Proactive Context Transfer and Forced Handover in IEEE 802.11 Wireless LAN Based Access Networks*

Ha Duong *Ha.Duong@postgrads.unisa.edu.au*
Arek Dadej *Arek.Dadej@unisa.edu.au*
Steven Gordon *Steven.Gordon@unisa.edu.au*
Institute for Telecommunications Research, University of South Australia, Australia

In recent years, many protocols have been developed to support user mobility in wireless networks, e.g. Mobile IP suite of protocols designed to support IP routing to mobile nodes. However, support for truly seamless mobility requires more than just routing: every service associated with the mobile user needs to be transferred smoothly to the new access network. In this paper, we will concentrate on the problem of transferring service state (context) at both the link and the IP layers. We propose a method to estimate the best moment in time for transferring context information associated with the mobile user. As one of key issues in the proactive context transfer scheme is accuracy of handover prediction, we suggest and describe a new concept of Forced Handover that can provide very low handover latency. The simulation results and the following discussions show that our scheme is helpful in ensuring seamless mobility, while keeping the number of unnecessary handovers resulting from the proactive nature of the scheme at a controllable level.

I. Introduction

The mobility of wireless users has created a number of technological challenges, especially when a Mobile Node (MN) changes the point of attachment to the network. In recent years, a great deal of research effort has been spent on the issue of mobility, and resulted in development of the general framework, as well as specific mechanisms and protocols supporting mobility. For example, the IETF Mobile IP Working Group (WG) has developed a solution officially named **IP mobility support**, and commonly known as **Mobile IP** [4].

Mobile IP and other mobility support protocols are intended to solve the problem of IP routing (i.e. finding the IP path) to the MNs. Typically, however, the access network may also need to establish and keep service state information (service context) necessary to process and forward packets in a way that suits specific service requirements, for example, Quality of Service (QoS) state or Authentication, Authorization and Accounting (AAA) state. Another example of context information is the header compression state established and maintained between an Access Router (AR) and the MN to reduce the large IP header overhead of short (e.g. voice) packets sent over a bandwidth-limited wireless link. To provide truly seamless mobility for real-time applications, both IP layer connectivity and the relevant context information have to be established or re-established as quickly as possible after a handover. However, the current research [7] indicates that it is impossible to re-establish both IP connectivity and service context within the time constraints imposed by real-time applications such as Voice over IP. Therefore, Context Transfer (CT) has been suggested as an alternative way of restoring the service context at the new access network.

As a means for CT, the IETF Seamoby WG is currently developing the Context Transfer Protocol (CTP) [8]. The CTP describes a simple way to transfer context information from the old AR to the new AR, to enable faster re-establishment of services. The CTP is expected to save time and bandwidth, and consequently to improve handover performance. However, even with CTP, many issues still must be solved in order to support specific services. For example, as CTP only specifies the transfer procedure between two ARs, it is not clear what CTP can or should do in cases when the service involves a number of other network entities. Unfortunately, most services such as AAA, QoS, or security, require participation of not only ARs, but also other network entities (e.g. authentication servers, resource managers). Therefore, some additional time will be required to re-establish service state after the new AR receives context information by means of CTP. This limitation can be considered a result of following a reactive CT approach where

*A part of this work was presented at the 2nd ACM International Workshop on Wireless Mobile Application and Services on WLAN Hotspots (WMASH 2004) under the title "Proactive Context Transfer in WLAN-based Access Networks".

the context transfer takes place immediately after handover, and provides a good motivation to examining the alternative proactive approach to CT, i.e. transferring the necessary context before the handover is needed or occurs.

I.A. Related Works on Proactive Context Transfer

Even though CT can reduce the delay of context reestablishment, the resulting delay is still a significant component of handover delay. Therefore, recently, a number of researchers have become interested in proactive CT. Pagtzis [11] suggested a proactive IP mobility model where MN's IP connectivity and other context are established at the new point of attachment in advance of the actual handover (transition between points of attachment). The key point in this model is the Mobility Neighbour Vector (MNV) - Routing Neighbour Vector (RNV) mapping. The MNV represents a collection of cells within the neighbourhood reachable from the current cell; while RNV is a collection of routers associated with MNV. Discovery of the MNV-RNV mapping is achieved incrementally by means of dynamic learning i.e. MN's handover transitions between Access Points (AP) and ARs. While Pagtzis' work focused on proactive mobility at the IP layer level, Mishra [2] focused on proactive context caching at the link layer level. In the link level model, after the MN associates with an AP, the AP will forward MN's context information to neighbour APs. Each AP learns about its neighbours through previous re-associations of MNs.

The shortcoming of the above works is that the authors did not consider the waiting time of the transferred context at the new access network. Timing is an important aspect in CT, especially in the case of QoS context. If QoS context is transferred too early in respect to handover, resources held (reserved) in the neighbouring access networks will be wasted until the MN re-establishes IP connectivity at its new point of attachment. Therefore, it is desirable that context information is transferred as close as possible to the handover time.

I.B. An Overview of Our Work

In this paper, we propose a proactive scheme for CT that attempts to estimate the best moment for proactive CT. With proactive approach, the typical question is: when should the proactive context reestablishment start? The starting time should be directly related to the time the handover occurs, which in turn depends on handover prediction mechanism. In mobile networks, a handover typically occurs when a new point of attachment offers a better service (e.g. communications link, error rate) than the current point of attachment. Handover prediction techniques try to estimate when a better point of attachment will become available. If handover prediction and proactive context reestablishment are used together, the context reestablishment may be a waste if the handover prediction fails. On other hand, it is desirable that the time between the context reestablishment and the actual handover is as short as possible; otherwise resources will be wasted at the new access network. In other words, the proactive CT should start and be completed as close as possible to the time of actual handover. The above requirement has led us to propose a new concept of Forced Handover i.e. a handover forced to occur at a "planned" moment. The philosophy of the Forced Handover is that instead of guessing the handover time, we force the handover at a planned time. With Forced Handover, instead of wasting resources in case of handover prediction failure, we may experience unnecessary handovers. However, the number of unnecessary handovers can be kept at a controllable level, as will be shown later in our analysis. At the expense of some unnecessary handovers, the Forced Handover will enable proactive CT and reduce handover latency.

In short, our contributions in this paper are:

- A new proactive scheme for integration of Mobile IP and its enhancements, Context Transfer Protocol and Candidate Access Router Discovery (CARD) protocol [10].
- Forced Handover as a key component of the proactive scheme.
- Demonstrated feasibility and effectiveness of the proposed scheme in IEEE 802.11 WLAN-based access networks.

