S-MIP: A Seamless Handoff Architecture for Mobile IP

Robert Hsieh, Zhe Guang Zhou, Aruna Seneviratne
School of Electrical Engineering and Telecommunications
The University of New South Wales
Sydney, 2052, Australia
{roberth, zheguang}@ee.unsw.edu.au, a.seneviratne@unsw.edu.au

***Abstract*—As the number of Mobile IP (MIP) [2] users grow, so will the demand for delay sensitive real-time applications, such as audio streaming, that require seamless handoff, namely, a packet lossless Quality-of-Service guarantee during a handoff. Two well-known approaches in reducing the MIP handoff latency have been proposed in the literature. One aims to reduce the (home) network registration time through a hierarchical management structure, while the other tries to minimize the lengthy address resolution delay by address pre-configuration through what is known as the fast-handoff mechanism. We present a novel seamless handoff architecture, S-MIP, that builds on top of the hierarchical approach [4] and the fast-handoff mechanism [3], in conjunction with a newly developed handoff algorithm based on pure software-based movement tracking techniques [16]. Using a combination of simulation and mathematical analysis, we argue that our architecture is capable of providing packet lossless handoff with latency similar to that of L2 handoff delay when using the 802.11 access technology. More importantly, S-MIP has a signaling overhead equal to that of the well-known 'integrated' hierarchical MIP with fast-handoff scheme [4], within the portion of the network that uses wireless links. In relation to our S-MIP architecture, we discuss issues regarding the construction of network architecture, movement tracking, registration, address resolution, handoff algorithm and data handling.**

Keywords – Mobile IP, Hierarchical Mobile IPv6, Seamless Handoff, Fast-handoff (Low Latency Handoff), Software-based Mobile Device Tracking

I. INTRODUCTION

Mobile IP (MIP) [2] describes a global mobility solution that provides host mobility management for a diverse array of applications and devices on the Internet. In Internet (IP) environments, when a mobile node moves and attaches itself to another network, it needs to obtain a new IP address. This changing of the IP address requires all existing IP connections to the mobile node be terminated and then re-connected. This is necessary as the IP routing mechanisms rely on the topological information embedded in the IP address to deliver the data to the correct end-point. Mobile IP overcomes this by introducing a level of indirection at the network (IP) layer. This indirection is provided with the use of network agents and does not require any modification to the existing routers or end correspondent nodes. With MIP, each mobile node is identified by a static home network address from its home network, regardless of the point of attachment. While a mobile node is away from its home network, it updates a special entity, a home agent, with information about its current IP address. The home agent intercepts any packets destined to the mobile node, and tunnels them to the mobile node's current location. Thus, it is necessary for a mobile node to register its location at the home agent. The time taken for this registration process combined with the time taken for a mobile node to configure a new network care-of address in the visiting network, amounts to the overall handoff latency. Thus the handoff latency in Mobile IP is primarily due to two procedures, namely, the address resolution and the (home) network registration.

There have been numerous proposals for minimizing the handoff latency of MIP. These can be broadly classified into two groups. The first group aims to reduce the network registration time by using a hierarchical network management structure while the second group attempts to reduce the address resolution time through address pre-configuration. The former is generally referred to as hierarchical handoff and the latter as fast-handoff or low-latency handoff. IETF drafts [4] and [3] incorporate the concepts of hierarchical and fast-handoff mechanisms in the IPv6 network, based on Mobile IPv6 [1]. However, although it has been shown that the combined use of hierarchical handoff and fast-handoff improves the performance, it is nonetheless not sufficient in providing a packet lossless handoff environment at IP layer, since an approximate delay of 300 to 400 milliseconds has been observed [10]. Furthermore, this combined scheme does not address the mobile 'ping-pong' movement problem effectively with respect to handoff delay.

This paper presents an architecture which minimizes the handoff latency, in large indoor open environments, to one that is similar to that of the L2 handoff schemes, thus virtually eliminating packet loss at the L3 IP layer. The architecture is able to achieve the above with handoff signaling overheads no greater than the well understood integrated hierarchical and fast-handoff scheme, whilst being scalable, highly available and sustains fault tolerance. S-MIP explores the intersection between mobile device tracking techniques, handoff algorithms and hierarchical Mobile IP Architecture and synthesizes together key advantages from each of those in achieving a seamless handoff architecture. The rest of the

paper is organized as follows. Section II provides the background and the related work. Then we describe the architecture which we refer to as S-MIP in Section III. The evaluation of S-MIP is presented in Section IV and we provide some concluding remarks in Section V.

II. BACKGROUND AND RELEATED WORK

A. Hierarchical Mobile IPv6

Hierarchical handoff schemes separate mobility management into micro mobility and macro mobility management, otherwise known as intra-domain mobility and inter-domain mobility management respectively. The central element of these schemes is the inclusion of a special conceptual entity called Mobility Anchor Point (MAP) [4]. It is normally placed at the edges of a network, above a set of access routers, which constitute its network domain. The MAP is a router or a set of routers and maintains a binding between itself and mobile nodes currently visiting its network domain. Thus, when a mobile node (MN) attaches itself to a new network, it is required to register with the MAP serving that network domain (MAP domain). The MAP intercepts all the packets addressed to the MN it serves and tunnels them to the MN's on-link[1] care-of address (LCoA). If the MN changes its LCoA within the same MAP domain, new LCoA binding with the MAP is required. If a MN moves into a separate MAP domain, it needs to acquire a new regional address (RCoA) as well as a LCoA. Usually, the MN will use MAP's address as the RCoA and LCoA can be formed according to methods described in [11]. After forming these addresses, the MN sends a regular MIPv6 Binding Update (*BU*) to the MAP, which will bind the MN's RCoA to the LCoA. In response, the MAP will return a binding acknowledgement (*BAck*) to the MN. Furthermore, the MN must also register its new RCoA with its home agent by sending another *BU* that specifies the binding between its home address and the new RCoA. Finally, it may send *BU* to its current corresponding nodes, specifying the binding between its home address and the newly acquired RCoA.

