCF) features contention free access and therefore is suited to provide QoS guarantees. The PCF is a centralized scheme, in which the access point grants transmission permission to the stations that request it. It has never been commercially available, but it has originated some recent work on different polling schemes [6][7], or call admission control mechanisms [8]. Nevertheless, the dominant access scheme is the distributed coordination function (DCF), which features contention access whereby offering greater flexibility compared to the PCF. In favor of the DCF, previous work has shown that the PCF mode can have poor performance both alone and in cooperation with the DCF [9]. A shortcoming with the DCF is the lack of QoS support now when the interest in supporting multimedia services in WLAN is steadily increasing. Current standardization work by the IEEE 802.11e task group is addressing the modifications to the DCF in order to support priority schemes that are suitable for multimedia transmissions [10][11]. The approach is to provide delay sensitive traffic priority access to the channel by differentiating the DCF parameters. It does not protect against overloading the cell, however, and our scheme can be used alone or as a complement to the draft supplement of IEEE 802.11e for preventing congestion in a cell.

In this paper, we present the distributed admission control for the standard DCF together with an evaluation of its

Manuscript received April 28, 2006. This work has been partly supported by the European Union under the E-Next Project FP6-506869 and KTH Graduate School of Telecommunications. Corresponding author's email: gk@kth.se.

All authors are with the Laboratory for Communication Networks at KTH, the Royal Institute of Technology, Stockholm, 100 44, Sweden.

performance. The control scheme prevents overload of the cell and limits thereby both transmission delay and packet loss. It works by performing a short, non-disturbing probe, which estimates what the average MAC service time would be if the new flow enters the cell. This average service time is then compared to an admission threshold. The threshold might be dynamically calculated at the access point and broadcasted in the beacons to limit the loss rate in the cell, and not only the service time. A new flow is admitted if its MAC service time during the probing is below the threshold. The protection scheme offers a trade-off between cell utilization on one hand and delay and packet loss due to congestion, on the other hand. Our evaluation addresses this tradeoff as well as the protection that the scheme offers. The scheme also works in case the AP belongs to an operator who is not interested in admission control. It provides a test for arriving users on whether the AP offers enough capacity for their connection. Only in the case the test is positive will the user go on with establishing the connection. In this mode, the service time threshold is statically configured in the client instead of being broadcasted in the cell beacon. This is a form of self-admission control that protects the user from making an association to an AP that cannot provide necessary resources (for self-admission control see [12]).

To the best of our knowledge, our work is the first to suggest distributed probing for admission control in IEEE 802.11 wireless LANs. There has been more work on probing for admission control in wired networks, for instance [13][14][15]; however, the work is not directly applicable to the distributed queuing system such as the DCF in the 802.11 MAC. Admission control for WLAN has been suggested in the literature, for instance in [16], but none of the proposals are distributed and they might require modifications of the coordination function.

Limitations of this work are as follows. *i)* It assumes a stable radio channel and does not consider effects like fading, multi-path propagation or interference from neighboring cells, which could increase the losses. *ii)* All stations and the access point are identical with respect to transmission rate and traffic characteristics. *iii)* The transmission rates are constant and do not change due to mobility.

The remainder of the paper is organized as follows. Section II presents the MAC service time in WLANs. Section III describes the admission procedure based on estimation of the MAC service time by probing. Section IV shows the evaluation via simulations of the admission control. Section V contains the future work items. Section VI summarizes the main findings.

II. THE MAC SERVICE TIME

The delay of a packet crossing a wireless LAN cell can be split into three parts: the delay at the medium access control (MAC), the transmission delay and the propagation delay. The propagation delay can be neglected due to the small size of the cells. The transmission delay can be easily determined from the packet size and the bit rate used. The MAC delay is the most difficult to calculate because it depends on the traffic

sent by the other stations in the cell. We divide the MAC delay into two parts: the service time and the queuing time. The service time is the time to gain access to the shared channel for the transmission of the packet following the rules specified in the IEEE Std. 802.11. The queuing time is the time spent in the queue waiting for earlier packets to be transmitted.

We analyze in this section the service time for the IEEE 802.11 MAC protocol. First, we describe the channel access procedure and then provide the analysis based on simulations. We identify the number of stations competing for the channel as an important factor that affects the service time and show that the service time may vary several orders of magnitude between consecutive packets sent from the same station. We also determine the average MAC service time and its standard deviation for different cell loads and numbers of stations. We will use this information to design the admission control.

A. Channel access procedure

The channel access procedure is defined in the IEEE Std. 802.11 [5] and is common to all supplements such as Std. 802.11a, b and g. As justified in the introduction, we focus our work on the DCF, one of the two channel access functions described in the standard. The DCF is a distributed access scheme in which all stations, including the access point, execute the same procedure to compete for the channel. The station that gains the channel transmits a single packet and enters the competition again if it has more packets to transmit.

The competition for the channel in the DCF works as follows. A station that has a packet ready to transmit senses the channel for one interval called the distributed inter-frame space (DIFS). The station may send the packet if the channel is idle during the entire interval; otherwise it is backlogged. All backlogged stations choose a random number called backoff from an interval called the congestion window. The initial congestion window goes from 0 to 31. The backoff represents the number of time slots that the station must sense the channel to be idle before it can start transmitting. The length of a slot depends on the physical layer (e.g. a slot is 20 μ s in IEEE 802.11b and 9 μ s in IEEE 802.11a). Since different stations are likely to choose different random numbers, this scheme is often collision free. However, two stations may choose the same backoff and a collision would occur. The chances of this event increase with the number of stations, but it is rare and hardly impacts the overall performance except for large numbers of stations [17]. When it occurs, it is solved via retransmission at the MAC layer.