As WLANs, particularly IEEE 802.11 are the most popular access technology in wireless Internet due to low deployment cost and hight date rates, in this paper we will examine proactive CT and Forced Handover in WLAN-based access networks. As part of future work, we intend to investigate these techniques in other types of access networks (e.g. cellular networks).

The rest of the paper is organized as follows. In the next section, we provide background information including handover at WLAN MAC layer, the IETF proposed Mobile IP standard and its enhancements, CTP

and CARD protocols. Our contributions, proactive CT and Forced Handover, are described in section III. Then, in the following two sections, we discuss performance metrics useful in evaluating our proposed scheme, with the emphasis on imperfect handovers and handover latency reduction. We also present results from the analysis of the proposed scheme (via simulation), and subsequent discussions. Finally, we make some concluding remarks and comments on the areas of research intended for future work.

II. Overview of Handover Process in IEEE 802.11 WLAN based Access Networks

A handover between points of attachment may occur at the link layer level or at the network layer level. In all cases, the change should be transparent to the user. In this section, after defining the terminology used throughout this paper, we provide background on how handovers may be performed in a WLAN-based access networks. This includes a description of the IEEE 802.11 MAC (link layer) handover procedure, as well as the key protocols in performing a network level (network layer) handover, Mobile IP, Context Transfer Protocol and Candidate Access Router Discovery Protocol.

II.A. Terminology

In this paper we define the network entities and handover concepts as follows:

- **Access Point (AP)** - A radio transceiver via which a MN obtains link layer (Layer 2) connectivity to the access network. An AP is typically a bridge between the wireless link and a wired link.
- **Access Router (AR)** - An IP router residing in an access network, connected to one or more APs, and offering IP (Layer 3) connectivity to the MN.
- **Inter-AP handover** - The process of switching (handing over) from one AP to another. This is a Layer 2 handover - in the IEEE 802.11 standard, this process is called *re-association*.
- **Inter-AR handover** - The process of switching from one AR to another AR, i.e. a Layer 3 handover.

Following from these definitions, an inter-AP handover will only result in an inter-AR handover when the old AP and the new AP are not connected to the same AR. As an example, Figure 1 illustrates a MN roaming in an area served by AR1 and AR2. The MN encounters an inter-AP handover when it moves from the cell served by AP1 to the cell served by AP2. In contrast, when the MN performs an inter-AP handover from AP2 to AP3, an inter-AR handover from AR1 to AR2 is also performed.

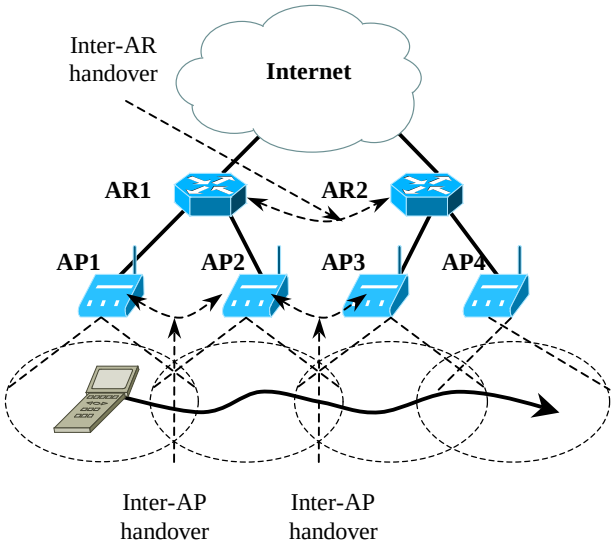

Figure 1: Inter-AP handovers and Inter-AR handovers.

II.B. Inter-AP handovers in 802.11 WLANs

The inter-AP handover process in a WLAN is a sequence of events occurring between APs and MN, and resulting in a switch of physical and link connectivity from one AP to another. In IEEE 802.11, there are four key components of the handover process: the handover trigger; AP discovery and selection; (re-)authentication; and (re-)association (Figure 2). For brevity, we will discuss the first two components as they are most directly related to handover performance improvement offered by our proposed scheme. Readers can refer to the IEEE 802.11 standard [3] for details of the two others.

II.B.1. Handover Triggers

In general, handover triggers are obtained from observations that the link with the current AP is deteriorating. These observations could include failed retransmissions of packets, missing beacons, or degradation of Signal-To-Noise Ratio (SNR) below a threshold. In addition, handover triggers may be based on service parameters other than link quality. Load-balancing between APs is an example. If the current AP is

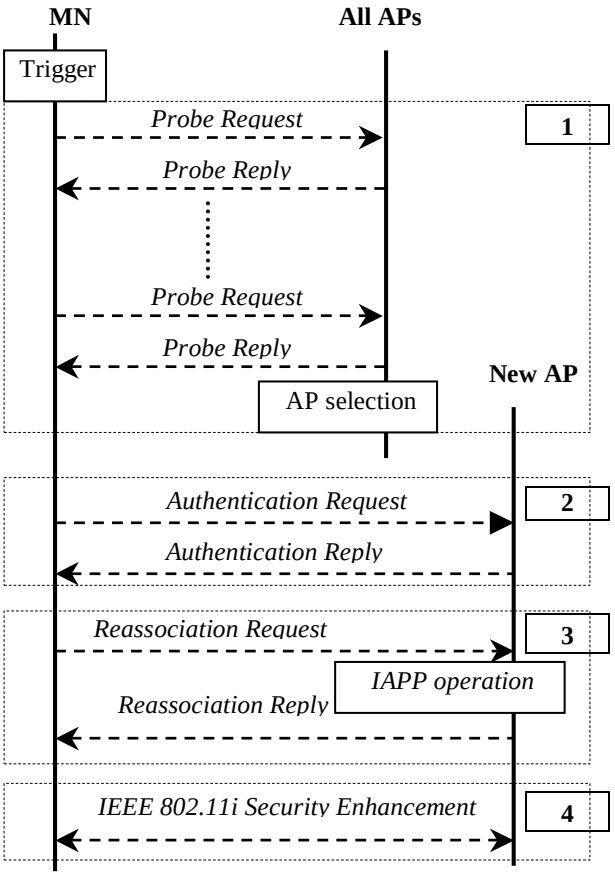

Figure 2: Inter-AP handovers process.

highly loaded, while another AP is serving a lighter load, then a handover to the lightly loaded AP may be triggered. Ideally, a combination of factors would be used to trigger a handover such that a satisfactory level of service is maintained in the network. In practice, there are various different implementations, but nearly all include at least some link quality factors in triggering handovers.

