An Empirical Analysis of the IEEE 802.11 MAC Layer Handoff Process

Arunesh Mishra  
Dept of Computer Science  
University of Maryland  
College Park, MD, USA  
arunesh@cs.umd.edu

Minho Shin  
Dept of Computer Science  
University of Maryland  
College Park, MD, USA  
mhshin@cs.umd.edu

William Arbaugh  
Dept of Computer Science  
University of Maryland  
College Park, MD, USA  
waa@cs.umd.edu

ABSTRACT
IEEE 802.11 based wireless networks have seen rapid growth and deployment in the recent years. Critical to the 802.11 MAC operation, is the handoff function which occurs when a mobile node moves its association from one access point to another. In this paper, we present an empirical study of this handoff process at the link layer, with a detailed breakup of the latency into various components. In particular, we show that a MAC layer function - probe is the primary contributor to the overall handoff latency. In our study, we observe that the latency is significant enough to affect the quality of service for many applications (or network connections). Further we find a large variation in the latency with from one handoff to another and also among APs and STAs used from different vendors. In this study, we account for this variation and also draw the guidelines for future handoff schemes.

General Terms
Measurement, Performance, Experimentation

Keywords
IEEE 802.11, Handoff, Performance, Scanning, Probe, Association, Authentication, Latency

1. INTRODUCTION
IEEE 802.11 based wireless local area networks (WLANs) have seen immense growth in the last few years. The predicted deployment of these networks for the next decade resembles that of the Internet during the early 90s. In public places such as campus and corporations, WLAN provides not only convenient network connectivity but also a high speed link up to 11 Mbps (802.11b). In this paper, we are concerned with the IEEE 802.11b network which operates in the 2.4 GHz range.

The IEEE 802.11 network MAC specification [4] allows for two operating modes namely, the ad hoc and the infrastructure mode. In the ad hoc mode, two or more wireless stations (STAs) recognize each other and establish a peer-to-peer communication without any existing infrastructure, whereas in infrastructure mode there is a fixed entity called an access point (AP) that bridges all data between the mobile stations associated to it. An AP and associated mobile stations form a Basic Service Set (BSS) communicating on the unlicensed RF spectrum.

A collection of APs (connected through a distribution system DS) can extend a BSS into an Extended Service Set (ESS) refer figure 1).

A Handoff occurs when a mobile station moves beyond the radio range of one AP, and enters another BSS (at the MAC layer). During the handoff, management frames are exchanged between the station (STA) and the AP. Also the APs involved may exchange certain context information (credentials) specific to the station. Consequently, there is latency involved in the handoff process during which the STA is unable to send or receive traffic.
Because of the mobility-enabling nature of wireless networks, there is opportunity for many promising multimedia and peer-to-peer applications (such as VoIP [3], 802.11 phones, mobile video conferencing and chat). Also, many believe that WLANs may become or supplement via hot spots the next generation 4G wireless networks. Unfortunately, the network connection as perceived by the application can suffer from the jittery handoff latencies. As a matter of fact, our measurements not only show that the latencies are very high, but also show that they vary significantly for the same configuration of STAs and APs.

Despite the growing popularity of WLANs, there has been no prior measurement based analysis of the handoff process. There is prior work on performance measurement in ATM-based wireless networks ([12], [8], [15]) and cellular wireless networks ([13]). In [2], Balachandran et. al. present an empirical characterization of user behavior and network performance in a public wireless LAN where they show the varying number of handoffs with time. There has been work on new handoff schemes in [14], [10], [11] and [7] focusing on reducing WLAN handoff latency, but none of these efforts have measured the current handoff latency.

In this study, we conduct experiments to accurately measure the handoff latency in an in-building wireless network. The measurements are done on two co-existing wireless networks (utilizing APs from two popular vendors), and using three wireless NICs from different vendors. We analyze the handoff latencies by breaking down the whole process into various phases to assess the contribution of each phase to the handoff latency. Our results show that the probe phase is the significant contributor to the handoff latency and the variations in the probe-wait time account for the large variations in the overall handoff latency.