There are some rules for decrementing slots from the backoff. First, the station must sense the channel idle during a complete slot. If a transmission starts during the slot, the whole slot is considered used. Second, a packet transmission includes some periods in which the channel is idle. No slots can be decremented during these periods. For instance, the channel is idle during some microseconds between the data packet transmission and its acknowledgment. This time cannot be used to decrement slots. Third, after a transmission of a frame is completed (including the acknowledgment) the station must sense the channel to be idle during a full DIFS

before decrementing new slots.

A station can transmit when its backoff reaches zero. In the basic access, the station transmits a single packet and waits for the acknowledgment to confirm its reception. If Request to Send/Clear to Send (RTS/CTS) signaling is enabled, the station transmits a RTS frame to notify the rest of the stations that it is about to start a transmission. The access point confirms the RTS reception by sending a CTS frame. When the CTS frame reaches the station, it transmits the data frame and waits for the acknowledgement. The RTS/CTS signaling is more efficient than sending a long data frame when collisions are likely.

The station retransmits the frame if the acknowledgment is not received. On each retransmission, the same access procedure is used with the exception that the congestion window is doubled. A packet can be retransmitted a maximum of 7 times, although the congestion window is only doubled the first 5 times. The side effect of doubling the congestion window interval is that stations retransmitting gain the channel less often than stations transmitting a packet for the first time. This is called the capture effect and affects the fairness of the access scheme when some stations often need to retransmit due to poor radio conditions. Nevertheless, it may increase the link throughput because stations with good channel conditions tend to transmit more often.

The DCF is fair in the sense that each station receives the same number of transmission opportunities in the long term. It is noteworthy that the access point behaves like any station and does not get more transmission opportunities; it is the bottleneck for symmetric bi-directional flows from the stations [18]. If all stations use similar bit rates and experience comparable number of retransmissions, then they achieve the same throughput. This also implies that each station's throughput decreases in equal proportion when the offered traffic exceeds the cell's capacity. Our admission control addresses this problem by limiting the offered traffic so that congestion is avoided.

B. Analysis of the service time

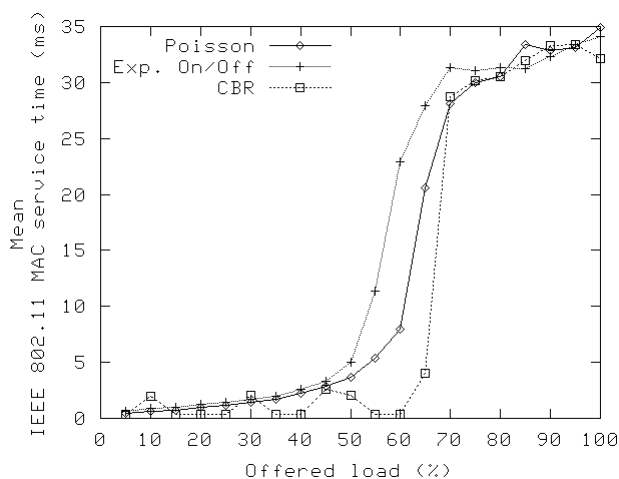


Fig. 1. Comparison of the mean MAC service time for different traffic patterns and loads.

The growing interest in quality of service has motivated the publication of models for the delay in the IEEE 802.11 MAC protocol. The analysis of MAC service time is a key part of these models. A common assumption for early models is that stations are always ready to transmit (saturation). There are several examples of such models. Chatzimisios et al. studied the packet delay in presence of transmission errors [19]. Their model is an evolution of Bianchi's model based on Markov chains for throughput analysis in ideal channel conditions [17]. Tay and Chua suggested a different model based on stochastic analysis that provides throughput and packet delay [20]. Carvalho and Garcia presented another model for the MAC delay as a function of the channel state probabilities [21]. Unfortunately, these probabilities can only be calculated under the assumption of saturation. All these models provide mean service time in saturation; they cannot be used for non-saturating loads. Therefore they are not suitable for designing an admission function that would prevent saturation. Nevertheless, their output values can be used as upper bounds for the service time. Banchs suggested an approximated expression for the distribution of the backoff delay (equivalent to the service time) in saturation [22]. His work permits a better understanding of the service time in saturation compared to previous models that only provided the mean service time. Again, it is only valid for saturation.

The saturation condition was relaxed in two models published recently. Tickoo and Sikdar presented a queuing model for the average service time valid for non-saturating loads and arbitrary arrival patterns [23]. Their model determines the mean service time from average inter-arrival times for traffic sources. Li and Battiti suggested another model for non-saturation in which the mean service time can be derived from the probability that a station's transmission queue is empty after the successful transmission of a packet [24]. An assumption in both models is that the number of stations in the cell is large enough so that the probability of packets colliding is constant and independent of the transmission time. The number of stations is however small on

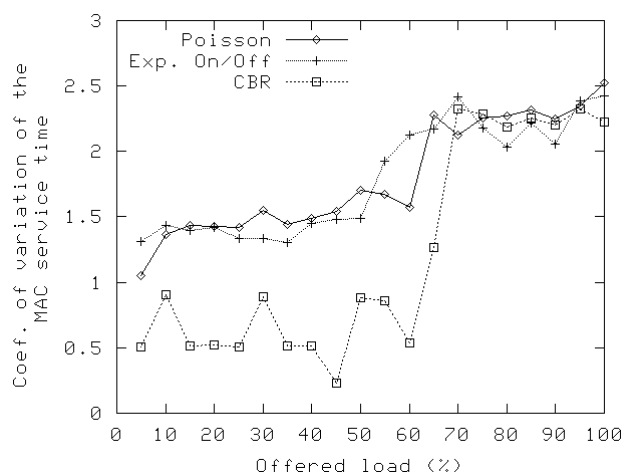


Fig. 2. Comparison of the coefficient of variation for different traffic patterns.

most operational WLANs. Hegde et al. provide an expression

for the service average service time for an arbitrary number of users and for non-saturated nodes for bi-directional voice calls of constant bit-rate [18], and Ergen and Varaiya provide a throughput model for unsaturated cells with arbitrary many users, and it includes physical layer for 802.11a [16]. An alternative to the analytical models is measuring the service time with real equipment. However this is not possible because the packet timestamps in commercial WLAN cards are not accurate enough and include other delays not related to the MAC protocol operation.