II.B.2. AP Discovery

Whenever a handover trigger occurs, the MN starts the AP discovery process (box 1 in Figure 2) by cycling through the possible radio channels in search for a new, more suitable AP. The IEEE 802.11 standard [3] specifies two types of scanning, namely passive and active. In the passive scanning mode, the MN switches to a new channel, waits to receive beacons for a period of time $T_{channel}$, and then switches to the next channel. Normally, $T_{channel}$ is set slightly greater than twice the beacon interval (T_{beacon}) so that the MN can receive two beacons on a channel. After cycling through all available channels (possibly several times), the MN can select a new AP based on information gathered from the beacons (discussed below).

In the active scanning mode, the MN follows the above process of cycling through channels but instead of waiting for beacons, the MN sends a Probe Request frame and waits for a Probe Response from an AP. Once a Probe Request has been sent, the MN starts a timer, and if there is no activity on the channel within a given time ($T_{MinChannel}$), the MN switches to the next channel. If the channel has activity, the MN waits until time $T_{MaxChannel}$, processes all received Probe Response frames and then scans the next channel. The MN uses information collected during the scanning process to select a new AP.

The selection of an AP to handover to is typically based on the same factors as the trigger, i.e. the quality of service offered by the AP. In most implementations (e.g. [9]), quality of the communication link, i.e. signal strength or signal-to-noise ratio (SNR), is used to make the handover decision, although in some other implementations (e.g. [5]) the current load on the APs is also taken into account.

In both active and passive scanning modes the scanning cycle is repeated every Scanning Interval (T_{SI}) until the MN eventually finds a new AP better than the current one. As an example, Figure 3 shows the change in SNR at two APs as measured by a MN (which is moving away from AP1 and toward AP2). In this case, SNR is used as both the trigger for AP discovery (SNR_1 goes below a cell search threshold, SNR_{CST}) and the criterion for AP selection and handover initiation (the SNR of a new AP must be greater, by the threshold Δ, than the SNR of the old AP). At point 1, the cell search threshold is satisfied, triggering the AP discovery process. The MN repeats the scanning cycles until it finds that AP2 provides a better SNR than the current AP by the amount of positive hysteresis Δ (point 4 in Figure 3).

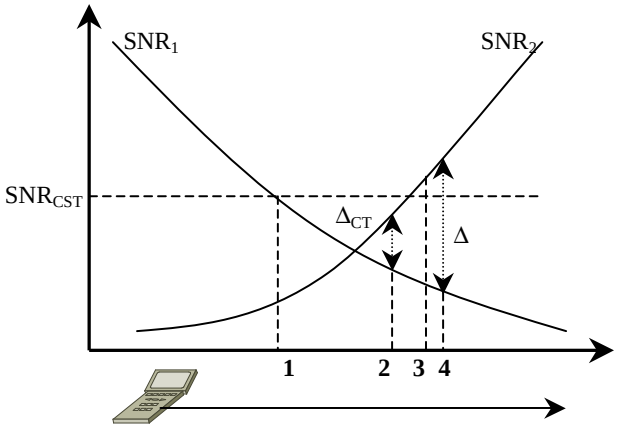

Figure 3: SNR change between AP1 and AP2

In summary, the condition for the inter-AP handover is as follows

$$\begin{cases} SNR_1 < SNR_{CST} \\ SNR_2 > SNR_1 + \Delta \end{cases} \quad (1)$$

II.C. Inter-AR handovers

Many proposals have been developed to support the routing process during a handover between ARs, with Mobile IP [4] being the most prominent approach. Here we will give a brief overview of Mobile IP. In addition to routing, support is needed to expedite the discovery of potential new ARs and transfer service state to those ARs selected. Therefore, following the Mobile IP overview, we introduce solutions to AR discovery and transfer of service state, i.e. the Context Transfer Protocol and the CARD protocol.

II.C.1. Mobile IP

The main characteristics of Mobile IP[1] include transparency to applications and transport layer protocols, scalability, and macro mobility. Mobile IP introduces two new entities into the network, Home Agent (HA) and Foreign Agent (FA). These two entities can reside anywhere within the subnet where they are serving. For simplicity, we assume that they co-locate with ARs. Mobile IP includes two main functions, registration and tunnelling. For the purpose of this paper, we will briefly describe the registration procedure as follows.

Whenever a MN discovers that it is moving into a new subnet, it sends a Registration Request message, which includes the MN home address, care of address and HA address, to the new FA. The new FA relays the message to the HA after retrieving information necessary for serving the MN in the future. In response to the Registration Request, the HA will send a Registration Reply to the new FA. In turn, the FA sends the Registration Reply to the MN to confirm (or reject) the MNs registration at the new FA. Following this registration procedure, all communications to and from the MN go via the new FA.

II.C.2. Context Transfer Protocol

The objective of the Context Transfer Protocol (CTP) [8], developed by the IETF Seamoby Working Group, is to transfer the service state between ARs to enable seamless mobility. CTP has a request-response mechanism for transferring the service state (or context), as well as mechanisms for triggering the CT and activating the context once at the new AR. The protocol can be initiated by either MN or AR, depending on the CT trigger. The CT trigger is still an open issue as it depends on specific link layer technology. As shown later in section III.C, our proactive scheme will use the condition from Eq. (3) as the CT trigger. In network-initiated scenarios, if the CT trigger is detected at the old AR, this AR will send the CT Data (CTD) to the new AR; otherwise the new AR will request the old AR to transfer context (CT Request). Upon receiving CTD, the new AR optionally may reply back to the old AR (CTDR - CT Data Reply). In both cases, the MN will send the CT Activation Request (CTAR). In mobile-initiated scenarios, the MN will send the CTAR upon receiving a CT trigger, usually from the link layer. Then, the new AR can request CT from the old AR.

Several issues arise when applying the CTP to specific services. For example, the CTP is insufficient in case of services involving network entities other than ARs. Intuitively, reestablishment of these services will require more time; hence reactive reestablishment may not be well suited to real-time applications. This limitation provides a good motivation to considering an alternative, namely the proactive approach to CT. In this approach, potential ARs for handover have to be discovered before the proactive CT can be carried out.

II.C.3. CARD Protocol

The purpose of the Candidate Access Router Discovery (CARD) protocol [10], another draft resulting from the work of the IETF Seamoby WG, is to identify (discover) the IP addresses of candidate ARs (CARs) for handover, and to discover their capabilities. Our proactive scheme will make use of the first CARD function mentioned above which, by CARD recommendations, can be implemented in either centralized or decentralized manner. The result of address mapping is included in the CARD Reply message that is sent back to the current AR. The reader should refer to [10] for more details of the mapping scheme.

