KUNGL TEKNISKA HÖGSKOLAN
Royal Institute of Technology

Techniques to Reduce IEEE 802.11b MAC Layer Handover Time

Héctor Velayos,
Gunnar Karlsson

April 2003

TRITA-IMIT-LCN R 03:02
ISSN 1651-7717
ISRN KTH/IMIT/LCN/R-03/02--SE

Laboratory for Communication Networks
Department of Microelectronics and Information Technology (IMIT)
KTH, Royal Institute of Technology
Stockholm, Sweden

Techniques to reduce IEEE 802.11b MAC layer handover time

Héctor Velayos Gunnar Karlsson
KTH, Royal Institute of Technology
P.O. Box Electrum 229
SE-16440 Kista, Sweden
{hvelayos,gk}@imit.kth.se

Abstract— **In this paper we analyze link-layer handover times in wireless local area networks based on the IEEE 802.11b MAC protocol. Our measurements indicate that detection and search phases are the main contributors to the handover time. We show that detection time can be reduced by reacting quickly to packet losses and by using shorter beacon intervals. We also show that search time can be reduced by using active scanning. In this case, we calculate values for the two timers that control the duration of active scanning in order to reduce search time. Several simulations illustrate the achieved reduction in handover time.**

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 test-bed, summarized in Table 1 and described later, indicate that current implementations of MAC layer handover do not meet the needs of real time traffic.

In this paper, we propose and evaluate via simulations techniques to reduce the IEEE 802.11b handover time. We describe the handover procedure and divide it into three phases. Our main contribution is a set of techniques to reduce the two longer phases, detection and search, without modifying the current IEEE 802.11b standard. The rest of the paper is organized as follows. Section II describes the handover procedure and our measurements of current handover implementations. Sections III, IV and V contain our proposals to reduce each of the handover phases, including simulation results to measure the time reduction. Section VI studies the phases which can be run in parallel. Finally, in Section VII we summarize our findings.

Table 1: Handover time for different IEEE 802.11b cards

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

II. HANDOVER PROCEDURE

Link-layer handover is the change of the access point (AP) to which a station is connected. In the case of IEEE 802.11b wireless LANs, handover 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 handover time.

The handover procedure aims to reduce this time as much as possible so that upper layers do not notice the handover, except for a temporarily higher delay on the link. Loss of packets is avoided by buffering the frames in the station and in the old AP during handover. 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 handover time for uplink and downlink traffic, the latter always being longer. Several authors have proposed solutions to make uplink and downlink handover time equal based on an adequate design of the distribution system [2] and cooperation of access points [3]. In this paper we assume that such solutions are in place and thus downlink and uplink handover times are the same.

The mechanisms to perform the handover are specified in the Medium Access Control (MAC) protocol of the IEEE 802.11 standard and are common to IEEE 802.11a and IEEE 802.11b supplements. Therefore, in general, our work on handover optimization can apply to both. However, our measurements and simulations focus on IEEE 802.11b.

We propose to split the handover process into three phases to simplify the analysis. The first phase, called detection, is the discovery of the need for the handover. The second phase, search, covers the acquisition of the information needed to perform the handover. Finally, the handover is performed during the third phase, execution. The following sections detail the events that occur during each phase.

The durations of the phases were 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 handover 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 driver were monitoring the activity of the radio channels. We developed software that captured the frames on the corresponding channel and calculated the duration of each handover phase. We selected four commercial IEEE 802.11b cards with different chipsets and measured their handover time as an average of 10 repetitions. There was one station at a time and it generated a flow of packets with the characteristics of voice over IP. Handover was forced by temporarily switching off the radio transmitter of the AP to which the station was connected. The handover time is the time during which the traffic was interrupted, that is, from the first non-acknowledged data frame until the transmission of the first frame via the new access point.

Our measurements are presented in Table 1. 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 a handover. 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 802.11b standard specifies the mechanisms to implement these phases, but their combination and duration are left unspecified. This allows manufacturers to balance between fast reaction and low power consumption.