The rest of the paper is organized as follows. Section 2 gives details about the handoff process as specified by the standard. Section 3 explains the methodology used for taking the measurements. We present the analysis and results in section 4. Section 5 concludes the study.

2. THE HANDOFF PROCESS

The handoff function or process refers to the mechanism or sequence of messages exchanged by access points and a station resulting in a transfer of physical layer connectivity and state information from one AP to another with respect to the station in consideration. Thus the handoff is a physical layer function carried out by at least three participating entities, namely the station, a prior-AP and a posterior-AP. The AP to which the station had physical layer connectivity prior to the handoff is the prior-AP, while the AP to which the station gets connectivity after the handoff is the posterior-AP. The state information that is transferred typically consists of the client credentials (which allow it to gain network access) and some accounting information. This transfer can be achieved by an (currently draft [5]) Inter Access Point Protocol (IAPP), or via a proprietary protocol. For an IEEE 802.11 network that has no access control mechanism, there would be a nominal difference between a complete association and a handoff/reassociation. Looking at it another way, the handoff-latency would be strictly greater than association latency as there is an additional inter-access point communication delay involved.

2.1 Logical steps in a handoff

The complete handoff process can be divided into two distinct logical steps: (i) Discovery and (ii) Reauthentication as described below. Later we shall see that the actual sequence of messages exchanged perform either one of these two functions.

1. Discovery: Attributing to mobility, the signal strength and the signal-to-noise ratio of the signal from a station's current AP might degrade and cause it to loose connectivity and to initiate a handoff. At this point, the client might not be able to communicate with its current AP. Thus, the client needs to find the potential APs (in range) to associate to. This is accomplished by a MAC layer function: scan. During a scan, the card listens for beacon messages (sent out periodically by APs at a rate of 10 ms), on assigned channels. Thus the station can create a list of APs prioritized by the received signal strength.

There are two kinds of scanning methods defined in the standard: active and passive. As the names suggest, in the active mode, apart from listening to beacon messages (which is passive), the station sends additional probe broadcast packets on each channel and receives responses from APs. Thus the station actively probes for the APs.

2. Reauthentication: The station attempts to reauthenticate to an AP according to the priority list. The reauthentication process typically involves an authentication and a reassociation to the posterior AP. The reauthentication phase involves the transfer of credentials and other state information from the old-AP. As mentioned earlier, this can be achieved through a protocol such as IAPP [5]. In the experiments detailed in this paper, we do not have the draft standard IAPP communication setup but the proprietary inter-access point communications were allowed (between APs of the same vendor). Thus the authentication phase is just a null authentication in our experiments.

Figure 3: The Handoff Measurement Setup

Figure 2 shows the sequence of messages typically observed during a handoff process. The handoff process starts with the first probe request message and ends with a reassociation response message from an AP. We divide the entire handoff latency into three delays which we detail below.
1. **Probe Delay**: Messages A to E are the probe messages from an active scan. Consequently, we call the latency for this process, probe delay. The actual number of messages during the probe process may vary from 3 to 11.

2. **Authentication Delay**: This is the latency incurred during the exchange of the authentication frames (messages E and F). Authentication consists of two or four consecutive frames depending on the authentication method used by the AP. Some wireless NICs try to initiate reassociation prior to authentication, which introduces an additional delay in the handoff process and is also a violation of the IEEE 802.11 [4] state machine.

3. **Reassociation Delay**: This is the latency incurred during the exchange of the reassociation frames (messages G and H). Upon successful authentication process, the station sends a reassociation request frame to the AP and receives a reassociation response frame and completes the handoff. Future implementations will include additional IAPP messages during this phase which will further increase the reassociation delay.