As mentioned earlier, three protocols, Mobile IP, CTP and CARD are expected to work together to facilitate seamless handover. Our contribution offers a way to combine these three protocols into a proactive CT scheme. To ensure smooth operation of the proactive scheme, we also suggest the concept of Forced

[1] In this paper, we use Mobile IPv4 to demonstrate the feasibility of our proposal, but we believe that the proposal is also applicable to Mobile IPv6 (however that is out of the scope of the paper). In the remainder of the paper, the term Mobile IP refers to Mobile IPv4.

Handover. We will describe these proposals in detail in the next section.

III. Proposed Scheme for Proactive Context Transfer in Forced Handover

In section I, we stated that proactive CT should start and be completed as close as possible to the time of actual handover. To do this, we propose a method of estimating the time when the proactive CT should occur (section III.A), as well as the concept of Forced Handover (section III.B), where we take advantage of a predicted handover at the expense of incurring a small number of unnecessary handovers. In section III.C, we describe the complete proactive process, including CTP and CARD operations.

III.A. Estimation of Proactive Context Transfer Time

Assuming CT can be completed within a scanning interval, the best time to start the CT is at the scanning interval closest to (but before) the actual handover. We propose the following procedure to estimate which scanning interval (or cycle) the CT should start at. We also assume that SNR is used for triggering and AP selection because of its wide acceptance as a handover criterion.

When in the cell-search state, after every scanning cycle, the MN estimates the time until handover as follows

$$T_{until\_handover} = \frac{\Delta - (SNR_2 - SNR_1)}{R_{SNR_2} - R_{SNR_1}} \qquad (2)$$

where R_{SNR_1} and R_{SNR_2} are rates of SNR change for signals from the current AP and the scanned AP respectively. These rate values are obtained and updated on the basis of SNR measurements performed as part of the current and previous scanning cycles.

If the $T_{until\_handover}$ is less than or equal to the T_{SI} (point 3 in Figure 3), the current scanning cycle is likely to be the second last (now called *scanning-to-CT*), and in the next scanning cycle (now called *scanning-to-handover*), the handover condition is likely to be satisfied. In short, the MN identifies the scanning-to-CT by

$$T_{until\_handover} \leq T_{SI} \qquad (3)$$

To reduce computations, the MN may start to estimate the $T_{until\_handover}$ when the following condition is satisfied

$$\begin{cases} SNR_1 < SNR_{CRT} \\ SNR_2 > SNR_1 + \Delta_{CT} \end{cases} \qquad (4)$$

where Δ_{CT} is less than Δ.

Δ_{CT} (point 2 in Figure 3) should be selected such that there is at least one scanning cycle before scanning-to-handover; therefore it can be defined from the following formula

$$\frac{\Delta - \Delta_{CT}}{R_{SNR_{2max}} - R_{SNR_{1max}}} = T_{SI} \qquad (5)$$

where $R_{SNR_{1max}}$ and $R_{SNR_{2max}}$ are maximum rates of SNR change from the current AP and the scanned AP. The rate values of interest can be learnt (estimated) from previous measurements, or pre-set.

III.B. Forced Handover

The above technique can produce a good estimate of the time for proactive CT (scanning-to-CT) and the time for handover (scanning-to-handover). This will be later confirmed by simulation. However, there is not a 100% guarantee that the handover condition in Eq. (1) will be satisfied at the time of scanning-to-handover. One can argue that if the Eq. (1) is not satisfied at the time of scanning-to-handover, the MN may wait until the next scanning. However, in this case we need to set up a longer waiting time for the transferred context at the new AP (if inter-AP handover) or new AR (if inter-AR handover), and consequently there may be more resources wasted. We suggest a Forced Handover, i.e. **the MN will make the handover after the scanning-to-handover time is reached, regardless of whether the handover condition in Eq. (1) is satisfied or not.** The main advantage of Forced Handover is that the MN knows exactly when the handover will happen, and therefore can set up an appropriate waiting time for the transferred context at the new access network. The Forced Handover at the link layer level also allows the MN sufficient time to prepare for the IP level handover. For example, the MN may use the Forced Handover as a trigger to send an Agent Solicitation message [4] to acquire the Router Advertisement message [4] from the new FA; therefore it can reduce the Agent Discovery time (T_{AD}) to as low as one Round Trip Time (RTT) between the MN and the new FA. The T_{AD} can be further reduced if the MN is able to receive the information needed for registration with the new FA through the CARD Reply message (more details will be given in the next subsection). This information enables the MN to create a Registration Request message [4] in advance and send it immediately after the MN re-establishes the link level connection to the new access network. The shortcoming of Forced Handover is that in some cases,

handover is forced to happen when the handover condition Eq. (1) is not yet satisfied; therefore, the number of unnecessary handovers may increase. However, this number can be kept at a reasonably low level as shown in simulation results (see Section V).

III.C. Description of the Proactive Process

Now we will describe the proactive scheme for CT. Assume that MN is moving into an area where the SNR from the current AP drops below the SNR_{CST} (see Figure 3 and box 1 Figure 4):

1. The MN starts a scanning cycle every T_{SI} seconds (i.e. box 2 Figure 4) until the condition Eq. (4) is satisfied.

2. The MN starts estimation of the $T_{until\_handover}$ and continues scanning cycles until at least one of the scanned APs satisfies $T_{until\_handover} \leq T_{SI}$ (box 3 Figure 4).

3. The MN collects L2 addresses of scanned APs satisfying the condition Eq. (3) (now we call them candidate APs), and sends them to the current AR via a CARD Request message.

4. Upon reception of the CARD Request message, the current AR resolves address mapping as described in section II.C.3 and identifies the expected handover type i.e. inter-AP or inter-AR (box 4 Figure 4). If the expected handover is inter-AP, the AR instructs the current AP to send the L2 context information to the new AP as specified in the IEEE 802.11 Inter-Access Point Protocol [6]; otherwise, the AR sends the CT Data message to the neighbour ARs. In case of the inter-AR handover, the current AR may ask the new AR to provide information necessary for the MNs registration with the new FA. Normally, this information is available via the Agent Advertisement message [4] broadcasted periodically by FAs. Now, the current AR informs the MN about the result, including the handover type, the selected AP, the selected AR, and the registration information (if inter-AR handover) via the CARD Reply message.

5. In the next scanning cycle (box 5 Figure 4), the MN performs handover at the link layer level to the AP specified in the CARD Reply message. If the inter-AR handover and the registration information are specified in the CARD Reply message, the MN creates a Registration Request message [4] and sends it to the FA immediately upon the reestablishment of link layer connectivity.

   Now, assume that the handover is inter-AR

6. When the MN gets connected to the new AR i.e. has received the Mobile IP Registration Reply message, it sends the CT Activate Request (CTAR) message to activate the transferred context at the new AR.