The differences in behavior during detection and search could be analyzed by looking at the frames captured during the handovers. This type of analysis produced the following conclusions. The need for handover is detected after several non-acknowledged frames. The number of failed frames is the main factor in controlling the duration of the detection phase and it varies with each card model. When a frame is not acknowledged, the station cannot differentiate whether the reason was a collision, congestion in the cell or access point 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. 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 more detailed measurements previously reported are [5]. Our measurements and conclusions are in line with that work. Nevertheless, numerical comparison is difficult because different card and AP models were used. Additionally, their definition of handover time does not include the 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 should be carefully analyzed and reduced.

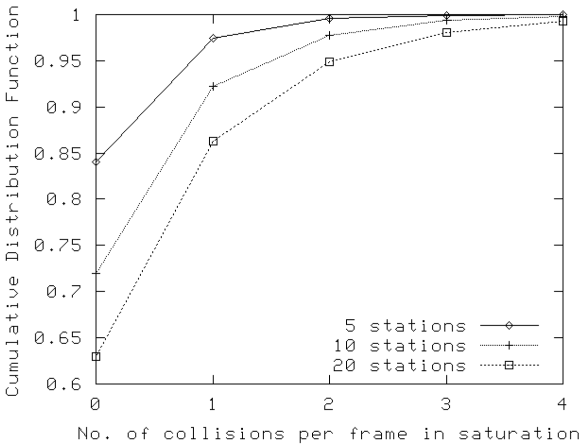

Figure 1: No. of collisions per frame in saturation

III. REDUCING THE DETECTION PHASE

The actions during the detection phase vary depending on which entity initiated the handover. When the handover is network initiated, the detection phase consists of a single disassociation message sent by an access point to the station. This is the fastest detection phase. However, the most common handover is the one initiated by the station, in which stations have to detect the lack of radio connectivity based on failed frame transmissions. 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 explicitly probe the link by sending probe requests after a series of unsuccessful transmissions. Different cards showed different detection times depending on the number of failed frames allowed and the number of probes sent.

As Table 1 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 collision can be excluded as reason for failure. If the reason was a temporary signal fading, the selected access point after the search would likely be the current one and the handover 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 need to reduce the search phase.

Let C be the random variable representing the number of collisions per frame transmission. Its cumulative distribution function (CDF) is given by:

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

Where p is the probability, seen by the station, that its transmitted frame collides. This probability can be calculated with the non-linear system reported by Bianchi

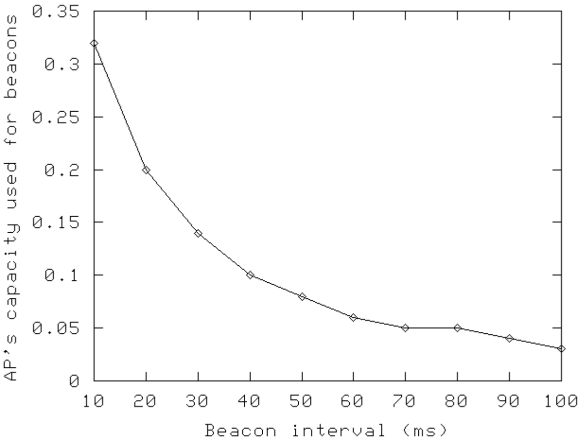

Figure 2: AP's capacity used for beacon transmission

in [6] for saturated conditions, i.e. the worst case for collisions. The CDF of the number of collisions per frame is plotted in Figure 1. This figure shows that three consecutive collisions is a rare event, even in saturation. Therefore, if a frame and its two consecutive retransmissions fail, the station can discard collision as the cause of failure and start the search phase. Hence, there is no need to probe the link.

A special situation happens when stations are not sending traffic at time of handover, 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 some beacons are missing since beacons can collide. 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 beacon transmissions. We have used ns-2[1] to evaluate this trade-off. Figure 2 shows the result of our simulations for a saturated IEEE 802.11b cell.

This result confirms the expected behavior and allows selecting a proper beacon interval. Currently, commercial IEEE 802.11b access points are shipped with a default 100 milliseconds beacon interval. This means that approximately 4% of capacity is used for beacons. Figure 2 indicates that the beacon interval can be reduced to 60 ms without noticeably increasing the used capacity. Further reductions of the beacon interval are possible, but with the cost of increased used bandwidth.

Finally, we noted in our measurements that some WLAN implementations take advantage of the information provided by the physical layer and can completely skip the detection phase. These stations start the search phase when the quality of radio signal degrades below a configurable threshold. In this case, the search starts before any frame has been lost.