As a note, according to our analysis presented above, the messages during the probe delay form the discovery phase, while the authentication and reassociation delay form the reauthentication phase. Apart from the latencies discussed above, there will potentially be a bridging delay caused by the time taken for the MAC address updates (using the IEEE 802.1d protocol) to the Ethernet switches which form the distribution system (the backbone Ethernet). The results in our experiments will not reflect this latency. In the next section we describe the details of the experiment.

![Handoff Latencies - Lucent STA with Lucent AP](image)

### 3. DESIGN OF THE EXPERIMENT

As mentioned earlier, the experimental setup consists of two in-building wireless networks, a mobile wireless client, and a mobile sniffer system. As shown in figure 3, the basic methodology behind the experiments, is to use the sniffer...
(physically well within RF range of the client at all times) to capture all packets related to the client for the analysis.

**Wireless Network Environment:** All the experiments were done in the A.V. Williams Building at the University of Maryland, College Park campus. The building hosts two co-existing wireless networks namely *cswireless* and *umd*. The experiments were done in the overlapping coverage area of both networks. The *cswireless* network consists primarily of Lucent APs while the *umd* network consists of Cisco APs. The *cswireless* network density is approximately 6 APs per floor of the building while that of *umd* is approximately 8 APs per floor. The channel allocation for the networks is done so that there is no interference between adjacent APs i.e. the proper channels are set for the radio transmission and reception of APs so that no adjacent APs are using the same channel. In this experiment, channel 1, 6 and 11 are used for the wireless communication.

**Client Setup:** For the mobile station, we used *OpenBSD 8.1* on a HP OmniBook 500 with Pentium III 700 MHz and 384 MB RAM. The following wireless cards were used at the mobile station during the experiment: *Lucent Orinoco Gold*, *Cisco Airnet 340* and *ZoomAir Prism 2*.

The experiments were done in the following manner. A person with the mobile station walks through the building following a fixed path of travel (to minimize effects from the layout of APs) during each run. The duration of the walk, which is the duration of a single run of the experiment is approximately 30 minutes. Each experiment is characterized by the (i) Wireless NIC used at the mobile station and (ii) the Wireless network used. The mobile client sends negligible periodic ICMP messages to the network to maintain and display connectivity. Thus as the station moves, it performs handoffs as it leaves a BSS and enters another.

**Collection of Data:** During a handoff, a set of management frames such as *probe*, *authentication* and *reassociation* frames, are exchanged between the APs and the mobile station. By collecting every management frame from the RF medium (with timestamps) we compute the handoff delay as the interval between the first *probe request* frame and *reassociation response* frame (figure 2). Also the time spent for each phase such as probe, authentication and reassociation phase was obtained. This analysis is done offline.

In order to capture every management frame in the RF medium we designed a separate IEEE 802.11 sniffing system that is also mobile and in close proximity so that they
4. RESULTS AND ANALYSIS

Figure 9: Handoff Latencies - ZoomAir STA with Cisco AP

The wireless cards based on the Intersil Prism 2 chipset have a monitor mode [1] which enables applications to read raw IEEE 802.11 frames on one particular channel. Thus by capturing traffic from three cards (on channels 1, 6, 11), we are able to sniff all packets transmitted by participating entities in the common RF medium. Other approaches that use one wireless NIC and hop among channels, are bound to miss up to an upper bound of 66% of the traffic. During our experiments using the Cisco Aironet card to capture packets [9], we observed a miss-rate of around 30% (from experiments by sending parallel traffic on all three channels).

To sniff multiple channels, we set up two Linux machines, one with one wireless card and the other with two wireless cards which sniff three different channels independently. To preclude the inaccuracy caused by the inconsistencies of the system clock in the two machines, we synchronized their times using the Network Time Protocol (NTP) through an ethernet connection between the machines. Throughout the experiment, we maintained a clock accuracy of 80 µs or better between the machines (an error of less than 0.08% for latency of 100 µs). These Linux machines we used are IBM ThinkPad laptops with Pentium III 866 MHz and 256 MB RAM. A network sniffer program, etheral and Prism 2 wireless cards in Hostap mode are used for sniffing the IEEE 802.11 management frames.

