Communications, 2004 IEEE International Conference on
Volume 7, 20-24 June 2004 Page(s):3844 - 3848 Vol.7
Digital Object Identifier 10.1109/ICC.2004.1313272

Techniques to reduce the IEEE 802.11b handoff time

Héctor Velayos and Gunnar Karlsson
Dept. of Microelectronics and Information Technology
KTH, Royal Institute of Technology
Stockholm, Sweden
{hvelayos,gk}@imit.kth.se

***Abstract*—We analyze the link-layer handoff process in wireless LANs based on the IEEE 802.11b standard and suggest how to reduce its duration. Firstly, we divide the process into three phases: detection, search and execution. Our performance measurements indicate that the detection and search phases are the main contributors to the handoff time. We show that the link-layer detection time can be reduced to three consecutive lost frames. We also show that the search time can be reduced at least by 20% using active scanning with the two timers that control its duration set to 1 ms and 10.24 ms. Several simulations illustrate the achieved reduction in handoff time.**

Keywords-wireless LAN; handoff; performance

I. INTRODUCTION

Wireless LANs based on the IEEE 802.11b standard are the predominant option for wireless access to the Internet. The performance of the cells permits the use of real time services, such as voice over IP, when admission control is added and the MAC scheduler is modified [1]. However, experimental measurements in our testbed, which are summarized in Table I and described later, indicate that current implementations of link-layer handoff do not meet the needs of real time traffic.

In this paper, we propose and evaluate via simulations techniques to minimize the IEEE 802.11b handoff time. We describe the handoff procedure and divide it into three phases. Our main contribution is a set of techniques to reduce the two longer phases, detection and search. The rest of the paper is organized as follows. Section II describes the handoff procedure. Section III presents our measurements of current handoff implementations. Sections IV, V and VI contain our proposals to reduce each of the handoff phases, including simulation results to assess the time reduction achieved. Finally, Section VII summarizes our findings.

II. HANDOFF PROCEDURE

Link-layer handoff is the change of the access point (AP) to which a station is connected. In the case of IEEE 802.11b wireless LANs, the handoff implies a set of actions (e.g. change of radio channel, exchange of signaling messages) that interrupt the transmission of data frames. The duration of this interruption is called handoff time. The handoff procedure aims to reduce this time as much as possible so that upper layers do not notice the handoff, except for a temporarily higher delay on the link. Loss of packets during handoff is avoided by buffering frames in the station and in the old AP. When data transmission is resumed, these frames must be transmitted via the new access point. In addition, the infrastructure connecting the APs, typically a set of Ethernet switches, must be notified of the new position of the station in order to route the frames properly. These two actions lead to different handoff time for uplink and downlink traffic, the latter always being longer. Several authors have proposed solutions to make the uplink and downlink handoff time equal based on an adequate design of the distribution system [2] and the cooperation of access points via their wired interfaces [3]. Since the design of the infrastructure connecting the APs is outside the scope of this paper, we assume that such solutions are in place and thus downlink and uplink handoff times are the same.

The signaling to perform the handoff is specified in the Medium Access Control (MAC) protocol of the IEEE 802.11 standard and is common to the IEEE 802.11a, IEEE 802.11b and IEEE 802.11g supplements. Therefore, in general, our work on handoff optimization can apply to all of them. However, our measurements and simulations focus on IEEE 802.11b.

We propose to analyze the handoff process by splitting it into three sequential phases: detection, search and execution. The detection phase is the discovery of the need for the handoff. The search phase covers the acquisition of the information necessary for the handoff. Finally, the handoff is performed during the execution phase. The following sections detail the events that occur during each phase.