B. Fast-handoff mechanism

The basic operation of the Fast-handoff [3] is illustrated in Figure 1. Fast-handoff introduces seven additional message types for use between access routers and the MN. An access router is the last router between the wired network and the wireless network where the MN is situated. These seven messages are: Router Solicitation for Proxy (*RtSolPr*), Proxy Router Advertisement (*PrRtAdv*), Handover Initiation (*HI*), Handover Acknowledgement (*HAck*), Fast Binding Acknowledgement (*F-BAck*), Fast Binding Update (*F-BU*) and Fast Neighbor Advertisement (*F-NA*). In addition, the old Access Router (oAR) is defined as the router to which the MN

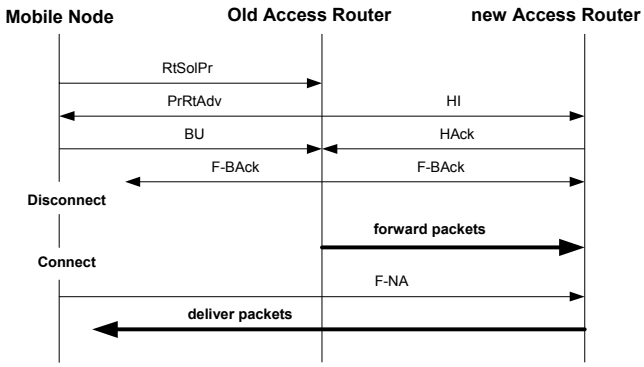

Figure 1 Fast-handoff Message Interaction

is currently attached, and the new Access Router (nAR) as the router to which the MN is about to move to. A fast-handoff is initiated on an indication from a wireless link-layer (L2) trigger. The L2 trigger[2] indicates that the MN will soon be handed off. Upon receiving an indication, the fast handoff scheme anticipates the MN's movement and performs packet forwarding to the nAR(s) accordingly. This is achieved by the MN sending a *RtSolPr* message to the oAR indicating that it wishes to perform a fast-handoff to a new attachment point. The *RtSolPr* contains the link-layer address of the new attachment point, which is determined from the nAR's beacon messages. In response, oAR will send the MN a *PrRtAdv* message indicating whether the new point of attachment is unknown, known or known but connected through the same access router. Further, it may specify the network prefix that the MN should use to form the new CoA. Based on the response, the MN forms a new address described using the stateless address configuration described in [11]. Subsequently, the MN sends a *F-BU* to the oAR as the last message before the handover is executed. The MN receives a *F-BAck* either via the oAR or the nAR indicating a successful binding. As the exact handoff instance is unpredictable, the oAR sends a duplicated *F-BAck* to the nAR to ensure the receiving of *F-BAck* by the MN. Finally, when the MN moves into the nAR's domain, it sends the Fast Neighbor Advertisement (*F-NA*) to initiate the flow of packets at the nAR. In addition to the message exchange with the MN, the oAR exchanges information with the nAR to facilitate the forwarding of packets between them and to reduce the latency perceived by the MN during the handoff. This is realized by the oAR sending a *HI* message to the nAR. The *HI* message contains MN's requesting CoA and the MN's current CoA used at the oAR. In response, the oAR receives a *HAck* message from the nAR either accepting or rejecting the requested new CoA. If the new CoA is accepted by the nAR, the oAR sets up a temporary tunnel to the new CoA. Otherwise, the oAR tunnels packets destined for the MN to the nAR, which will take care of forwarding packets to the MN temporarily.

[1] Three different addressing scopes exist in IPv6. A global address uniquely identifies a node on the Internet. A regional address is a global address that is specific to a particular region/domain on the Internet. An on-link address is an address local to a domain and it is only a unique identifier inside a specific domain.

[2] L2 trigger is implementation dependent, i.e., based on the access technology used. Cross protocol layer messaging techniques are used.

C. *Related work*

Some work has already investigated the feasibility and performance of hierarchical and fast-handoff schemes. In [5] the signaling cost of hierarchical schemes and deployment of Mobile IPv6 is addressed. In [6], the 'prehandoff registration', similar to fast-handoff, is shown to perform better than the standard Mobile IPv4 handoff. Moreover, [10] shows that hierarchical together with fast-handoff schemes greatly reduces the overall handoff latency to around 300 to 400 milliseconds. However, [10] also shows that hierarchical with fast-handoff is still far from offering a seamless handoff environment, as packet loss at the IP layer still exists, hence impacting on packet lost sensitive applications. Furthermore, the 'ping-ponging' issue is not addressed adequately. In this paper, we propose the S-MIP architecture that not only addresses 'ping-ponging', but also reduces the handoff latency to where at the IP-layer, the end devices perceives a seamless connectivity, that is, no packet lost.

III. S-MIP: A SEAMLESS HANDOFF ARCHITECTURE

S-MIP Architecture builds on the fast-handoff [3] and hierarchical schemes [4] and introduces the use of an 'intelligent handoff' mechanism. This intelligent handoff mechanism is operable on the existing hierarchical frameworks [4] without any modifications. In the following sections, the design of S-MIP, its intelligent handoff mechanism, and the mathematical proof of its validity are presented.

A. *Design*

We design the S-MIP, seamless handoff architecture for Mobile IP, with the following goals and assumptions:

- **Extreme Low Handoff Latency.** The S-MIP is to achieve packet lossless handoff latency at the Internet IP layer. We aim to obtain a handoff latency similar to that of the L2 handoff delay, which is in the order of tens of milliseconds.
- **Minimal Handoff Signaling.** The S-MIP will have a signaling mechanism with overhead cost no greater than the well understood integrated hierarchical MIPv6 with fast-handoff [4] scheme (within the wireless domain portion).
- **Indoor Large Open Space Environment.** With increasing demand for wireless network access under indoor large open spaces, such as convention halls, hotel ballrooms and airports, S-MIP and its intelligent handoff algorithm are designed to function optimally under such environment. We assume free space propagation model and low mobile device movement velocity, i.e. 1m/s.
- **Scalability, High Availability, Fault Tolerance.** S-MIP is to be scaled incrementally for ease of deployment. The architecture is to be able to tolerate failure gracefully and hide them from the end mobile devices. Moreover, control entities are to be distributed, thus achieving load balancing and high availability.