7. Upon receiving the CTAR, the new AR starts context reestablishment at the new access network and sends the CT Cancel (CTC) to the current AR (now the old AR).

8. Upon receiving the CTC, the old AR takes appropriate action to delete old context.

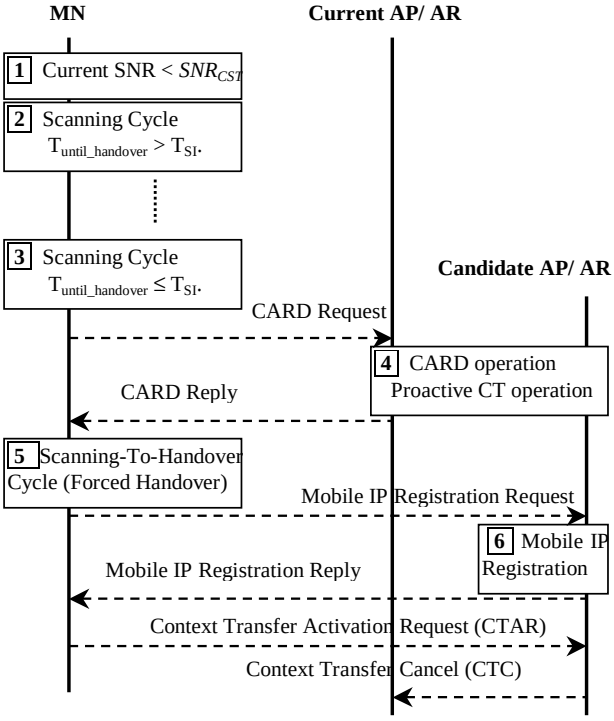

Figure 4: Time diagram of the proactive CT scheme.

Table 1: Scenarios of inter-AP handovers and inter-AR handovers.

	Current AR	Neighbour ARs	Handover Type
One candidate AP (CAP)	CAP	None	Inter-AP
	None	CAP	Inter-AR
Two or more CAPs	All CAPs	None	Inter-AP
	At least one CAP	Other CAPs	Inter-AP*
	None	All CAPs	Inter-AR

*An inter-AR handover is also possible but an inter-AP handover is preferred.

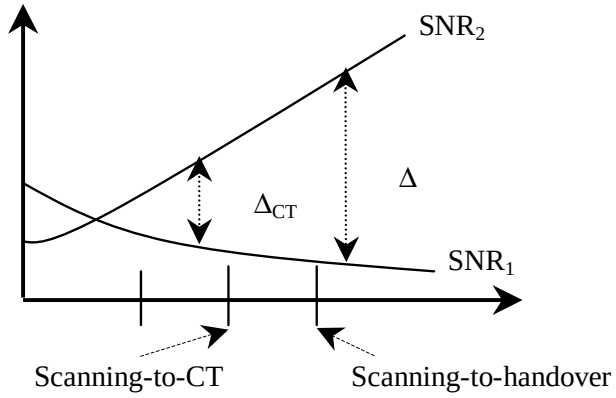

Figure 5: Time diagram of the premature handover.

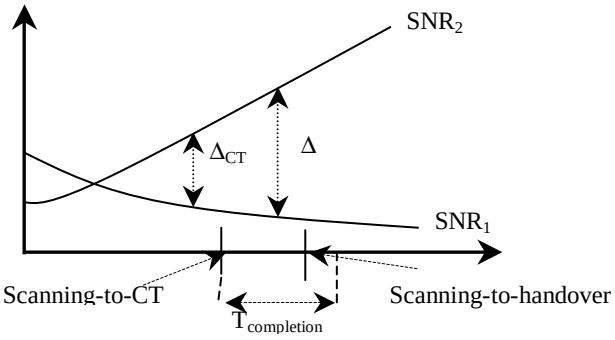

Figure 6: Time diagram of proactive scheme incompletion.

In the step 4, there may be a number of different scenarios, depending on how many APs satisfy the condition Eq. (3) (i.e. number of candidate APs). We summarize these scenarios in Table 1.

IV. Metrics for Performance Evaluation

The underlying design trade-off in our proposed scheme is to reduce the handover delay, at the expense of increasing the number of unnecessary handovers. Intuitively, the handover delay is reduced by discovering the potential AP/AR (using CARD) and transferring the necessary context (using CTP) in parallel with the WLAN handover process. This avoids delays due to CT and agent discovery after the WLAN handover process is complete, therefore reducing total handover time. On the other hand, predicting and forcing the handover results in an increase in the number of untimely and unnecessary handovers (due to the fact that the predictions may be wrong). Such handovers lead to a waste of resources (processing, memory) as well as additional handover delays (e.g. a second handover may be needed to compensate for the first erroneous handover). In order to evaluate our proactive CT scheme, we attempt to quantify this performance trade-off.

In this section, we first look at potential drawbacks of the scheme, namely the waste of resources due to unnecessary handovers and incomplete CT, and analyse these effects in further detail, providing evidence of the overall benefits of proactive CT and Forced Handover. Finally, we discuss how the forced handover can reduce handover latency.

IV.A. Unnecessary Handovers

The unnecessary handover can be defined as a handover that should not happen (but actually happened because of forcing). The indication of a possible unnecessary handover is that it is a "premature" handover as illustrated in Figure 5. In the premature handover, MN estimated that $T_{until\_handover}$ is less than or equal to T_{SI} at the time of the scanning-to-CT, but eventually found that the handover condition Eq. (1) is not satisfied yet at the time of scanning-to-handover. As the MN forces handover to occur anyway, we would like to see whether this Forced Handover is necessary or not i.e. whether the MN would eventually perform a handover in the near future. The unnecessary handover happens when the MN moves in such a way that the handover condition Eq. (1) is never satisfied. For example, the MN changes the direction of movement or stops after the scanning-to-CT cycle.

IV.B. Proactive Scheme Completion

Recall from the description of the proactive process in section III.C, that at the time of scanning-to-CT, the MN sends the CARD Request message to trigger proactive CT. After that, the MN waits for the CARD Reply message from the current AR as an indication of proactive scheme completion. If the MN does not receive this message until the scanning-to-handover (Figure 6), the proactive process can be seen as incomplete. Incomplete proactive process has the following effects:

1. If there is more than one candidate AP, the MN has to decide which candidate AP to switch to in the scanning-to-handover. The decision would be based on results obtained from the scanning-to-CT (e.g. SNR measurements). As the current AR is unsure whether the MN received the CARD Reply message, it has to initiate CT to all candidate APs; therefore there will be resources wasted at those candidate APs to which the MN

will eventually not connect. To reduce this waste, the new AR or AP, upon reestablishment of connectivity with the MN, should notify the candidate APs to release "unused" context. If there is only one candidate AP, there is no problem of wasting resources. Therefore, we may be interested in finding the probability of having only one candidate AP.