III. MEASUREMENTS OF HANDOFF TIME

The duration of each handoff phase was measured in our testbed. It consists of two co-located IEEE 802.11b access points belonging to the same wireless LAN and connected to an Ethernet switch. Thus, stations can perform link-layer handoffs between APs. Each access point is a PC equipped with a D-Link wireless LAN card running Linux and the Host AP driver [4]. During the experiments, other PCs with the same

TABLE I. LINK-LAYER HANDOFF TIME FOR DIFFERENT IEEE 802.11B CARDS

	D-Link 520	Spectrum24	ZoomAir	Orinoco
Detection	1630ms	1292ms	902ms	1016ms
Search	288ms	98ms	263ms	87ms
Execution	2ms	3ms	2ms	1ms
Total	1920ms	1393ms	1167ms	1104ms

driver were monitoring the activity on the radio channels. For the monitoring PCs, we developed software that captured the frames on the corresponding channel and calculated the duration of each handoff phase. Four commercial IEEE 802.11b cards with different chipsets were selected to measure their handoff time as an average of 10 repetitions. Each station's handoff was measured independently. During the tests, the only traffic in the cells was a flow of packets generated by the station with the characteristics of voice over IP.

We noted in preliminary measurements that commercial wireless LAN cards take advantage of the information provided by the physical layer and completely skip the detection phase. These cards start the search phase when the strength of the received radio signal degrades below a certain threshold. Since we were interested in measuring the performance of the handoff using link-layer detection (i.e. without support from the physical layer), the handoff was forced by abruptly switching off the radio transmitter of the AP to which the station was connected. This allows assessing the importance of using the signal strength in deciding to start the handoff. Handoff measurements using physical layer information have already been reported by Mishra et al. [5]. It can be expected that a wireless LAN card implements both physical and link-layer detection. The later would be preferred in some situations such as wireless LANs featuring admission control or load balancing. In these cases, stations can loose the right to continue transmitting via the current AP regardless of the received signal strength.

As indicated above, we define the handoff time as the time during which the traffic was interrupted. Thus, in our experiments we measure the handoff time from the first non-acknowledged data frame until the transmission of the first frame via the new access point.

Our handoff measurements are presented in Table I. From them we can draw the following conclusions. First, different stations showed different performance, but none matched the delay requirements of real time applications during handoff. Second, detection is the longest phase in all cases, while execution could be neglected. And third, detection and search times widely vary among different models. This was expected since the IEEE 802.11 standard only specifies the mechanisms to implement the handoff, but their combination and duration are left unspecified. The purpose was to allow the manufacturers some freedom to balance between different tradeoffs such as fast reaction or low power consumption.

The length differences in detection and search could be explained by analyzing the frames captured during the handoffs. This type of analysis produced the following conclusions. The need for handoff is detected at the link-layer after several non-acknowledged frames. The number of allowed failed frames is the main factor in controlling the duration of the detection phase. It varies with each card model because when a frame is not acknowledged, the station cannot differentiate whether the reason was a collision, congestion in the cell or the access point being out of range. Different cards use different assumptions depending on their purpose. For instance, the D-Link 520 is designed for a desktop PC, thus it assumes that the AP is always in range and retransmits for a longer period than the Orinoco card designed for laptops. Nevertheless, it was common to all the cards to reduce the bit rate and use the RTS/CTS mechanism after failed frames to overcome possible radio fading or collisions in an overloaded cell. Surprisingly, none of the analyzed models used the lack of beacon reception to discover that the access point was not in range. Regarding the search phase, all cards performed active scanning based on broadcasting probes to locate APs. The duration's variance is due to the different number of probe requests sent per channel and more significantly due to the time to wait for probe responses.

The most detailed handoff measurements previously reported are [5]. Our measurements are in line with that work but numerical comparison is difficult because different cards were used. Additionally, their definition of handoff time does not include the link-layer detection phase. In their experiments, stations voluntarily started the search phase when the signal from the AP became weaker than a threshold.

The main conclusion from our measurements is that detection and search phases are the main contributors to handoff time. Therefore, we analyzed them in the following sections and suggest how they can be reduced.

IV. REDUCING THE DETECTION PHASE

The actions during the detection phase vary depending on which entity initiated the handoff. When the handoff is network initiated, the detection phase consists of a single disassociation message sent by an access point to the station. However, the most common handoff is the one initiated by the station, in which stations have to detect the lack of radio connectivity based on weak received signal reported by the physical layer or failed frame transmissions. QoS-concerned stations implement the former method because the handoff is initiated before any frame is lost. This method assumes that the density of APs is high and therefore, there is a better AP in range as soon as the received signal gets weak. On the other hand, the latter method produces less handoff events because the handoff is not triggered by temporary radio fading or interferences, but only when transmission is actually interrupted.