Since we are interested in the service time of each packet in a flow for bursty traffic sources, we simulate the access link to get these results. We present a statistical analysis of the IEEE 802.11 MAC service time from the simulation results. Our analysis extends the mathematical models providing packet level information for small number of stations. We have used the Network Simulator 2¹ (ns-2) for our simulations. The only modification was the addition of a monitoring agent to measure the MAC service time. A single cell is simulated where each station has the same contribution to the cell load and all stations transmit at the same bit rate. The cell load is measured as the addition of the traffic generated by the stations normalized to the bit rate of the transmissions. The cell load ranges from 5% up to 100%, increasing in steps of 5%. The number of stations in the cell n is a parameter in our simulations.

In addition to the cell load, the traffic pattern of the sources affects the MAC service time. We have compared sources generating traffic according to three different patterns: constant bit rate (CBR), exponentially distributed on/off with 20 ms average on-time and 35 ms average off-time, and Poisson inter-arrivals. Fig. 1 shows the mean MAC service time for the three source types. The service time exhibits a state change for all source types. Numerical values are similar for the different sources for loads below 50% and in saturation (above 65% load). A small difference occurs as the load approaches saturation because the sources have slightly different saturation values. Fig. 2 shows the coefficient of variation of the service time for the three source types. As expected, the variability of the CBR source in non-saturation is the lowest, while the variability of the other two sources is similar. All sources show the same variability in saturation. Based on these results, we have selected exponential on/off sources for further statistical analysis. Their high variability and highest mean service time close to saturation make them good benchmark sources.

Hence, each station generates bursty traffic according to an exponentially distributed on-off traffic source. As stated above, the average on-time t_{on} is 20 ms and the average off-time t_{off} is 35 ms. If the desired cell load is l , the rate of each station during the on period is calculated as follows:

$$rate = \frac{(t_{on} + t_{off})}{t_{on}} \frac{l}{n}. \quad (1)$$

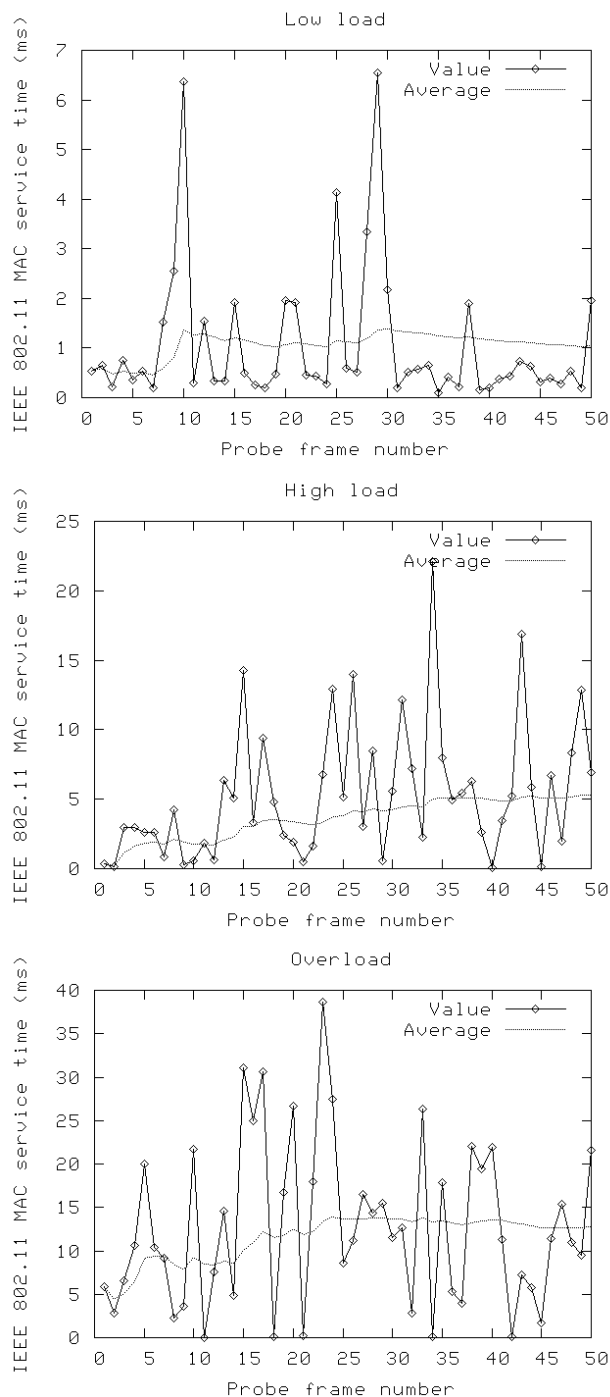


Fig. 3. IEEE 802.11 MAC service time per packet for different cell loads.

During the simulations, $n - 1$ stations are generating traffic and we measure the service time of the n -th station when it starts its transmission. The results reported in the next section correspond to the measurement of the first 50 packets per simulation run. We repeat the simulation with different random seeds 30 times to calculate the average service time per packet. All simulations use a fixed packet size equivalent to the default TCP segment size (approximately 500 bytes),

1. <http://www.isi.edu/nsnam/ns/>

and there are no retransmissions due to poor radio conditions.

The following figures present two types of results for the MAC service time. First, we show the service time per packet for different load levels. Then we show the average service time and its standard deviation for different load levels using the number of stations as a parameter.