From prior analytical study on Hierarchical MIP and fast-handoff [10], we classify packet losses as being either due to losses within the MAP and the access routers (segment packet loss) or between the last access routers and the mobile device (edge packet loss). These edge packet losses occur due to mobility of a mobile node and transmission errors. The segment-packet lost is due to the non-deterministic nature of handoffs and the resulting switching of the data stream at the MAP upon receiving the MAP Binding Update. Thus the design of S-MIP is aimed at minimizing the edge packet and segment packet losses. Edge packet loss is minimized by keeping the 'anchor point' for the forwarding mechanism [4] as close to the mobile node as possible. Hence, it is located at the access router that bridges the wireless network and the wired network. Segment packet loss is minimized by using a newly developed Synchronized-Packet-Simulcast (SPS) scheme and a hybrid handoff mechanism. The SPS simulcasts packets to the current network that the mobile node is attached to, and to the potential access network that the mobile node is 'asked' to switch onto. The hybrid handoff strategy is 'mobile node initiated, but network determined'. This allows the mobile node, which has the best knowledge regarding its current location to initiate the handoff, yet allow the network system to unambiguously determine which network a mobile node should switch to. The decision as to which access network to handoff is formulated from the movement tracking mechanism, which is based on a synchronized feedback mechanism. The movement tracking and handoff algorithm aims to distinguish the following three conditions, namely, is the mobile device currently moving linearly, moving stochastically or is stationary at the center of overlapping area formed by two network coverage areas. Stationary and not near the center of the overlapping areas need not be considered, as there is no ambiguity as to which network a mobile device should belong to. By determining the movement conditions, different handoff strategy can be applied regarding a handoff (details in Section III-C). More importantly, this movement tracking is to be achieved by using only readily available infrastructure, i.e., the ARs, in performing location tracking. No additional hardware infrastructure is necessary.

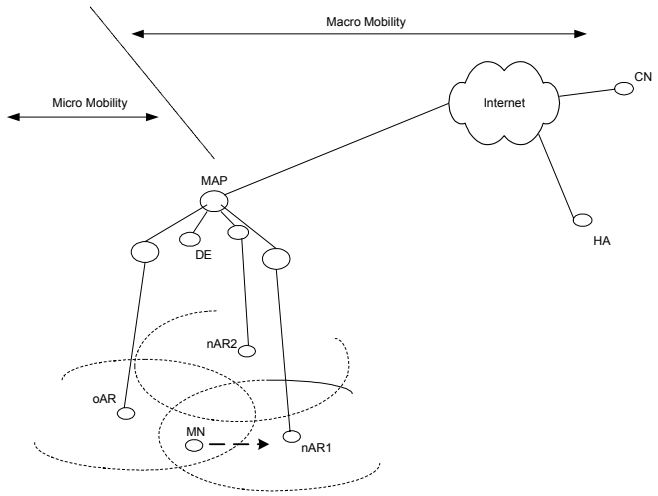

Figure 2 S-MIP Architecture

In order to ensure the full benefits of SPS a coarse-grain packet sequencing scheme inside access routers are required. To ensure that no packets are lost, packets need to be forwarded at the oAR. However, as SPS introduces the simulcast packets (s-packets), it is advantageous to separate these from the forwarded packets (f-packets), to ensure seamless handoff. The reason is that f-packets are most likely to be chronologically 'older' than the s-packets received from the MAP (see analysis in Appendix D). As there is no packet sequencing capability in IP, this coarse-grain separation between these two types of packets is necessary to minimize the re-ordering of packets.

B. S-MIP Network Architecture

Figure 2 illustrates the architecture of S-MIP. It virtually retains the hierarchical handoff framework and extends it by adding a Decision Engine (DE) entity, and the Synchronized-Packet-Simulcast (SPS) scheme with an intelligent hybrid handoff protocol. Identical with the hierarchical handoff schemes, *Mobility Anchor Point* (MAP) separates the mobility scheme into micro mobility and macro mobility. The new Access Router (nAR) and old Access Router (oAR) retain the same functionality and meaning. The Decision Engine (DE) is similar to a MAP in its scope, and makes handoff decision for its network domain. Through periodic feedback information from individual ARs, the DE maintains a global view of the connection state of any mobile devices in its network domain, as well as the movement patterns of all these mobile devices. The DE is also capable of offering load-balancing services where it is able to instruct ARs, responsible for fewer mobile devices, to take on new mobile device, rather than ARs that are closer to the mobile device.

Figure 3 shows a typical operation of S-MIP which defines six new additional messages to the messages used in hierarchical and fast-handoff scheme [4]. These six additional messages are:

- Current Tracking Status (*CTS*) message from the MN to DE. It contains location tracking information.
- Carrying Load Status (*CLS*) message from the ARs to DE. The *CLS* message contains the information regarding how many mobile devices an AR is currently managing.
- Handoff Decision (*HD*) message from the DE to ARs. The *HD* message contains the outcome of the handoff decision at the DE, namely which AR a MN should handoff to.
- Handoff Notification (*HN*) message from the oAR to MN. The *HN* contains the indication from the oAR to the MN, directing exactly which nAR the MN should handover to (this is sent in combination with the *PrRtAdv*). The oAR derives the content of the *HN* message from the received *HD* message.
- Simulcast (*Scast*) message from oAR to MAP. The *Scast* message triggers the start of the SPS process.
- Simulcast Off (*Soff*) message from nAR to MAP. This message terminates the SPS process.

Furthermore, the Router Advertisement message is modified to include the DE reply option similar to that of the MAP discovery option [4].