Our study focuses on the optimization of detection based on failed frames, i.e. link-layer detection without physical layer information. The main difficulty is to determine the reason for the failure among collision, radio signal fading, or the station being out of range. We have observed in our measurements that stations firstly assume collision and retransmit several times using lower bit rates. If transmission remains unsuccessful, then radio fading is assumed and probe request are sent to check the link. Only after several unanswered requests, the station declares the out of range status and starts the search phase. Different cards showed different detection times depending on the number of failed frames allowed and the number of probes sent. As Table I indicates, this type of detection procedure tends to be long, so we suggest a different approach: stations must start the search phase as soon as collisions can be excluded as the reason for failure. If the actual reason was a temporary signal fading, the selected access point after the search would likely be the current one and the handoff

will not be executed. This means that independently of the duration of the fading, the data flow will be interrupted for the duration of the search phase, which further motivates the reduction of that phase.

Therefore, a key factor in our detection algorithm is the number of collisions that a frame can suffer before it is transmitted. Let C be the random variable representing the number of collisions per successfully transmitted frame. Its cumulative distribution function (CDF) is given by

$$\Pr ob(C \le k) = \sum_{i=0}^{k} (1-p)p^i = 1 - p^{k+1} \qquad (1)$$

where p is the probability, seen by the station, that its transmitted frame collides. This probability depends on the number of stations competing for the medium, and it can be calculated with the non-linear system reported by Bianchi in [6] for saturated conditions (i.e. all stations always have a frame ready to transmit) that is the worst case for collisions. The CDF of the number of collisions per transmitted frame is plotted in Fig. 1. This figure shows that three consecutive collisions is a rare event, even in saturation. Therefore, our link-layer detection algorithm can be formulated as follows: if a frame and its two consecutive retransmissions fail, the station can discard collision as the cause of failure and start the search phase; there is no need to explicitly probe the link. In the same conditions we used during our measurements, this time would be around 3 ms, which is approximately 300 times shorter than the fastest measured detection phase. A drawback of this link-layer detection algorithm is that its duration increases with the cell load and the transmitted frame length.

A special situation happens when stations are not sending traffic at the time of handoff, but only receiving. In this case, stations must track the beacon reception to differentiate between the situation when the access point has no traffic addressed to them or the AP is out of range. Stations must start the search phase after three beacons are missing and no traffic to other stations was received. Stations should not react at the first missing beacon because beacons can also be lost due to collisions. This converts the beacon period into another key factor to reduce the detection time. The shorter the period is, the shorter the detection time would be. But as the beacon period is reduced, more capacity is used for sending beacons

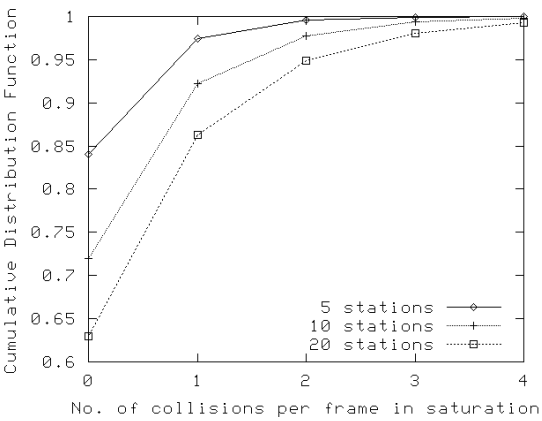

Figure 1. No. of collisions per transmitted frame in saturation

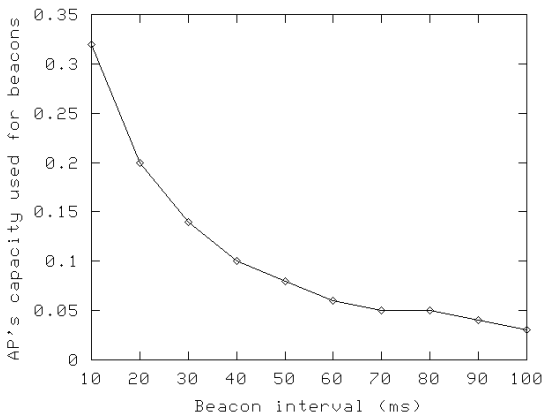

Figure 2. AP's capacity used for beacon transmission in a saturated cell

instead of data frames. We have added beacon transmission to the ns-2[1] IEEE 802.11 module to evaluate this trade-off. Fig. 2 shows the result of our simulations for a saturated IEEE 802.11b cell.