Fig. 3 shows the MAC service time per packet in a cell with 10 stations for three load levels representative of the different cell status. We have selected *low* load to denote an offered load equivalent to 15% of the channel capacity, *high* load to 60% of the channel capacity, and *overload* to 90% of the channel capacity. The cell operates without losses only at low and high loads. In overload, the traffic in excess is dropped. The subplots in Fig. 3 show the service time per packet as well as the running average, i.e. the average of the service time including all packets previously transmitted; it is updated after each packet transmission.

We can draw the following conclusions from Fig. 3. There are always packets whose service time is below 1 ms regardless of the offered load. For these packets, the station can decrement all the slots in the congestion windows without interruptions due to transmission from other stations. This case is always possible, although its probability decreases when the load increases. This is confirmed comparing the subplots in Fig. 3. The number of occasions in which the service time is below 1ms is much higher in the first subplot than in the last one.

The maximum service time for a packet depends on the cell load. The higher the load, the higher the service time could be. The peaks shown for the service time correspond to cases in which the congestion window decrement is interrupted by transmissions from other stations. The probability of this event grows with the load. The difference between peaks in the same subplot is related to how many packets interrupt the slot count down.

Despite the differences in service time between packets, the running average stabilizes quickly; around 20 packets are enough to get an accurate value. Nevertheless, the average is a poor indicator of the service time for individual packets as shown in the subplots. Fig. 5 shows the mean service time for different load levels and its standard deviation using the number of stations as a parameter. Thirty independent simulation runs were used to compute the mean for each number of stations. Both subplots in Fig. 5 indicate that the number of stations starts to affect the service time significantly only when the offered load is above 60%. If the offered load is below 60%, the mean service time is small, but the standard deviation is of the same order of magnitude as the mean. This indicates that the service time of the packets is widely spread around the mean, with some packets experiencing service times in the order of tens of milliseconds. An example of such packets is visible in the middle subplot of Fig. 3. Therefore, the service time cannot be neglected for some applications with limits on the end-to-end delay for every packet. A typical application of this type is voice over IP that requires a maximum of 150 ms end-to-end delay for every packet. A MAC service time of around 20 ms for some of the packets

may be excessive and might lead the receiver to consider them as losses.

There is no correlation between the service times of consecutive packets. This was expected from the access channel procedure since each packet is transmitted independently of any other using the same rules. Fig. 3 intuitively shows the lack of correlation. Fig. 4 shows the result of the analysis of autocorrelation as described in [25]. It presents the autocorrelation in high load, plotted for lags of 1 to 100. The autocorrelation plot contains three horizontal reference lines. The middle one is at zero. The other two are 99% confidence bounds. Since almost all autocorrelation values fall within the 99% confidence limits and there is no visible pattern in the plot, the data is random; there is no significant autocorrelation. Therefore, the service time of a packet cannot be calculated from the service time of previous packets. We have published a deeper analysis of the service time in [26].

III. ADMISSION CONTROL

The purpose of our admission control is to determine whether a new flow can be accepted to the cell without causing overload for the existing flows (i.e. increasing their delay and loss due to congestion above a given target). The decision is taken at the station originating the new flow, since we aim at a distributed scheme without signaling to the access point or to other stations in the cell. There are various possibilities to take an admission decision.

One possibility is that the station with a new flow tries to estimate the cell's load by listening to the channel. Unfortunately, the station may not receive all transmissions in the cell. Therefore, this method would likely result in an under estimation of the cell's load. Another possibility is that the station starts the new flow and monitors its packet losses. If the losses exceed a certain limit, the flow should be stopped. Although this scheme would work, existing flows in the cell are likely to suffer similar losses since the MAC protocol gives the same number of transmission opportunities to each station. The new flow would simply be blocked too late. Instead, we propose that the station with the new flow

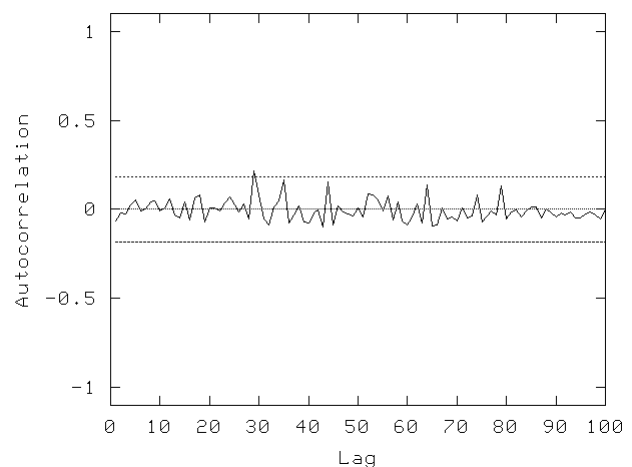


Fig. 4. Autocorrelation of the MAC service time for high load.

performs an active measurement to determine if the new flow can be accommodated. The main goal of our probing scheme is to perform these measurements in a way that gives enough information to perform an accurate admission decision, but does not produce significant disturbance to the ongoing traffic.

A. The admission process

Our admission decision is based on the relation between the mean MAC service time and the offered load shown in Fig. 5. We limit the offered load by limiting the mean service time. The service time is in a sense an aggregate measure of the number of stations, as well as their transmission bit rates and offered load (and the transmission bit rate is in turn a result of all physical layer conditions that a station experiences). We would also like the admission control to limit the packet loss in the stations so that the maximum load permitted in the cell is given by a target packet-loss probability we want to enforce. The only information the station requires in order to perform the admission decision is the admission threshold for the service time during the probing. We suggest that the access point includes this value together with the other cell information in the periodically broadcast beacon and that a default value is statically configured in the mobile client for a completely AP independent admission control.