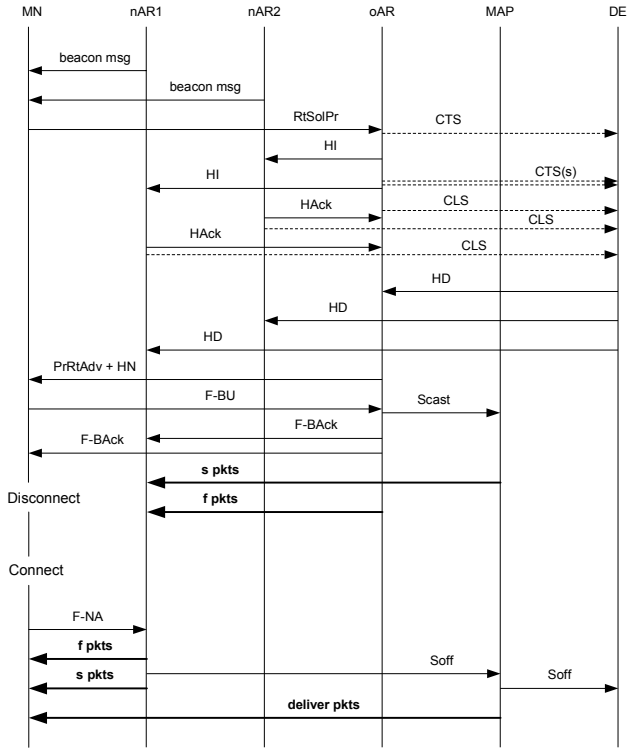

Figure 3 Mobile Node Initiated, Network Determined Linear Handoff

The behaviors of the entities involved in the exchange of these messages are described in the following. Upon receiving beacon advertisement messages from the newly discovered ARs, the MN initiates a handoff by sending the *RtSolPr* message to the oAR. This indicates that the MN desires to proceed with a seamless handoff to the new attach points nAR (e.g. nAR1 or nAR2). Upon receiving the *RtSolPr* message from the MN, the oAR sends *HI* messages, to all the potential nARs identified by the MN in the *RtSolPr* message. This *HI* message will contain within the requested care-of address (CoA) on the nAR and the current care-of address being used at the oAR. All nARs will respond to the *HI* message with a *HAck* message either accepting or rejecting the new CoA. As specified in [4], if the new CoA is accepted by the nAR, the oAR sets up a temporary tunnel to the new CoA. Otherwise the oAR tunnels packets destined for the MN to the nAR, which takes care of forwarding packets to the MN temporarily. In response to the *RtSolPr* message, the MN will receive a *PrRtAdv* messages from the oAR, similar to the hierarchical scheme [4].

ARs send *CLS* messages to the DE periodically (every 3 seconds approximately) as a reply to the modified Router Advertisement [13] message (with DE reply option). The *CLS* message indicates how many mobile devices are associated with a particular AR as well as the IP addresses of those MNs. A *CTS* message is generated by the MN every time it receives a L2 beacon advertisement from the ARs. Each *CTS* message contains the signal strength of the detectable AR and the respective Ids of the AR. The signal strength and AR Ids serve

as MN's location tracking information. The current AR will forward this tracking information (*CTS*) every second until the reception of the *HD* message from the DE. The MN may send the tracking information to other ARs, other than the current AR, if the connection to the current AR is poor. In this case, duplicated *CTS* messages will arrive at the DE, which will simply discard the duplicated messages. After analyzing the *CTS* and *CLS* messages, including tracking the mobile node's movement for a short period (minimum of 3 seconds, analysis shown in Appendix C), the DE sends *HD* messages to all participating ARs for the specific mobile node requesting the seamless handoff. In turn, the oAR sends a *HN* message together with the *PrRtAdv* to the MN indicating exactly which AR that the MN should switch to.

The movement tracking is achieved as follows. Firstly, if the mobile node is determined to be in the stochastic moving state, the *HD* messages will inform ARs to be in the anticipation-mode. In this mode, even though a MN might no longer wish to be associated with an AR, the AR will still maintain the MN's binding, in preparation for the MN returning (ping-ponging). This avoids the unnecessary 're-setup' resource and time costs. In this scenario, the *HN* message to the MN from the oAR will indicate that the MN is able to switch network freely, using Fast Neighbor Advertisement (*F-NA*) message, once the signal strength has decreased to a certain predefined threshold. The DE will send further *HD* messages to any of the participating ARs in cases where it determines that they are no longer required in the anticipation-mode. Secondly, if the MN is determined to be in the stationary state near the boundary between two network coverage areas, the *HD* message from the DE will instruct the action of multiple bindings between the MN and the ARs. (MN use more than one care-of address simultaneously, see [1]) Lastly, if the MN is determined to be moving in a linear fashion, the *HD* message will contain which AR the MN is to be handed off to. The ARs that are not selected for handoff by the DE will be notified to discontinue from further participation in this handoff process in its *HD* message.

The MN will only send a *F-BU* message to the oAR after receiving the *HN* message and the formation of the new care-of address as described in [11]. This *F-BU* message binds MN's current on-link address to the new care-of address. When the oAR receives this *F-BU*, it will send the *Scast* message to the MAP initiating the simulcasting of packets. Every subsequent packet from the CN arriving after the reception of the *Scast* message at the MAP will be duplicated and send to both the oAR and the nAR simultaneously. These packets will be marked with a *S* bit, as an option parameter, in the IP header. Furthermore, as a reply to the MN's *F-BU* message, a *F-BAck* message will be sent by the oAR to both its current network and the new network. This is to ensure that the MN receives the *F-BAck* message, as the precise moment that the MN will switch networks is unpredictable.

The nAR maintains two distinct buffers in the S-MIP architecture, namely a 'f-buffer' and a 's-buffer'. The f-buffer contains packets forwarded from the oAR (f-packets) while the s-buffer contains packets that are marked with the *S* bit (s-packets). The nAR will start delivering buffered packets to the MN after it receives the *F-NA* message, from the MN, signifying that the MN have arrived at its network. The nAR will attempt to transmit the f-buffer and empty it before beginning to transmit from the s-buffer. Meanwhile, at the oAR, it will only forward those packets which do not have the *S* bit marked to the nAR. In addition, all packets will be sent on the wireless channel. In case that the mobile node does not switch network immediately, it will therefore still be able to receive packets from the oAR. The nAR will send the *Soff* message to the MAP after the f-buffer has been emptied, indicating the termination of packet simulcast. Upon receiving the *Soff* message, the MAP performs the binding update of associating the new on-link address of the MN with its regional care-of address. A *Soff* message will also be forwarded to the DE by the MAP. The DE will not allow the MN to perform another seamless handoff before the current seamless handoff has been completed.