This result confirms the expected behavior and allows selecting an adequate beacon interval. Currently, commercial IEEE 802.11b access points are shipped with a default 100 ms beacon interval. This means that approximately 4% of the AP's capacity is used for beacons and that detection time based on three missed beacons would be 300 ms. Fig. 2 indicates that increasing the used capacity only to 6% would reduce the beacon interval to 60 ms. This would reduce by 60% the detection time (i.e. to 180 ms). Further reductions of the beacon interval, and thus the detection time, are possible but at the cost of noticeably decreasing the AP's capacity.

V. REDUCING THE SEARCH PHASE

The search phase includes the set of actions performed by the station to find the APs in range. Since APs can operate in any channel of the allowed set, the IEEE 802.11 standard mandates that all allowed channels must be scanned. The standard also specifies two methods to scan a channel, active and passive scanning. In passive scanning, stations listen to each channel for the beacon frames. The main inconvenience of this method is how to calculate the time to listen to each channel. This time must be longer than the beacon period, but the beacon period is unknown to the station until the first beacon is received. Another problem is its performance. Since the whole set of allowed channels must be scanned, stations need over a second to discover the access points in range with the default 100 ms beacon interval. There are 11 allowed channels in USA, thus it would take 1.1 seconds. In most of Europe, there are 13 channels, so it would take 1.3 seconds.

When faster scanning is needed, stations must perform active scanning. Active scanning means that stations will broadcast a probe-request frame on each channel and wait for the probe response generated by the access point. The time to wait for responses depends on the channel activity after the probe transmission. If the channel is idle during

[1] ns-2 is a network simulator developed at the Information Science Institute, USC. (http://www.isi.edu/nsnam/ns/)

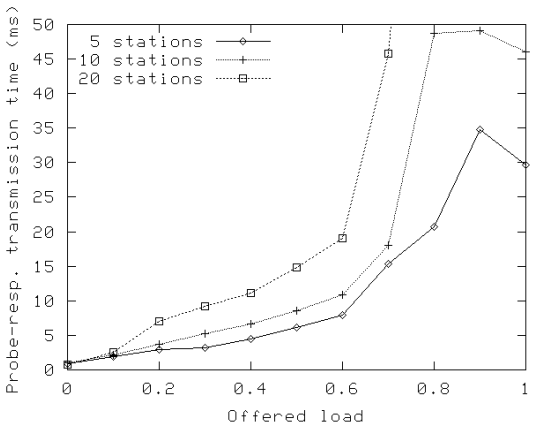

Figure 3. Probe response transmission time (ms)

MinChannelTime, i.e. there is neither response nor any kind of traffic in the channel, the scanning is finished and the channel is declared empty. If there is any traffic during this time, the station must wait MaxChannelTime. MaxChannelTime should be large enough as to allow the AP to compete for the medium and send the probe response. Both MaxChannelTime and MinChannelTime are measured in Time Units (TU). The IEEE 802.11 standard defines a TU to be 1024 microseconds. Note that scanning stations might not be able to sense other stations communicating with the AP, but they will always receive the acknowledgments sent from the AP and thus they will wait MaxChannelTime for probe responses.

Despite that MinChannelTime and MaxChannelTime control the duration of the scanning, the IEEE standard does not specify their values. We indicate in the rest of this section how to calculate them to minimize the search phase. First, we calculate MinChannelTime that is the maximum time an access point would need to answer given that the access point and channel were idle. If we neglect propagation time and probe response generation time, the IEEE 802.11 medium access function establishes that the maximum response time is given by

$$MinChannelTime = DIFS + (aCWmin \times aSlotTime) \qquad (2)$$

where DIFS is the Distributed InterFrame Space, aCWmin is the maximum number of slots in the minimum contention window, and aSlotTime is the length of a slot. These values are defined in the IEEE 802.11b standard and inserting them in (2), we obtain 670 μs. Since MinChannelTime must be expressed in Time Units, we can conclude that MinChannelTime should be one TU (i.e. 1024 μs).