2. Upon link reestablishment, the MN has to send the Agent Solicitation message in order to reduce the Agent Discovery time (T_{AD}), as it does not know the expected handover type (i.e. inter-AP or inter-AR). If the expected handover is inter-AP, sending of the Agent Solicitation message will waste bandwidth.

3. In the case of inter-AR handover, T_{AD} can be reduced only to the RTT between the MN and the new AR, not to zero, as explained earlier in section III.B.

In summary, in the case of incomplete proactive scheme, there would be wasted effort of CT if there is more than one candidate AP, wasted bandwidth for sending the Agent Solicitation message, and limited reduction of Agent Discovery delay.

IV.C. Reduction of Handover Latency

Normally, the handover latency in WLAN-based access networks includes two elements: latency of link switching (or link handover), and latency of network layer handover. In this section, we explain how the Forced Handover can reduce these delays, in particular the Probe delay and Agent Discovery time.

The overview of inter-AP handover in IEEE 802.11 WLANs reveals three main factors that contribute to link handover latency: Probe Delay ($T_{ProbeDelay}$): the time between a MN initiating AP discovery and the MN selecting a new AP within a scanning cycle, Authentication Delay ($T_{Authentication}$), and Reassociation Delay ($T_{Association}$).

The two factors that contribute to network layer handover latency are Agent Discovery and Mobile IP Registration:

1. Agent Discovery (T_{AD}): The time required for the MN to discover that it has moved to a new subnet. Like AP discovery, T_{AD} depends on interval of sending Agent Advertisement messages, $T_{ADV\_interval}$. As $T_{ADV\_interval}$ is constrained due to bandwidth consumption, the MN may require information of inter-AP handover to speed up discovery by sending the Agent Solicitation message. Upon reception of this message, the new FA should respond with an Agent Advertisement message. Such approach reduces T_{AD} to the Round Trip Time (RTT) between the MN and the new FA that is normally much smaller than $T_{ADV\_interval}$.

2. Registration delay mainly consists of packet transmission delay from the MN to HA via new FA. The processing time at the new FA and HA is quite small, and normally can be ignored.

Experimental results have shown that Probe delay is the dominating component, accounting for more than 90% of link handover delay [1]. The main reason here is that the MN has to scan every channel from total N channels (for FCC regulatory domain, applied in North American, N = 11), and the total probe delay would be bounded by

$$T_{MinChannel}N \leq T_{ProbeDelay} \leq T_{MaxChannel}N \quad (6)$$

where $T_{MinChannel}$ and $T_{MaxChannel}$ are two parameters that determine the duration of scan for each channel. To reduce $T_{ProbeDelay}$ we need to selectively scan some of N channels (selective scan) instead of scanning all N channels (full scanning), based on information about channel allocation at neighbour APs. In [2] two algorithms for selective scan were presented. In the Forced Handover, there is no need to scan APs at the handover moment ($T_{ProbeDelay} = 0$) as the scanning process has been completed in previous cycles, and the decision of AP selection has been made based on estimation. As authentication context information is to be proactively transferred immediately after the scanning-to-CT cycle, authentication delay[2] is expected to show a significant reduction.

Forced Handover can also significantly reduce agent discovery time T_{AD}. Recall from the description of the Forced Handover, that the MN can detect a new FA and obtain information needed for registration via the CARD Reply message; hence T_{AD} is reduced to zero. In the worst case i.e. the MN missed the CARD Reply message, the MN can still follow the approach of L2 trigger (sending Agent Solicitation message). In this case, T_{AD} is equal to RTT between the MN and the FA. We will present numerical results of handover latency reduction in section V.B.

[2]We are currently investigating proactive key distribution for AAA Context Transfer.

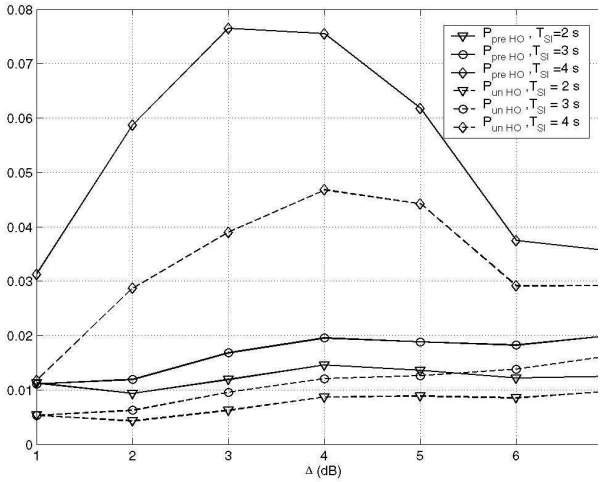

Figure 7: Probability of premature handovers (P_{pre_HO}) and probability of unnecessary handovers (P_{in_HO}).

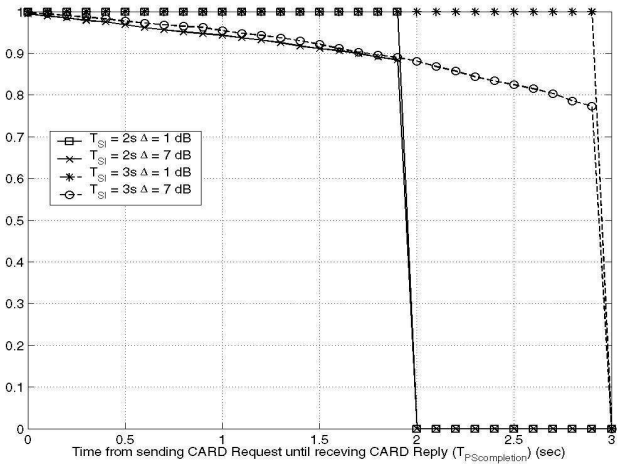

Figure 8: Probability of proactive scheme completion ($P_{PS\_completion}$).

V. Simulation Results and Discussions

In this section, we describe the simulation scenario, present simulation results and follow up with discussions. The simulation objective is to investigate the performance metrics discussed in the previous section, i.e. the probability of unnecessary handovers (P_{un_HO}), the probability of proactive scheme completion ($P_{PS\_completion}$), and handover latency. Firstly, using MATLAB we evaluate P_{un_HO} and $P_{PS\_completion}$. Secondly, we will show via OPNET simulations that our proposed scheme can reduce significantly handover latency.