4. RESULTS AND ANALYSIS

Figures 4, 5 and 6 show the handoff latencies for the three client cards (Lucent, Cisco, ZoomAir) with Lucent APs. The X axis is the handoff number (i.e. handoffs in order of occurrence) while the Y axis is the handoff latency breakup among the three delays. Figures 7, 8 and 9 show the results for the Cisco APs. Each graph is a single run of the experiment through the building. Below we itemize the conclusions from these results:

1. **Probe delay is the dominating component:** Looking at the six graphs, (figures 4,5,6,7,8,9) it’s clear that the probe delay accounts for more than 90% of the overall handoff delay, regardless of the particular STA, AP combination. Also even in the number of messages exchanged between the STA and the APs involved, the probe phase accounts for more than 80% of these in all cases. Thus any handoff scheme that uses techniques/heuristics that either cache or deduce AP information without having to actually perform a complete active scan clearly stand to benefit from the dominating cost of the scan process.

2. **The wireless hardware used (AP,STA) affects the handoff latency:** Looking at the differences in the Y scale among the six graphs, one can readily draw this conclusion. We can warrant this conclusion by observing two facts. Firstly, keeping the AP fixed, we can see that the client wireless card affects the latency. Figure 11 compares the average values of the latency among all six configurations. In each half of the figure (i.e keeping the AP fixed), we can see a maximum average difference of 335.53 µs (Lucent STA and Cisco STA with Cisco AP). This is a huge variation by just changing the client card being used. Secondly, keeping the client card fixed, the AP also affects the latency but to a much lower extent (around 50% less). This can be inferred by looking at figure 11 and noting that the maximum average difference (between the two APs

\^1 Host AP is a software implementation of AP functionality for Prism II wireless cards.
for any fixed client) is 186.47 ms (Lucent AP vs Cisco AP for ZoomAir STA) and a minimum of 32.64 ms.

3. **There are large variations in the handoff latency**: Apart from the variations in the latency with different configurations, we can see significant variations in the latency from one handoff to another within the same configuration. Figure 13 shows the maximum difference between any two latency values within a particular AP-STA configuration (i.e. the $|\text{Max} - \text{Min}|$). Also 12 shows the standard deviation values of the handoff latencies for all configurations. From these graphs, it is again clear that the particular AP-STA being used affects the extent of variations. For example, the maximum variation for a fixed AP (i.e maximum of $|\text{Max} - \text{Min}|$) happens between Lucent and ZoomAir STAs with Lucent AP (a variation of 178.96 ms). Also for the same STA, we can see that the Cisco APs have a lower variation (standard deviation) than the Lucent APs.

4. **Different wireless cards follow different sequence of messages**: This is an observation from looking at the traces offline. We found that the ZoomAir Prism 2 cards follow a slightly different procedure than the rest, as shown in figure 10. The figure shows that the card sends a reassociate message prior to authentication which it does when the AP sends a deauthentication message. The figure also shows the modified semantics of the reassociation delay and the authentication delay for the ZoomAir cards. This sequence of messages for a reauthentication, violates the IEEE 802.11 state machine as specified in the standard [4].

The probe delay being accountable for the high handoff latency and the variations, we present further analysis of this process. The probe is essentially an active scan (the wireless NICs do by default), and hence an analysis of the messages and latencies in the active scan is discussed below.