For situations where the MN is moving in a stochastic fashion, the handoff procedure is the same as described previously until the packet simulcast stage, where the MAP will perform packet simulcast to *all* potential nAR identified by the DE. The MN will be allowed to switch to any one of these potential nAR, by sending a *F-NA* message, provided that the signal strength of its current AR is below the allowable threshold. This threshold value and the available ARs (for handoff) are indicated in the *HN* message to the MN via the oAR. Until further *HD* messages from the DE to any of the ARs, the packet simulcast will continue even if the f-buffer of the nAR has been emptied. Upon receiving the *HD* message from the DE indicating the termination of packet simulcast, an AR (picked by the DE) will be responsible for sending the *Soff* message to the MAP and the MAP will subsequently forward the *Soff* message to the DE signifying the completion of the process. For situations where the MN is stationary near the boundary of network coverage areas, the handoff procedure is similar to the stochastic case, except that the MN now binds simultaneously to more than one access router by using two or more care-of-address [1]. Thus no *F-NA* message is required to be sent, by the MN to the AR, before requesting the buffered packets from ARs.

As mentioned previously, modification to the standard Router Advertisement message [13] is necessary. Similar to that described in [4], a MAP discovery option is added to the Router Advertisement message to enable the discovery of the MAP for the MN. Furthermore, the DE reply option is added to the Router Advertisement. This option synchronizes the timing of the *CLS* message from each individual AR to the DE. The reception of such Router Advertisement with the DE reply option triggers the *CLS* message replies from the ARs. This synchronized *CLS* messaging combined with the periodic *CTS* messages is critical to the precise calculation of the movement tracking and handoff algorithm. If the MN moves into a different MAP domain, similarly to [4], state information will need to be transferred. The movement tracking and pattern information can also be exchanged in a similar fashion between the DEs.

C. *Movement Tracking and Handoff Algorithm*

As described earlier, movement tracking is an integral part of the S-MIP architecture. Movement tracking is a two-phase process consisting of location tracking and followed by movement pattern detection. Therefore, the movement tracking scheme in S-MIP enables the determination of types of movement, namely, linear, stationary or stochastic, in 802.11 networking environments. This tracking process is executed continuously once a MN moves within zones, where movement detection is necessary, defined by various coverage areas described below.

1) Coverage Areas and Zone Definitions

The coverage area of an 802.11 AR can be defined in terms of Signal Strength (SS) or Signal-to-Noise Ratio (SNR). In S-MIP, the movement tracking mechanism uses the SS as the metric, rather than the SNR, which is prone to random fluctuations due to noise. When a MN is near to an AR, the SS from the AR is strong and the link quality is high. Therefore the probability of losing packet is very low (usually zero [8]). When the MN moves further away, the SS decreases, which can be described using a negative log function [16]. As the S-MIP architecture is aimed at large indoor environments, there will be a minimum of three ARs that provides the required coverage as shown in Figure 4. The coverage area of each access router is divided into four different areas. First, referred to as the 'effective' coverage area, has high SS and results in no packet loss. The second is defined as the 'marginal' coverage area, and corresponds to areas outside the effective area, which nonetheless maintain a low percentage of packet loss (below 5%). The third area, referred to as the 'poor' coverage area represents the rest remaining area. The fourth area is a logical area inside the effective coverage area referred to as the 'good' coverage area. For any overlapping set of three ARs, the good coverage area of these ARs intersects at a single point (in theory). The difference in radius between the good and effective coverage area radius, *d*, has a direct influence on the minimum overlapping distance, inside the overlapping zone II. Figure 4 also shows the three different zones formed by the three ARs' effective coverage areas. They are, zone I where the MN can only receive signals from only one AR, zone II where the MN is able to receive from two different ARs, and zone III where the MN can receive from all three ARs.

It is known that the attenuation factor, due to human obstruction or device orientation is 6.4dB and 9.0dB respectively [16]. Therefore, when assuming that attenuation is approximated by a negative log function, 6.4dB equates at least to a distance of 5 meters, therefore a radius difference, *d*, of 2.5 meters. As a result, we assume that distance *d* is at least 2.5m. Described in Appendix A, it can be shown that a *d* value of 1m will result in a minimum overlapping distance *A'B'* of 3.28m and 2m results in a *A'B'* of 6.25m. (For $r = 2.5$, d is > 7m.) This is more than enough for movement pattern detection, since that the sampling interval is defined to be once every second and a minimum of 3 samples are sufficient in determining the movement pattern by the Decision Engine as shown in Appendix C.

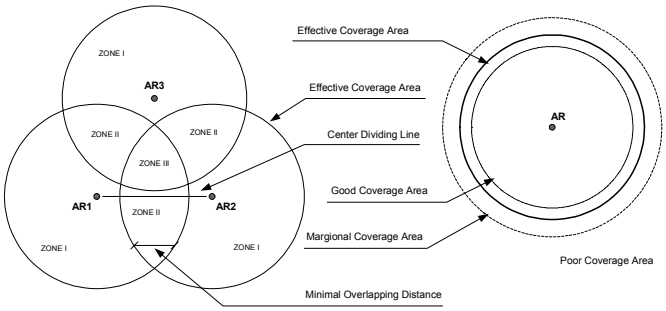

Figure 4 Coverage Model

2) Location Tracking

The location tracking is an integral part of S-MIP and requires no additional hardware or pre-measurements. Unlike [7] and [9], it is a pure software based approach that is capable of obtaining 1 to 2 meters accuracy in tracking mobile devices based on our prior work in [16]. We perform location tracking when a MN enters zone II and/or zone III.

In zone II, a MN receives two SS signals from two different ARs, AR1 and AR2. Location tracking mechanism will be activated upon a DE receiving its first *CTS* message containing both access router SS values from the MN via its current AR. The MN will continue updating the DE with the *CTS* message every second. However, within zone II, a specific location cannot be determined precisely by using only two effective SS, this is because the two effective SS distance circles intersect at two points. It is not possible to differentiate one from another without an auxiliary reference. Nevertheless, the top portion of the combined coverage area is partially covered by AR3 (see Figure 4). Therefore, the DE is still able to infer which intersection point represents the MN's current location, through the *CLS* message from AR3. In cases where a MN is located near the center of zone II, where location inference using AR3's CLS message might not be effective, the exact location can still be determined using marginal coverage area calculation inference. As shown in Appendix B, the marginal coverage area should cross the center of zone II defined in Figure 4 as line AR1-AR2. This means that there exists no 'dead corner' inside zone II and location tracking is achievable anywhere in zone II, with 95% accuracy at worst.