V.A. Unnecessary Handovers and Proactive Completion Scheme

In the MATLAB-based simulations, the simulated area is covered by 61 APs distributed uniformly at a distance of 200m from each other. Transmission power of all APs is the same and at such level that there is no gap between coverage areas. The simulation area is assumed an open outdoor environment, therefore we can limit radio propagation model to path loss modelling. As future work, we intend to investigate more complex scenarios of semi-open or indoor environments i.e. we have to take into account fading channels or obstructions. Every AP, excluding the APs residing close to the edges, has 6 neighbour APs. In the simulation area, the MN is moving according to the random waypoint model as follows. After randomly selecting a destination, the MN moves towards the selected destination with a constant velocity v (the velocity v is randomly selected from a range of (0.5 m/s - 5 m/s)). After reaching the destination, the MN stops for the duration of pause time and then selects another destination and speed and moves again. The MN is always associated with an AP, and keeps monitoring SNR with this associated AP. As soon as the SNR drops below the threshold SNR_{CST}, the MN starts to follow the procedure described in section III.C. Such scenario was repeated in simulation 10000 times to ensure that collected data are statistically valid.

The target performance parameters are investigated in the context of different scanning intervals T_{SI} and hysteresis Δ. As we will see later, smaller T_{SI} usually give better performance. It is expected, since with smaller T_{SI}, the estimation of $T_{until\_handover}$ is performed more frequently, hence produces more accurate predictions. On the other hand, smaller T_{SI} mean that the MN has to interrupt the current communications more frequently in order to perform scanning.

Figure 7 presents the probability of premature handovers (P_{pre_HO}) and probability of unnecessary handovers (P_{un_HO}). As can be seen from the graphs, P_{un_HO} is always below the upper bound P_{pre_HO}, and remains under 1% when the scanning interval T_{SI} = 2 sec, and under 1.5% when T_{SI} = 3 sec. As P_{un_HO} dramatically increases with the T_{SI} = 4 sec, the scanning interval should be selected to be no more than 3 sec. It is also observed that the P_{un_HO} is lower with smaller T_{SI}. The reason, as discussed earlier, is that more accurate estimation results from smaller T_{SI}. However, it is undesirable to select too small a T_{SI} that will significantly affect transmissions carried out by the MN.

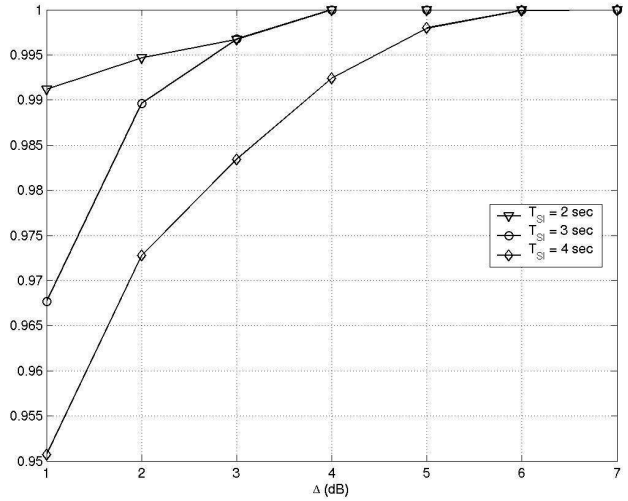

Figure 9: Probability of having one candidate AP (P_{one_AP}).

Figure 8 depicts the $P_{PS\_completion}$ when Δ = 1 dB and 7 dB, for two cases of scanning interval T_{SI} = 2 sec and 3 sec. We also obtained values of $P_{PS\_completion}$ with other values of Δ between 1 and 7 dB. Those values fall within the two graphs included in the figures and were not included to improve the readability of the figure. For small Δ such as 1 dB, the $P_{II\_PS\_completion}$ is 100% as long as the completion time of the proactive scheme $T_{completion}$ is less than T_{SI}. Small Δ implies that handovers occur in the area well covered by the current AP; therefore the completion of proactive scheme may only be prevented by the sequence of events whereby by the time the current AR sends the CARD Reply, the MN has already switched to the new AP.

Now we turn our attention to the probability of a given number of candidate APs. For instance, we are interested in probabilities of having one candidate AP (P_{one_AP}) (Figure 9).We note that these probabilities depend on positioning of APs in the simulation area. Firstly, we observe that higher P_{one_AP} results from larger value of hysteresis Δ. Larger Δ implies that the MN is likely to be close to only one candidate AP. Secondly; we also observe that higher P_{one_AP} results from smaller T_{SI}. However, the differences between the values of P_{one_AP} obtained with different T_{SI} are not great, and can be explained by the fact that smaller T_{SI} gives more accurate estimations.

The simulation results and the discussion lead to the following conclusion. With appropriate scanning interval T_{SI}, the probability of unnecessary handovers can be kept as low as 1% or 1.5%. At the expense of some waste of resources resulting from the unnecessary handovers, the remaining majority of handovers can be very accurately predicted; therefore, the MN has sufficient time to prepare for the handover. The results also reveal that the main factor preventing the completion of proactive scheme is switching to another AP too early, i.e. situation arising when the scanning interval T_{SI} is too short. In the simulations, the T_{SI} was selected from the range 2 - 3 s, and we believe that the proactive scheme can be completed within such scanning interval. Even when the proactive scheme fails to complete, i.e. when the MN is unable to receive the CARD Reply message, the proactive scheme still derives benefits, as context data is transferred to candidate APs and ARs. The disadvantage of not being able to complete the proactive scheme, as mentioned in the previous subsection, is that the MN does not know the expected handover type, therefore, there would be wasted bandwidth for sending the Agent Solicitation message, and limited reduction of Agent Discovery delay. In the case of having more than one candidate AP, some efforts to transfer context data to candidate APs would be wasted.

However, the second problem is insignificant because of the high probability of having one candidate AP (P_{one_AP}). In the case of having one candidate AP, the MN still knows the AP to switch to despite missing the CARD Reply message. However, after re-association, the MN will still need to send an Agent Solicitation message to discover whether it has moved to a new subnet.

V.B. Handover Latency Reduction

We have used OPNET (www.opnet.com) with WLAN MAC and Mobile IP models from the OPNET library for our simulation. We have added active scan and Forced Handover features as they were not implemented in the original OPNET WLAN MAC model. The modification also includes making information about Forced Handover at the WLAN MAC layer available to the Mobile IP module. Table 2 lists the parameter values used in the simulation. These are based on Cisco 340 WLAN card specifications.