The first step for a station is to measure the service time by probing the channel. The procedure works as follows: When a station wants to start transmitting a new flow, it sends 50 probe packets with a probe packet size of 500 bytes. Ideally, the probe size should be equal to the average packet size of the flow, but this information is unknown at the time of the probing. Instead, we use a fixed size that is close to the average packet size on the Internet backbones². The probe packets are generated at the MAC control function as part of the access scheme and can contain randomly generated data or signaling information to setup the flow end-to-end. There are two cases for the probing that relate to streaming and elastic flows, respectively. The probe rate is constant and equal to the peak rate of the flow for a streaming flow. This corresponds to the maximum load that the flow would be able to impose on the cell. The choice for an elastic flow is less distinct. We suggest sending the probe packets consecutively, i.e. the packets are placed in the access buffer and sent out as fast as possible. This choice corresponds to the maximum load that a TCP flow would impose on the cell, given that the access queue drops packets from its tail. The evaluation in this paper is limited to the streaming case.

The train of probe packets interacts with the ongoing flows, thus experiencing a service time similar to the service time of the new flow if it would be accepted. The station measures the service time of each probe packet and determines the mean service time of the probes. If Request-To-Send/Clear-To-Send signaling (RTS/CTS) is used, the RTS service time should be measured instead. The obtained mean is compared to the admission threshold received in the last beacon. If the mean is smaller than the admission threshold, the new flow is

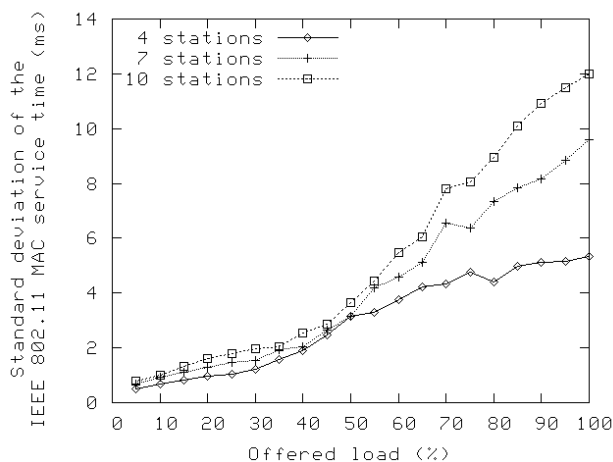
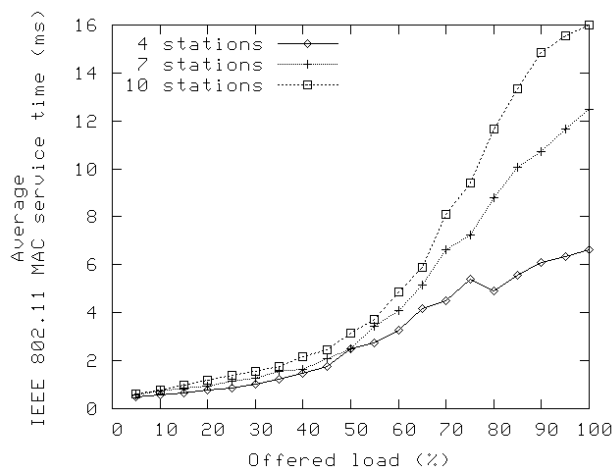


Fig. 5. Mean and standard deviation of the IEEE 802.11 MAC service time as a function of the cell load

accepted. Otherwise, the flow is rejected and has to back off for a certain time before performing a new attempt to enter the cell. The function that takes the admission decision is located in the MAC control function, where it can obtain the beacon information and may control the queuing in the interface to compute the MAC service time. Streaming flows require a second condition to be met: The rate achieved by the probe packets has to equal the peak rate of the flow, i.e. there should not be any queue buildup during the probing.

The effect of the probing phase on the ongoing sessions depends on the load in the cell and the number of stations. The evaluation of our admission control, presented in Section IV, shows that the probing does not significantly affect ongoing sessions, and yet it permits making correct admission decisions. The time required to complete the probing depends on the flow's peak rate. For instance, a flow with 1 Mb/s of peak rate would complete the probing with 500 byte packets in 200 ms if successful (probes facing high load take longer but will lead to rejection), while a typical VoIP application transmitting at 50 kb/s would take roughly 0.5 s to probe successfully with a typical VoIP packet size of 60 bytes. Most streaming flows have substantially longer durations and the overhead incurred by the probing is then low. The probe

2. Sprint IP monitoring project, <http://ipmon.sprint.com>

packets may contain user data or signaling information and hence the overhead is reduced, for instance with respect to short TCP flows. We discuss this aspect in the Section V on future work.

B. Finding the admission threshold

The limit on the service time, enforced by the admission control, is sufficient to prevent overload in the cell. For this, there could be a static limit in the cell. However, both streaming and elastic flows are usually sensitive to packet loss. It is possible to limit the packet loss probability by adjusting the admission threshold dynamically, as explained in this section. For this purpose we select the admission threshold to be the mean service time for which the mean packet-loss probability in the cell is equal to the desired target. The access point determines this threshold from the number of stations in the cell and the desired upper bound for the average packet loss probability. The number of stations is always available at the access point because it grants connection authorizations, and the bound for the loss probability must be provided as a configuration parameter. The values are updated dynamically and are broadcast in the beacons.