IV. REDUCING THE SEARCH PHASE

The search phase includes a set of actions performed by the station to find the APs in range. The IEEE 802.11 standard specifies two scanning modes, 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. Incidentally, the station cannot switch to another channel when the first beacon arrives and has to wait for the whole beacon period because several access points of different WLANs can operate in the same channel. Since the standard mandates that 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 (e.g. there are 11 allowed channels in USA, thus it would take 1.1 seconds).

When faster scanning is needed, stations must perform active scanning. Active scanning means that stations will broadcast a probe-request management frame in each channel and wait for probe responses generated by access points. The time that stations should wait for responses in each channel is controlled by two timers: MinChannelTime and MaxChannelTime. The first is the time to wait for the first response in an idle channel. If there is neither response nor traffic in the channel during MinChannelTime, the channel is declared empty (i.e. no access point in range). The second timer, MaxChannelTime, indicates the time to wait in order to collect all responses in a used channel. This limit is used when there was activity in the channel during MinChannelTime. Both timers are measured in *Time Units* (TU), which the IEEE standard defines to be 1024 microseconds. Exact values for these timers are not given

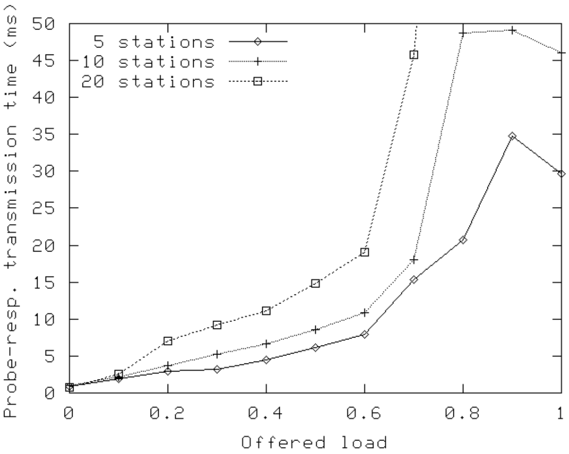

Figure 3: Probe response transmission time (ms)

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

in the standard, but we indicate below how to calculate them.

First, we calculate MinChannelTime. This value is bounded by 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 maximum response time is given in (2):

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

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. Table 2 contains these values for the IEEE 802.11b standard. Inserting them in (2), we obtain 670 µs. Since MinChannelTime must be expressed in *Time Units*, we can conclude that minimum MinChannelTime is one TU (i.e. 1024 µs).

Table 2: Physical characteristics for IEEE 802.11b standard

	IEEE 802.11b
aSlotTime	20 µs
aCWmin	31 slots
DIFS	50 µs

The calculation of MaxChannelTime is more complex. It is the maximum time to wait for a probe response when the channel is being used. In order to find an upper bound for MaxChannelTime, we have run simulations to measure the time to transmit the probe response. Figure 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 management frames.

Our simulations show that the transmission time of a probe response depends on 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 seem to be an adequate number to achieve a good cell throughput [6], Figure 3 indicates that 10 ms would be a reasonable choice for MaxChannelTime. It is important to maintain this value small to prevent an excessively long search phase in areas with several access points.

Now that we have determined MinChannelTime and MaxChannelTime and that both timers are shorter than reasonable beacon intervals, it is clear that active scanning is faster than passive scanning. Thus active scanning should be used to reduce the search phase.

Finally, we have to calculate the total search time. The IEEE standard requires that stations must scan all available channels during active scan. The available channels vary with the regions. For instance, there are 13 possible channels in most of the European countries, while there are 11 in the USA. 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 now determine T_u and T_e. When a channel is scanned, first a probe request is broadcasted and then the station waits for probe responses. Since the probe request is sent to the broadcast address, its reception will not be acknowledged. Therefore, at least two 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 suffer the cell transmission delay. Let T_d be the transmission delay, then we can calculate T_u and T_e as:

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

Total search time can be calculated with (3) and (4), as well as the values indicated before for MaxChannelTime and MinChannelTime. It increases with the number of used channels because MaxChannelTime is larger than MinChannelTime. This is illustrated in Figure 4, which shows the total search time versus number of used channels in range. To plot it, we obtained T_d from our delay simulations reported in Figure 5. 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 in Figure 4 to show the search time reduction achieved in comparison with our measurement in Table 1.