In zone III, the MN receives three different effective SS signals from three different ARs. The exact location of the MN can be determined by using the triangulation technique developed from our previous work in [16]. Triangulation technique is able to determine the location of the MN with reasonable accuracy (1 to 2 meters) by using *only* 3 access routers as reference points. In zone I, the exact location of the MN cannot be detected. However, this is non-critical as no handoff is required when a mobile device is in this zone. The *CTS* message-train will stop when the MN enters zone I, since the requirements for sending the *CTS* messages is that the MN must receive more than 2 different SS signals. Nevertheless, the AR is still able to determine how many MN are under its management through layer-2 link connectivity information and report this via *CLS* message to the DE.

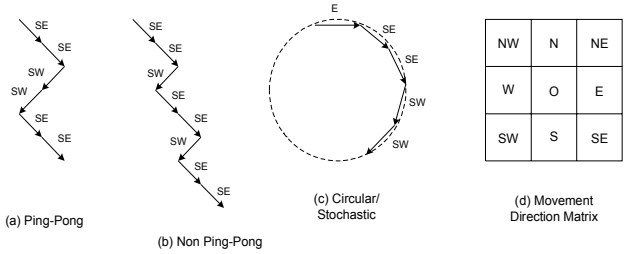

Figure 5 Movement Pattern

3) Movement Pattern Detection

For detecting the movement pattern, the DE needs to sample the location tracking information of the MN periodically. The accuracy of the movement pattern detection depends on the sampling period and the MN moving speed. In S-MIP, it is assumed that the MN moves with a speed of 1m/s, and the sampling period is defined to be one second, beginning with the first *CTS* message sent to the DE from the MN via the AR(s). As shown in Appendix C, under these conditions, a minimum of only 3 samples are required to establish the direction of movement. Once the direction of movement is known, the movement pattern can be established by looking at the history of the direction in which the MN has moved. The movement history is created by recording a series of movement directions, illustrated in Figure 5. The movement directions are based on the matrix in Figure 5 where the center, O, indicates the current location of the MN and the rest of the eight squares illustrate the next movement direction of the MN in terms of North (N), South (S), East (E), West (W), North-East (NE), South-East (SE), South-West (SW) and North-West (NW) respectively. The direction is calculated by firstly determining the location position using SS values and then the differential of the location position at two different time intervals [16]. Thus, to detect a linear movement pattern, all the movement direction information should be the same, meaning, the DE needs to detect a set of repeated directions of movement. Likewise, a MN being stationary is detected by all movement information indicating the center O, i.e. the MN is always at O in Figure 5. Furthermore, the SS values from ARs must be similar. Anything that is not detected to be linear or stationary is considered to represent stochastic movement.

D. Analysis of the S-MIP Architectural Validity

The S-MIP architecture is based on two fundamental observations. The first is that all f-packets are unlikely to be received at the new access router before any s-packets are received. The second is that the old access router is unlikely to start forwarding f-packets to the new access router before the MN switches to the new access network. A formal justification of these observations is given in Appendix D. This observation implies that it is necessary to provide a mechanism for packet re-sequencing and double buffering. In this paper, all analysis and experiments assume the simple case of single connection with no interference from other connections. Multiple connection analysis is more difficult in that packet inter-arrival time could have complex distribution and high dispersion [12].

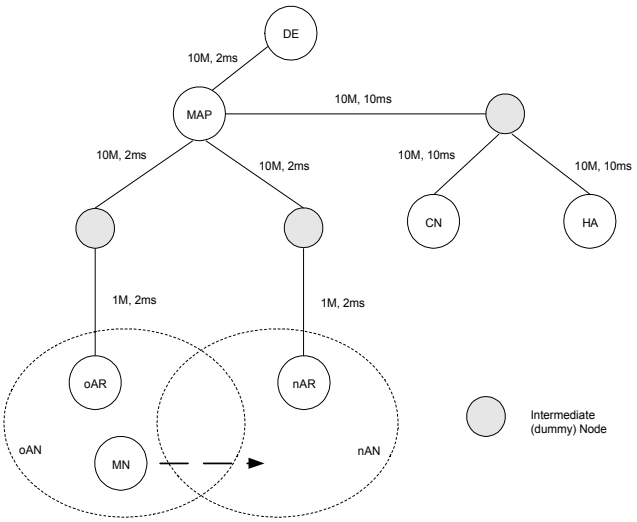

Figure 6 Simulation Network Topology

IV. PERFORMANCE EVALUATION

A. Simulation Performance result

Figure 7 illustrates our preliminary simulation result for the S-MIP architecture. The goal of our simulation was to examine the effect of S-MIP architecture on handoff latency of an end-to-end TCP communication session. In particular, we wanted to examine the packet loss and packet re-ordering behavior in S-MIP. The extensions to *ns* [15] described in [10] were used for the simulations described below.

Figure 6 shows the network topology used for the experiments. This topology depicts a simplistic version of a typical Mobile IP network topology and has been used extensively in MIP performance studies [4]. The link characteristics, namely the bandwidth (megabits/s) and the delay (milliseconds), are shown beside the link. The access routers are set to be 70 meters apart with free space in between to ensure only signal interference. A Lucent WaveLan Card running 802.11 protocols was simulated with a coverage area of approximately 40 meters in radius. We deliberately increased the radius to 40m (from 25m) to show that minimum overlapping distance tends to a constant, irrespective of the value of radius (analysis in Appendix C). As an initial proof of concept simulation, we only consider a linear mobile node movement pattern, where the mobile node moves linearly from one access router to another at a constant speed of 1m/s. Finally, TCP Tahao, which follows a 'go back-n model using accumulative positive acknowledgement with slow start, congestion avoidance and fast retransmission' model, was chosen as the default TCP flavor. A *ns* TCP source agent is attached to the corresponding node (CN) and a *ns* TCP sink agent is attached at the Mobile Node (MN). The MN is initially positioned near the old Access Router (oAR) inside the old Access Network (oAN) and starts to move towards the new Access Router (nAR) in the new Access Network (nAN) 6 seconds after the simulation starts. This is to enable the establishment of TCP communication and allowing it to