The admission threshold is calculated in two steps. In the first step, we determine the offered load that corresponds to the target loss probability. From our simulations, we have obtained the relation between mean packet-loss probability in the cell and offered load. This relation is shown in Fig. 7 for 4 and 10 stations. Each point of the plot was obtained by averaging the mean packet-loss probability for 30 simulation runs with different random seeds. The data in Fig. 7 show that there is no packet loss when the offered load is lower than 50% and that there is an offered load value after which the packet-loss probability increases roughly linearly. This value depends on the number of stations in the cell. The higher the number of stations, the lower the value is, because congestion is reached earlier. The slope at which the loss probability increases also depends on the number of stations. The smaller the number of stations, the higher the slope because there is less total buffer space in a smaller group of stations (each station has a fix buffer size of 50 packets). The linear behavior

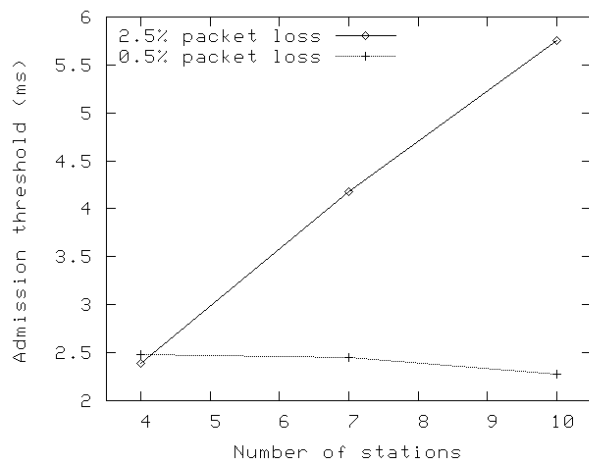


Fig. 6. Admission threshold as a function of the number of stations in the cell for 0.5% and 2.5% packet loss probability.

of the loss probability function permits us to determine the offered load corresponding to the target loss probability. For example, if we select a target loss probability of 2.5%, we obtain from Fig. 7 that the maximum offered load in the cell should be approximately 72% and 79% of the link capacity for 4 and 10 stations, respectively.

In the second step to calculate the admission threshold, we determine the service time that corresponds to an offered load equivalent to, or slightly lower than, the value obtained in the previous step. This service time value is the dynamically updated admission threshold that is announced in the beacons from the access point. Fig. 5 shows the mean service time as a function of the offered load. The obvious choice would be to take the mean service time that corresponds to the maximum offered load calculated in step 1. But the mean service time is not appropriate due to its high standard deviation. We rather choose the minimum mean value obtained during the 30 independent simulation runs that produced the mean service time shown in Fig. 5. This conservative option produces a lower admission threshold. We show in the evaluation below that this choice results in higher unwarranted blocking, it but virtually eliminates the wrong admissions.

The two-step procedure allows calculating the admission threshold for the desired target loss probability as a function of the number of stations. This function could be pre-computed and tabulated in the access point. All our figures refer to 500 bytes packet sizes for the flows, but the function could easily be computed for different packet sizes and tabulated for the most representative packet sizes in the cell to accommodate flows probing with actual data packets. An example of the result is shown in Fig. 6 for two target loss probabilities: 0.5% and 2.5%. For instance, the admission threshold would be 4.25 ms for the target packet loss probability of 2.5% if there are six stations in the cell. The threshold for 2.5% packet loss increases with the number of stations, while the number of stations does not affect the threshold for 0.5%. The first threshold corresponds to a loss probability high enough to permit a light congestion; therefore, some packets wait until the transmission of packets from other stations is completed. It is natural that the maximum service time (i.e. the admissible threshold) increases with the number of stations. Ten stations have more buffer space than 4 stations to store packets while the first in the queue is transmitted. The second threshold completely prevents the congestion so the maximum service time is independent of the number of stations. The offered load is low enough so that most of the packets are transmitted as soon as they reach the link layer. They do not have to wait in queues for other stations' transmissions to complete.

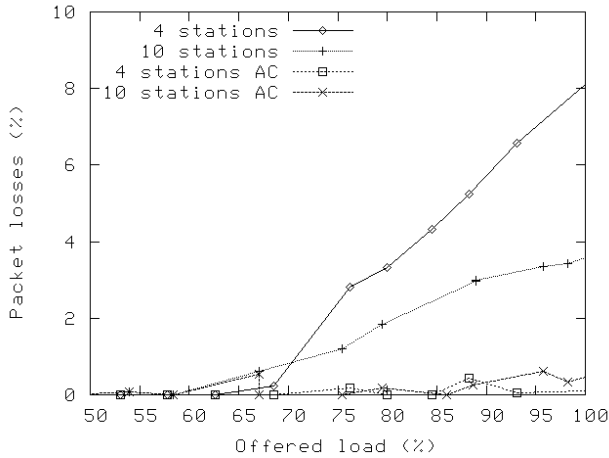


Fig. 7. Packet loss probability for 4 and 10 stations as a function of the offered load with and without admission control.

IV. EVALUATION

We have implemented our admission control in ns-2 to evaluate its behavior and performance. There were three goals for our evaluation: First, we wanted to show that our admission control effectively keeps the packet-loss probability below the target; second, we desired to compare the achieved link utilization with and without admission control; and third, we wanted to analyze the correctness of the decisions taken by the admission control.

Our evaluation is based on simulations. We study the different aspects of the admission control as a function of the offered load. The offered load is expressed as a percentage of the link capacity and varies from 5% to 100% in steps of 5%. For each offered load value, the simulation is run 30 times with different random seeds. The results in the plot are the mean of the 30 repetitions. The number of stations is used as parameter. Two values are shown: four and ten stations. In all simulations, we have used the same type of bursty sources described in the analysis of the MAC service time in Section II. The target packet-loss probability was set to 2.5% for all

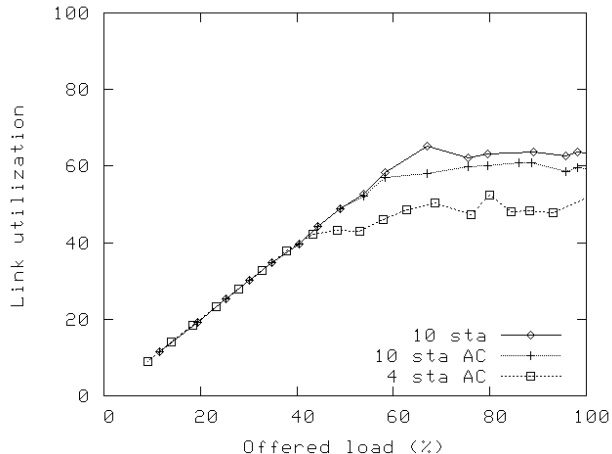


Fig. 8. IEEE 802.11 cell utilization as a function of the offered load with and without admission control.

simulations. Therefore, the admission thresholds given in Fig. 6 are used when admission control is enabled.