Table 2: Simulation Parameters

Layer	Parameters	Value
Wlan MAC	Number of Channels	11
	Beacon Interval (T_{beacon})	0.1 sec
	Channel Time	0.25 sec
	Min Channel Time ($T_{MinChannel}$)	0.017 sec
	Max Channel Time ($T_{MaxChannel}$)	0.038 sec
Mobile IP	Agent Advertisement Interval	3 sec

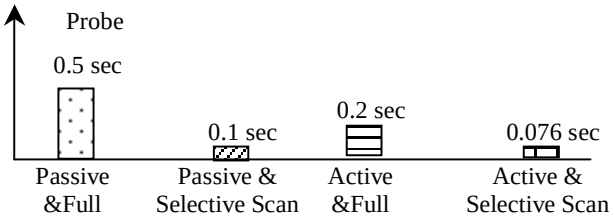

Figure 10: Probe delay in various types of scanning.

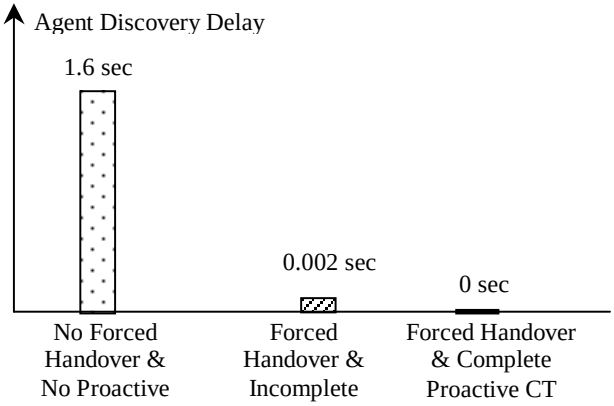

Figure 11: Comparison of Agent Discovery delays.

Probe delay was measured in scenarios of various scanning types, namely Passive & Full Scan, Passive & Selective Scan, Active & Full Scan and Active & Selective Scan. Typical results are presented in Figure 10. The number of channels in full scan is 11 (FCC Regulatory Domain of North America), while with selective scan, we assume that the MN just scans non-overlapping channels (3 channels). As the MN continuously monitors its current channel, the number of channels to be scanned in full mode and selective mode are 10 and 2 respectively. From the graph, it is quite clear that having knowledge of channel allocation, and therefore being able to scan selectively, can reduce significantly Probe delay in both passive and active modes. Active scan further reduces delay, but at the same time consumes more bandwidth because of Probe Request and Probe Reply frames.

Figure 11 presents results that confirm what we discussed early in section IV.C. Forced Handover can reduce significantly Agent Discovery delay, potentially to zero. Even in the case of incomplete proactive CT, there is still significant reduction.

Finally, we show results for overall handover latency (Table 3) in some typical scenarios as follows: Active & Full Scan, Non-Forced Handover (S1), Active & Selective Scan, Forced Handover, Incomplete Proactive CT (S2), and Active & Selective Scan, Forced Handover, Complete Proactive CT (S3).

It is noted that the Agent Advertisement Interval

Table 3: Handover Latency Components

	S1	S2	S3
$T_{ProbeDelay}$	212 ms	0 ms	0 ms
$T_{Authentication}$	41 ms	41 ms	0 ms
$T_{Association}$	21 ms	21 ms	21 ms
T_{AD}	500 ms	20 ms	0 ms
$T_{Registration}$	100 ms	100 ms	100 ms
Overall	847 ms	182 ms	121 ms

($T_{ADV\_interval}$) is reduced down to 1 sec, so that average Agent Discovery delay is equal to half of the $T_{ADV\_interval}$ i.e. 500 ms. However, it is still a significant delay, and can be reduced further by applying Forced Handover (scenarios S2 and S3).

VI. Conclusion and Future Work

In this paper, we presented a simple scheme for proactive CT with a new concept of Forced Handover in WLAN-based access networks. Based on observation of SNR changes, the proposed scheme predicts the best moment in time to perform the CT. In our scheme, the MN is forced to carry out a handover. As the handover is forced to happen at a "planned" moment in time, the network can prepare for such event by selecting the best AP and AR, and transferring service context information between APs and ARs; therefore the scheme facilitates seamless mobility. Thanks to proactive CT and Forced Handover scheme, the MN can significantly reduce Probe delay, authentication delay, and Agent Discovery delay; hence improve the handover latency. The improvement is achieved at the expense of small increase in the number of unnecessary handovers; this however can be kept at a reasonably low level by appropriate selection of scanning interval.

We intend to carry the research described in the paper further. Firstly, we intend to verify the proactive scheme for other simulation scenarios, i.e. characterised by different AP distributions and mobility models. Secondly, as pointed out in the discussion, the scanning interval is open to optimisation: lower T_{SI} gives better performance of the proposed scheme, but affects (interrupts) the user communications more. Therefore, we need to investigate optimisation of T_{SI}. One approach may be to use an adaptive scanning interval. For example, initially T_{SI} can be selected large, and then be reduced adaptively as the MN approaches handover.

References

[1] A. Mishra, M. Shin, and W. Arbaugh, An Empirical Analysis of the IEEE 802.11 MAC Layer Handoff Process, *ACM Computer Comm. Review*, vol. 33. no. 2, pp. 93-102.

[2] A. Mishra, M. Shin, and W. Arbaugh, Context Caching using Neighbour Graphs for Fast Handoffs in Wireless Networks, University of Maryland, Technical Report CS-TR-4477, 2003.

[3] ANSI/IEEE Standard 802.11, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, 1999 Edition.

[4] C. Perkins (editor), IP Mobility Support for IPv4, RFC 2002, IETF, January 2002.

[5] Cisco Systems Inc., Cisco AVVID Wireless LAN Design: Solution Reference Network Design, 2003.

[6] IEEE, Recommended Practice for Multi-Vendor Access Point Interoperability via an Inter-Access Point Protocol Across Distribution Systems Supporting IEEE 802.11 Operation, *IEEE Draft 802.11f/D5*, Jan 2003.

[7] J. Kempf (editor), Problem Description: Reasons for performing Context Transfers between nodes in an IP access network, RFC 3374, Sept 2002

[8] J. Loughney (editor) et al., Context Transfer Protocol, Internet draft (draft-ietf-seamoby-ctp-08.txt, work in progress), IETF, Jan 2004.

[9] Lucent Technologies Inc., Roaming with WaveLAN/IEEE 802.11, Tech. Rep. WaveLAN Technical Bulletin 021/A, Dec 1998.

[10] M. Liebsch, A. Singh, et al., Candidate Access Router Discovery, Internet draft (draft-ietf-seamoby-card-protocol-05.txt, work in progress), IETF, Nov 2003.

[11] T. Pagtzis, P. Kristein, S. Hailes and H. Afifi, Proactive seamless mobility management for future IP radio access networks, *Computer Comm.*, vol. 26, pp. 1975-1989, 2003.