The calculation of MaxChannelTime is more complicated. It is the maximum time to wait for a probe response when the channel is being used. This time is not constant since it depends on the cell load and the number of stations competing for the channel. In order to find an upper bound for MaxChannelTime, we have run simulations to measure the time to transmit the probe response. Fig. 3 presents the results of our simulations. The probe response time shown is the average over 10 transmissions for each load level with channel bit rate set to 2 Mbps, the maximum possible rate for the probe response in IEEE 802.11b.

Our simulations confirm that the transmission time of a probe response depends on the offered load and number of stations. In addition, they also show that MaxChannelTime is not bounded as long as the number of stations can increase. We suggest then to set a value for MaxChannelTime that would prevent overloaded access points to answer in time. Since 10 stations per cell seems to be an adequate number to achieve a good cell throughput [6], Fig. 3 indicates that 10 TU (10.24 ms) would be a reasonable choice for MaxChannelTime.

Now that we have determined MinChannelTime and MaxChannelTime and that both timers are shorter than feasible beacon intervals, it is clear that active scanning is faster than passive scanning. Thus, active scanning should be used to minimize channel-scanning time.

Finally, we have to calculate the total search time that includes the time to scan all available channels. The number of available channels varies with regions. For instance, there are 13 possible channels in most of the European countries, while there are 11 in USA. Considering that the time to scan a channel is different depending whether it is been used, the total search time s can be calculated as

$$s = uT_u + eT_e \qquad (3)$$

where u is the number of used channels (i.e with traffic) and T_u is the time needed to scan a used channel. Respectively, e is the number of empty channels and T_e is the time to scan an empty channel. We can now determine T_u and T_e. When a channel is scanned, a probe request is broadcasted and then the station waits for the probe response. Since the probe request is sent to the broadcast address, its reception will not be acknowledged. Therefore, at least two consecutive probe requests must be sent to overcome a possible collision. Each probe request must follow the same channel access procedure as the data packets, thus they will experience the transmission delay. Let T_d be the transmission delay, then we can calculate T_u and T_e as follows:

$$\begin{aligned} T_u &= 2\,T_d + MaxChannelTime \\ Te &= 2\,T_d + MinChannelTime \end{aligned} \qquad (4)$$

The total search time can be calculated with (3) and (4), as well as the transmission delay. It increases with the number of used channels because MaxChannelTime is larger than MinChannelTime. This is illustrated in Fig. 4, which shows the total search time versus number of used channels in range for different load conditions. To plot it, we obtained T_d from our delay simulations reported in Fig. 5. In Fig. 4, we used T_d for an offered load of 50% with 5 and 10 stations per cell. In addition, we included a no-load case that is comparable with our measurements conditions reported in Table I. This case shows that the search time can be reduced to 70 ms when handing over between two APs, which is 20% faster than the shortest search phase measured.

These values in the x-axis of Fig. 4 are particularly interesting: one channel used would be the case of a search phase started due to radio fading when there are no other access points in range; two channels used would be the case of a handoff between two access points, the current and the new; and three channels used is an interesting value since it is the

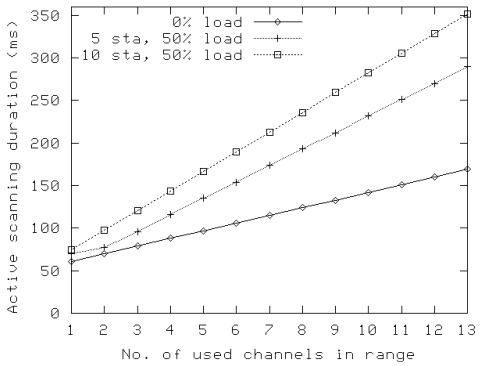

Figure 4. Total search time (ms)

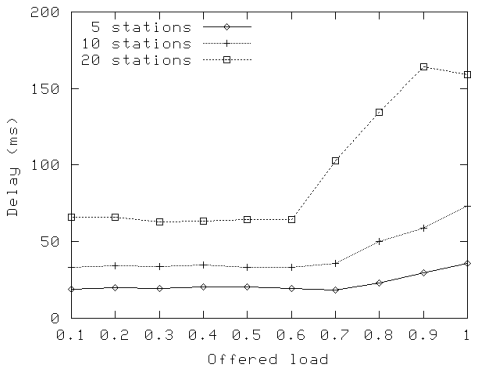

Figure 5. Delay versus load

maximum number of channels that can share the same physical location without mutual interference.