Fig. 7 shows that our admission control keeps the mean losses due to congestion below the target. It depicts the mean packet loss probability in the cell for different offered loads, with and without admission control, for 4 and 10 stations. The admission control is set to keep the packet-loss probability below 2.5%. The figure clearly shows that the packet loss probability grows almost linearly from an offered load of around 60% of the link capacity when there is no admission control. When our admission control is enabled, the packet-loss probability remains well below our target packet-loss probability of 2.5%, regardless of the offered load or number of stations.

Fig. 8 shows the link utilization with and without admission control for different offered loads. The utilization for 4 stations without admission control was omitted for clarity since it virtually overlaps with the utilization for 10 stations without admission control. It can be seen that there is a slight decrease in utilization at high offered loads when the admission control is enabled. This is due to the fact that some sessions are blocked to avoid the losses due to congestion. Lower link utilization is the main penalty of our admission control.

Another important observation is the effect that the number of stations has on the achieved utilization. The lower the number of stations, the lower the link utilization is. This effect is consequence of our simulated scenario in which each station generates the same amount of traffic. This means that blocking one out of four flows represents a higher reduction in offered traffic than blocking one out of 10 flows. This effect is then more related to the size of the sessions being setup than to the number of stations in the cell.

Finally, we show the performance of the admission control in 0 by classifying the admission decisions into three subsets: correct decisions, wrong decisions and unnecessary blocking. Wrong decisions are those in which the admission control fails to bound the packet loss ratio in the cell, i.e. after admitting a new flow the cell experiences a packet loss probability above

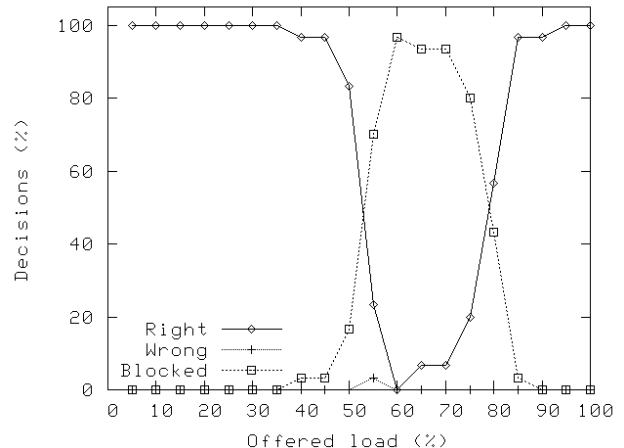


Fig. 9. Admission control performance as percentage of correct, wrong and unnecessary blocking decisions as a function of the cell's load

2.5%. Unnecessary blocking occurs whenever a new flow is not admitted, although the loss probability in the cell would not have grown over our target value had it been admitted. As a design decision, we wanted to minimize the number of wrong decisions, risking to increase the unnecessary blocking and thereby to decrease the utilization. The figure shows that our admission threshold meets our design choice. There are a negligible number of wrong decisions while the unnecessary blocking is only high near the theoretical limit of the load that should be accepted. Around 60% of offered load, new flows are rejected while some of them could have been accepted and still meet our packet loss criteria. However, most of the flows are correctly rejected once the offered load increases over 80%. A less restrictive admission control would achieve higher utilization, but would have risked admitting sessions that would suffer a higher packet loss probability than our target.

V. FUTURE WORK

There are a number of issues of our admission control for the DCF that needs further study. These are the more important:

We have fixed the probing length to 50 packets since it produced mean service time measurements accurate enough for our purposes. A deeper analysis of the required probing length may result in a reduction of the probing phase.

Our admission control is flow based. If a station sends several flows, the admission control is applied to each one independently. Our simulations focused in the case with one flow per station. Further simulations should validate that the admission controls also works for the case with several flows per station.

We justified our decision to design and validate our admission threshold with a single, representative traffic pattern: exponentially distributed on/off sources. Additional simulations are needed to validate that our admission control works regardless of the traffic pattern, and we need to include elastic flows in the evaluation.

We have not stated what would trigger the probe; it could be handled outside the MAC. Ideally, we would like to probe with the first packets of a new end-to-end flow to integrate it in our suggested host-based service differentiation scheme [14].

VI. CONCLUSION

In this paper, we have presented a distributed admission control for the DCF of IEEE 802.11 wireless LANs that prevents new sessions from pushing the state of a cell into overload. The admission decision is based on comparing the measured service time of a short, non-disturbing probe with a given admission threshold. The probing provides an estimate of the expected service time, which captures the offered load and transmission rates in the cell in one aggregate measure. It may be done independently by a station to test whether a cell has enough capacity for a connection; the station might search for an alternative access point if the probe result is negative.

We have also proposed to adjust the admission threshold in order to establish an upper bound on the mean packet-loss probability due to congestion, and have described how to calculate the threshold as a function of the number of stations and target loss probability. We have evaluated the performance of our admission function showing the percentage of correct and wrong decisions, as well as the amount of unnecessary blocking. The evaluation shows that 50 packets are enough to perform an accurate decision that effectively limits the loss rate in the cell. This may be performed in a fraction of a second and would constitute a minor part of the overall session duration.