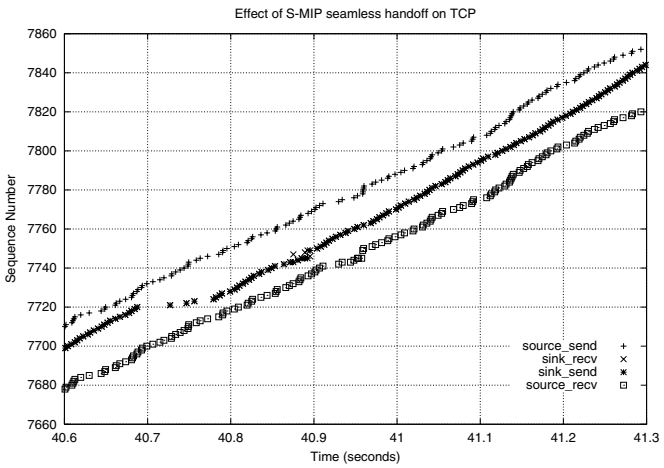

Figure 7 Handoff Result for S-MIP

stabilize, meaning, TCP is transferring data with a full window. A FTP session between the MN and the CN is started 5 seconds after the simulation has started. The bulk FTP data traffic flow is from the CN to the MN. Each simulation run in the experiment had a duration of 80 seconds, consists of 5 seconds of initial stabilization period, 70 seconds of linear movement from oAR to nAR, and the final 5 seconds being stationary near the nAR.

The results of these experiments are shown in Figure 7. The upper most curve is the *source_send* curve, indicating the CN's TCP sending buffer. The bottom most curve is the *source_recv* curve, showing the CN's TCP receiving buffer. The *sink_recv* and *sink_send*, close together in the middle, corresponds to MN's TCP receiving and sending buffer. These results show that S-MIP's handoff mechanism completely eliminates the L3 disruption perceived by the communication end-host. As can be observed, the TCP sender is essentially uninterrupted by the handoff mechanism despite the apparent L2 handoff occurring at approximately t = 40.70s. Furthermore, there is no packet loss at the IP layer. The re-transmission, shown in Figure 7 around t = 40.86s, is caused by the handoff, yet it did not activate the TCP congestion control mechanism due to the double-buffering mechanism in the S-MIP architecture.

B. Signaling Cost Analysis

The primary overhead cost of S-MIP is associated with its signaling. This signaling can be divided into two parts. First part is associated with setting up the seamless handoff. As can be seen from the S-MIP handoff example, in the wireless domain, the signaling cost is identical to that of the integrated HMIPv6 with Fast-handoff described in [4]. The functional role of the *RtSolPr*, the *HI* and the *HAck* messages from the fast-handoff framework remain unchanged. The *PrRtAdv* is extended to contain the handoff notification message. Therefore, S-MIP has an equal setup signaling cost. However, compared to fast-handoff, *PrRtAdv* message in S-MIP needs to wait for the arrival of the Handoff Decision (*HD*) message in addition to the expected Handoff Acknowledgement (*HAck*) messages. Fortunately, as shown in Appendix C, although a MN needs to wait approximately 3 seconds for the *HD* message, it is most likely still just arriving at the half way of the minimum overlapping distance (at the very most), namely half way between the overlaps. Therefore, the delay in receiving the *PrRtAdv* message by MN will not impact on the handoff performance considering the subsequent setup that follows can be achieved in the order of tens or, at most, hundreds of milliseconds.

The second part of the signaling requirement for S-MIP is associated with movement tracking, in particular, the *CLS* and *CTS* messages. The *CLS* messaging introduces additional overhead as ARs need to reply to the modified Router Advertisement message with the DE option, approximately once every 3 seconds (defined by the *MinRtrAdvInterval* in [13]). The *CTS* messaging introduce additional overhead as the MN is required to send *CTS* messages to DE, every second, via the current care-of access router. This messaging is a two-step process. Firstly, the MN piggybacks the tracking information (SS and AR Id) and sends it to the AR, as part of the L2 wireless messaging. The AR in turn then packages this information as a *CTS* message and sends it to the DE[3]. Therefore, the 'real' additional messaging overhead occurs only between the AR and the DE, and that *CTS* messaging only take place during a handoff. This is in fact what the design is intended to achieve, namely, to ensure signaling overhead (*CTS* and *CLS*) occur only at the wired portion of the S-MIP architecture. Wired network links, when compared with wireless, have far more bandwidth resources.

V. CONCLUSION

Analyses of Hierarchical MIP such as [10] and [5] showed that it is possible to further optimize the handover performance of Mobile IP. This can be done in large indoor areas by taking into consideration a hierarchical Mobile IP architectural layout together with the movement patterns of mobile devices. S-MIP provides a unique way of combining a novel location tracking scheme and the hierarchical MIP style handover scheme. In this paper, it was shown that this combined scheme can provide lossless handovers at the IP layer, with minimal increase in signaling overheads compared to [4]. Furthermore the paper provided an analytical proof of the viability of the proposed location tracking mechanism and the validity of the newly developed handover algorithm. Together these confirm the overall viability and show that S-MIP is capable of providing an effective seamless handover in Mobile IP. In future work, we plan to study S-MIP under complex multiple connection scenarios where connections interfere with each other, as well as, to examine the effect of this on the distribution of packet inter-arrival time at the MAP and the subsequent impact on the S-MIP architectural design.

[3] It must be noted that the number of piggyback messages sent from MN to AR, containing SS value and node Ids, are around four per second [14]. It is up to the AR to sample these values and package the 'best' representation to DE every second.

REFERENCE

[1] C. Perkins and D. Johnson, "Mobility Support in IPv6," Internet Draft, IETF, March 2002. Work in Progress.

[2] C. Perkins, "IP Mobility Support," RFC 2002, IETF, October 1996.