Two problems regarding the search time must be highlighted. First, all access points in a given location affect the handoff time of stations, even access points belonging to different wireless LANs, because MaxChannelTime will be spent scanning their channels. Second, in areas with a high density of access points, search time can increase over the limits of real time applications. Both problems could be addressed with a small modification to the standard: the active scanning should not scan all available channels in a region (e.g. Europe or USA), but a shorter list with the channels actually used in the wireless LAN to which the station is connected. This is feasible since most wireless LANs use a fixed subset of the available channels. The list could be distributed as an additional field in the beacons.

VI. REDUCING THE EXECUTION PHASE

The last phase is the execution of the handoff. To perform the handoff, the station sends a reassociation request to the new access point and the AP confirms the reassociation sending a response with a status value of "successful". This execution is the shortest possible, but the typical execution is longer because the new access point needs to authenticate the station before the reassociation succeeds.

The IEEE 802.11 standard specifies two authentication algorithms: open system and shared key. The open system is the default and equals to a null authentication algorithm. It involves the exchange of two frames, while the shared key algorithm requires a four-step transaction. Our measurements show that the execution phase using open system authentication is slightly over 1 ms for an empty cell, thus reducing the execution phase will not significantly reduce the total handoff time. Nevertheless, there are more complicated authentication schemes under study that require contacting an external server. In these cases, the authentication must be made before the handoff execution [7].

VII. CONCLUSIONS

We have measured, analyzed and suggested how to reduce the link-layer handoff time in IEEE 802.11b networks. The handoff process was split into three sequential phases: detection, search and execution. We studied the detection based on failed frames (link-layer detection) instead of weak signal because it produces less handoff events. We have shown that the link-layer detection phase can be reduced to three consecutive non-acknowledged frames when stations are transmitting. In the same conditions we used during our measurements, this time would be around 3 ms, which is approximately 300 times shorter than the fastest measured detection phase. When stations are only receiving, we identified the beacon interval as the key factor in reducing detection time. Our simulations suggest 60 ms as an adequate beacon interval. We have also shown that using active scanning with its timers MinChannelTime and MaxChannelTime set to 1 ms and 10.24 ms respectively can reduce the search phase by 20% compared to the shortest measured one. Finally, the execution phase can be reduced with pre-authentication, but our measurements indicate that it is a very short phase and its reduction will not significantly decrease the total handoff time when using the current authentication methods.

REFERENCES

[1] M. Barry, A. T. Campbell, A. Veres, "Distributed control algorithms for service differentiation in wireless packet networks", Proc. IEEE INFOCOM 2001, Anchorage, Alaska

[2] Amre El-Hoiydi, "Implementation options for the distribution system in the 802.11 Wireless LAN Infrastructure Network", Proc. IEEE ICC 2000, vol. 1, pages 164-169, New Orleans, USA, June 2000.

[3] Anne H. Ren, Gerald Q. Maguire Jr., "An adaptive real-time IAPP protocol for supporting multimedia communications in wireless LAN systems", Int. Conf. on Computer Communications, Japan, Sept. 1999.

[4] Host AP driver, http://hostap.epitest.fi/, last visit March 2004

[5] Arunesh Mishra, Minho Shin, Willian Arbaugh, "An empirical analysis of the IEEE 802.11 MAC layer handoff process", University of Maryland Technical Report, UMIACS-TR-2002-75, 2002

[6] Giuseppe Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function", IEEE Journal on Selected Areas in Communication, 18(3): 535 - 547, March 2000.

[7] Sangheon Pack, Yanghee Choi, "Pre-authenticated fast handoff in a public wireless LAN based on IEEE 802.1x Model", IFIP TC6 Personal Wireless Communications 2002, Singapore, October 2002