The main drawback of the proposed admission control is a reduction in the link utilization. We have shown that this reduction depends on the number of nodes in the cell and, of course, on the size of the sessions being set up. Nevertheless, we believe that the levels of utilization achieved are acceptable. As stated above, the admission control protects the cell from overload caused by new flows. There could however be overload caused by worsened transmission conditions for one or more station. We have addressed this issue in ref. [2] by means of load balancing. So the two mechanisms complement each other to ensure that a cell is operating well.

In conclusion, it is feasible to protect IEEE 802.11 cells from overload without modification of the distributed coordination function. The suggested admission control might limit losses as well as delay in the cell so that differentiation is not needed; it could also be applied in conjunction with service differentiation according to IEEE 802.11e.

REFERENCES

- [1] H. Velayos Muñoz, *Autonomic Wireless Networking*, Ph.D. thesis, KTH, June 2005. (See "publications" at URL www.ee.kth.se).
- [2] H. Velayos, V. Aleo, and G. Karlsson, "Load balancing in overlapping wireless lan cells," in *Proc. of IEEE ICC*, Paris, France, June 2004.
- [3] H. Velayos, G. Karlsson, "Techniques to Reduce IEEE 802.11b Handoff Time," in *Proc. of IEEE ICC*, Paris, France, June 2004.
- [4] J. W. Roberts, "A survey on statistical bandwidth sharing," *Computer Networks*, vol. 45, pp. 319–332, March 2004.
- [5] LAN/MAN Standards Committee of the IEEE Computer Society, "IEEE Std 802.11-1999, Wireless Lan Medium Access Control And Physical Layer Specifications," USA, August 1999.
- [6] J.-Y. Yeh and C. Chen, "Support of multimedia services with the IEEE 802.11 MAC protocol," in *Proc. of IEEE ICC*, pp. 600–604, 2002.
- [7] C. Coutras, N. Gupta, and N. Shroff, "Scheduling of real-time traffic in IEEE 802.11 wireless LANs," *Wireless Networks*, vol. 6, no. 6, pp. 457–466, 2000.
- [8] Y. Xiao, H. Li, and S. Choi, "Protection and guarantee for voice and video traffic in IEEE 802.11e wireless LANs," in *Proc. of IEEE INFOCOM*, Hong Kong, China, March 2004.
- [9] A. Lindgren, A. Almquist, and O. Schelén, "Quality of service schemes for IEEE 802.11: A simulation study," in *Proc. of IWQoS*, Karlsruhe, Germany, Jun 2001, Springer LNCS, vol. 2092.
- [10] IEEE 802.11 working group task e, "Draft supplement to standard 802.11-1999 wireless medium access control and physical layer specifications: medium access control enhancements for quality of service (QoS)," IEEE 802.11e/D2.0, November 2001.
- [11] Y. Xiao, "IEEE 802.11e: QoS provisioning at the MAC layer", *IEEE Wireless Communications*, vol. 11, issue 3, pp. 72-779, June 2004.
- [12] O. Hagsand, I. Más, I. Marsh and G. Karlsson, "Self-Admission control for IP telephony using early quality estimation," in *Proc. of IFIP Networking*, Athens, Greece, May 2004, Springer LNCS, Vol. 3042.
- [13] V. Fodor (née Elek), G. Karlsson, and R. Rönngren, "Admission Control Based on End-to-End Measurements," in *Proc. IEEE INFOCOM*, Tel-Aviv, Israel, March 26-30, 2000.

- [14] H. Lundqvist, I. Más and G. Karlsson, "Edge-based differentiated services," in *Proc. IWQOS*, 2005.
- [15] P. Key and L. Massoulié, "Probing strategies for distributed admission control in large and small scale systems", in *Proc. of IEEE INFOCOM*, 2003.
- [16] M. Ergen and P. Varaiya, "Throughput Analysis and Admission Control for IEEE 802.11a," Springer Mobile Networks and Applications, vol. 10, no. 5, October 2005, pp. 705–716.
- [17] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," IEEE Journal on Selected Areas in Communications, vol. 18, pp. 535–547, March 2000.
- [18] N. Hegde, A. Proutiere and J. Roberts, "Evaluating the voice capacity of 802.11 WLAN under distributed control," in *Proc. IEEE LANMAN*, 2005.
- [19] P. Chatzimisios, A.C. Boucouvalas, V. Vitsas, "Performance analysis of IEEE 802.11 DCF in presence of transmission errors", in *Proc. of IEEE ICC*, pp. 3854 – 3858, Paris, France, June 2004.
- [20] Y.C. Tay and K.C. Chua, "A capacity analysis for the IEEE 802.11 MAC protocol", *Wireless Networks*, vol 7, pp 159-171, Kluwer Academic Publisher, 2001.
- [21] M. Carvalho and J.J. Garcia-Luna-Aceves, "Delay analysis of IEEE 802.11 in single-hop networks", in *Proc. of the 11th IEEE ICNP*, 2003.
- [22] A. Banchs, "Analysis of the distribution of the backoff delay in 802.11 DCF: a step towards end-to-end delay guarantees in WLANs", in *Proc of QoFIS*, Barcelona, Spain, 2004.
- [23] O. Tickoo and B. Sikdar, "A queuing model for finite load IEEE 802.11 random access MAC", in *Proc. of IEEE ICC*, Paris, France, June 2004.
- [24] B. Li and R. Battiti, "Analysis of the IEEE 802.11 DCF with service differentiation support in non-saturation conditions", in *Proc. of QoFIS*, Barcelona, Spain, 2004.
- [25] NIST/SEMATECH e-Handbook of Statistical Methods, Dec. 2004. Available at <http://www.itl.nist.gov/div898/handbook/>.
- [26] H. Velayos and G. Karlsson, "Statistical analysis of the IEEE 802.11 MAC service time", in *Proc. of ITC 19*, Beijing, China, August 2005.