Some of the x-axis values in Figure 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

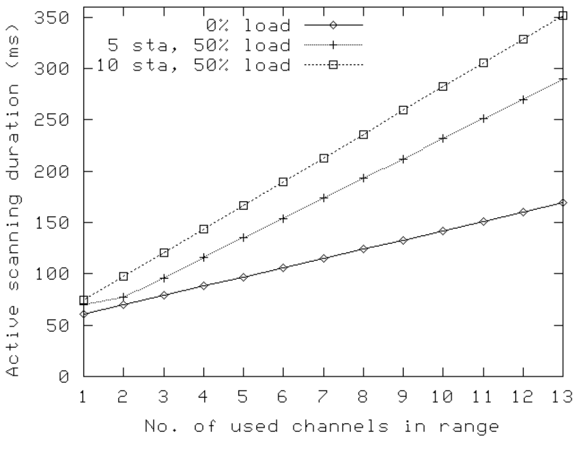

Figure 4: Total search time (ms)

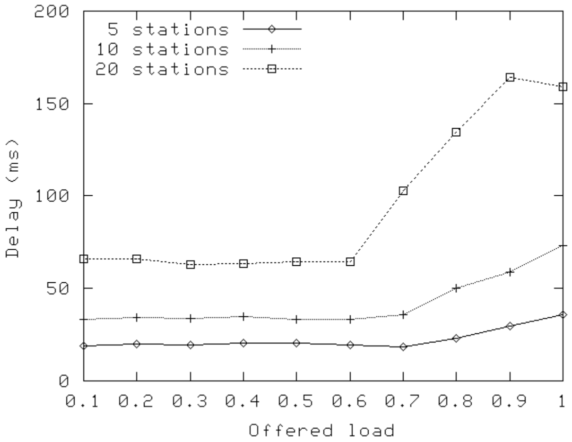

Figure 5: Delay versus load

the case of a handover between two access points, the current and the new. Three channels used is an interesting value since it is the maximum number of channels than 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 handover time of stations, even access points belonging to different wireless LANs to which stations cannot hand over. 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 of the standard: the active scanning should not scan all available channels in a region (e.g. Europe or the USA), but a smaller list of configured channels. 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 beacons.

V. REDUCTION OF THE EXECUTION PHASE

The execution is a two-step process. 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 802.11 standard specify 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, thus reducing the execution phase using pre-authentication will not significantly reduce the total handover time. Nevertheless, there are more complicated authentication schemes under study that require contacting an external server. In those cases, the authentication must be made before the handover execution [7].

VI. HANDOVER PHASES IN PARALLEL

The previous sections have considered the phases running in sequence, but further reduction of the total handover time can be achieved if they can run in parallel. Search must finish before execution because it provides the necessary information, and detection will always be before execution. Therefore, only detection and search can run in parallel.

Moreover, it is advisable to eliminate the search phase completely from the handover procedure since it can run in parallel with the data transmission. Stations can periodically scan one channel at a time actively to discover alternative access points while being connected to an AP. The duration of the one-channel scan was calculated before and is short enough not to impact on ongoing data transmission severely.

VII. CONCLUSIONS

We have measured, analyzed and suggested means to reduce the link-layer handover time in IEEE 802.11 networks. The handover was split into three phases, typically performed in sequence: detection, search and execution. We have shown that the 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 phased. When stations are only receiving, beacon interval plays a key role. A shorter beacon interval reduces the detection, but it also reduces the available capacity for data. This effect was evaluated in our simulations and 60 ms was suggested as adequate beacon interval. We have also shown that search phase can be reduced by using active scanning if its timers are set to the values we have deduced. Finally, 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 handover time when using current authentication methods. Additionally, further handover time reduction is possible if search phase is performed in parallel with data transmission.

ACKNOWLEDGMENTS

We would like to thank Constantin Adam, Alberto Gonzalez, Henrik Lundqvist, Evgeny Osipov, Enrico Pelletta and Jon-Olov Vatn for their helpful remarks and valuable suggestions to improve this work.

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 International Conference on Communications 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", Proc. of International Conference on Computer Communications, pp. 437 - 442, Japan, Sept. 1999.

[4] Host AP driver, http://hostap.epitest.fi/, last visit Jan 2003

[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.