**The probe function**: The probe function is the IEEE 802.11 MAC active scan function and the standard specifies a scanning procedure as follows (modified for brevity):

For each channel to be scanned,

1. Send a **probe request** with broadcast destination, SSID, and broadcast BSSID.
2. Start a **ProbeTimer**.
3. If medium is not busy before the ProbeTimer reaches $\text{MinChannelTime}$, scan the next channel, else when ProbeTimer reaches $\text{MaxChannelTime}$, process all received **probe responses** and proceed to next channel.
Figure 15: The Effect of the Number of Probe Responses on the Probe-Wait Time (Cisco STA and Cisco AP) - Bar Graph

Figure 16: The Effect of the Number of Probe Responses on the Probe-Wait Time (Lucent STA and Cisco AP)

Figure 17: The Effect of the Number of Probe Responses on the Probe Delay (Lucent STA and Cisco AP)

Figure 18: The messages in an active scan.

Figure 18 shows the messages in a probe phase. The STA transmits a probe request message and waits for responses from APs on each channel. Let Probe-Wait latency be the time an STA waits on one particular channel after sending the probe request. We measure this as the time difference between subsequent probe request messages. Thus according to the above procedure, the traffic on the channel and the timing of probe response messages affects the probe-wait time, i.e. the probe-wait time should be expected to be distributed between a MinChannelTime and a MaxChannelTime value. For the probe phase analysis, we study two experiments with the following setup: (i) Cisco STA and Cisco AP, and (ii) Lucent STA and Cisco AP.

Figure 14 shows the various probe-wait times with respect to the number of probe response messages received by the STA. This plot is for the Cisco STA and Cisco AP combination. The scatter-plot shows two clusters being formed, which more-or-less correspond to the MinChannelTime and MaxChannelTime values from the above active scan procedure. This clustering is further elucidated in figure 15 which shows that the probe-wait time tends to be within 0 and 20ms for less than 2 probe response messages, otherwise it tends to be within a short interval of 35 to 40ms. Thus the number of probe response messages can create a difference of at least 15ms (and an average of 25ms) per channel. This explains the high standard deviation that we observe in the overall handoff latencies.

To compare with Lucent STA and Cisco AP combination, figure 16 shows the variation in the probe-wait time with the number of probe responses. This graph shows that there is a positive correlation between the two quantities, but also there is large variation in the probe-wait time for the same number of probe responses. Figure 17 shows the overall probe delay with the number of probe responses. Thus on the overall, for Lucent STAs and Cisco APs, there is a positive correlation between number of probe responses and the probe delay. Contrasting with the earlier experiment, we do not find any clusters that correspond to the MinChannelTime and MaxChannelTime values, but rather we find an almost uniform distribution. Also the probe-wait times for the Lucent STA are on an average less than the corresponding values for Cisco STA with respect to Cisco APs. Albeit the contrasting results, we can see a maximum difference of around 18 ms among the various probe-wait times for the same number of probe responses and a maximum difference of around 25 ms regardless of the probe responses. This vari-
5. CONCLUSION AND FUTURE WORK

The primary contribution of this work is a detailed analysis of the handoff process, the factors that bring about the high latency and the variation and the various messages/steps involved. We find that out of the three basic functions (probe, authentication and reassociation), carried out by the STA, the probe phase has the dominant latency regardless of the AP-STA being used.

We also present a detailed analysis of the probe phase, and account for the large variation to the probe-wait time which essentially depends on the particular heuristic employed by the wireless client NIC being used.

In our experiments we used wireless PC cards from three vendors, namely Lucent Orinoco, Cisco Aironet, and ZoomAir and the APs from Lucent and Cisco. This gives us enough diversity in our experiments and we find that there is large variation in the latency with the particular AP-STA hardware being used. Also we find that the sequence of messages exchanged during the handoff process can also differ with the STA being used.

One of the more interesting results of our work is that current WLAN equipment will not meet the expectations (replacing or augment 4G systems) that many have. This is because the handoff latencies we measured far exceed guidelines for jitter in voice over IP (VoIP) applications where the overall latency is recommended not to exceed 50ms [6].

In the future, we plan to investigate methods to add a robust authentication mechanism to WLAN handoffs and reduce the overall latency of the handoff within acceptable bounds for VoIP applications.

6. REFERENCES