[3] G. Dommety et al., "Fast Handovers for Mobile IPv6," Internet Draft, IETF, March 2002. Work in Progress.

[4] H. Soliman, C. Castelluccia, K. Malki, and L. Bellier, "Hierarchical MIPv6 Mobility Management," Internet Draft, IETF, July 2002. Work in Progress.

[5] J. Hunskaar and T. Lunde, "Mobility in IPv6," Graduate Master Thesis, Hogskoleniagder, 2001.

[6] J. Kempf and J. Wood, "Analysis and comparison of Handoff Algorithm for Mobile IPv4," unpublished.

[7] N. B. Priyantha, A. K. L. Miu, H. Balakrishnan, and S. Teller, "The Cricket compass for context-aware mobile application," in *Procedings of MOBICOM*, 2001.

[8] "OriNoCo Manager Suit User Guide," http://www.orinocowireless.com.

[9] P. Bahl and V.N. Padmanabhan, "RADAR: An in-building RF-based user location and tracking system," in *Procedings of INFOCOM*, 2000.

[10] R. Hsieh and A. Seneviratne, "Performance analysis on Hierarchical Mobile IPv6 with Fast-handoff over TCP," in *Procedings of GLOBECOM*, Taipei, Taiwan, 2002.

[11] S. Thomson and T. Narten, "IPv6 stateless address autoconfiguration," RFC 2462, IETF, December 1998.

[12] T. Larsson, Y. Ismailov and A.A. Nilsson, "Performance Characteristics of Multiplexed TCP Connections," in Proceedings of SPECTS, 2000.

[13] T. Narten, E. Nordmark, and W. Simpson, "Neighbour Discovery for IP Version 6 (IPv6)," RFC 2461, IETF, December 1998.

[14] T. S. Rappaport, *Wireless Communications: Principles and Practice*, Prentice Hall, 2000.

[15] "The Network Simulator – *ns* (version 2) Website," http://www.isi.edu/nsnam.

[16] Z.-G. Zhou, R. Chen, P. Chumchu and A. Seneviratne, "A Software Based Indoor Relative Location Management System," in *Procedings of Wireless and Optical Communications*, Canada, 2002.

APPENDIX

A. Minimum Overlapping Distance

Problem.

Show that minimum distance of intersection of three access routers (AR) in terms of effective coverage radius, *r*, and the overlapping radius, *d*.

Assumption.

1. The center of the three ARs form an equilateral triangle ABC, with three good coverage circles intersect at point S as shown in Figure 8(a).
2. The coverage area of an AR is a pure circle with a radius r.
3. The overlapped radius, d, is less than half of the effective radius, *r*, i.e. $d < r/2$.

Solution.

Suppose the radius of these three ARs overlapped by an amount of *d*, then the intersection of the effective coverage circles forms an overlapping area with three vertexes, A', B' and C', as shown in Figure 8(a).

Since the overlapped radius is the same for all three ARs, the overlapping area forms an internal equilateral triangle ΔA'B'C', which is similar to ΔABC. Point S is also the center of ΔA'B'C'. We need to show that the distance, *A'B'*, in terms of *d* and radius, *r*.

Since ΔA'B'C' is an equilateral triangle, so ∠A'SB' is $2\pi/3$. We enlarge ΔA'B'C as shown in Figure 8(b), the distance of *A'B'* is given by the following. By sine rule:

$$\frac{\sin\left(\frac{2\pi}{3}\right)}{r} = \frac{\sin(a)}{r-d}$$

$$\therefore a = \sin^{-1}\left[\frac{\sqrt{3}}{2}\left(1-\frac{d}{r}\right)\right] \quad \text{and} \quad b = \frac{\pi}{3} - a$$

Thus, $L = r \cdot \sin(b)$

$$\therefore A'B' = 2L = 2r \cdot \sin\left\{\frac{\pi}{3} - \sin^{-1}\left[\frac{\sqrt{3}}{2}\left(1-\frac{d}{r}\right)\right]\right\}$$

Therefore, for typical *r* = 25m, if we assume that *d* = 1m, then *A'B'* = 3.28m, if *d* = 2m then *A'B'* = 6.25m.

B. Overlapping Distance of Marginal Coverage

Problem.

We want to compute the overlapping distance of two effective circles, i.e. the distance of *YC'*, *ZC'* and *YZ* in Figure 9, and show that the distance is short enough for the marginal coverage area to cover.

Solution.

Since three internal (good coverage) circles with radius *r-d* intersect at center point S and the centers of these three circles form an equilateral triangle ΔABC,

$$\therefore XS = r - d \Rightarrow YS = \frac{r-d}{2}$$

$$\therefore YZ = YS - d = \frac{r-d}{2} - d = \frac{r-3d}{2}$$

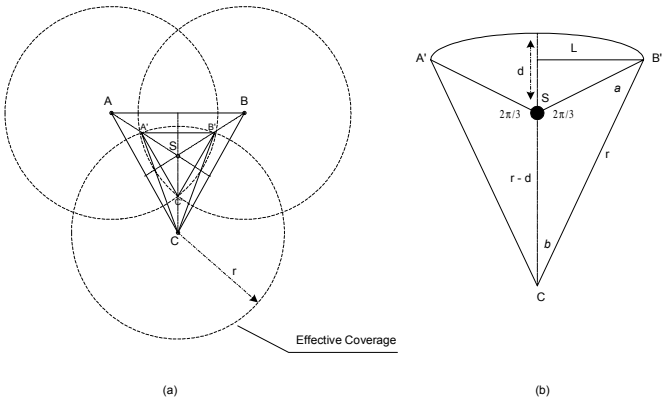

Figure 8 Minimum Overlapping Distance

From previous computations, we know that $A'B' = 2L$ and $\angle$A'SB' is $2\pi/3$. For an equilateral triangle,

$$\angle ZSB' = \frac{\pi}{3} \text{ and } SC' = SB'$$

$$\sin(\angle ZSB') = \frac{\frac{1}{2}A'B'}{SB'} = \frac{L}{SC'} \Rightarrow SC' = \frac{2\sqrt{3}}{3}L$$

$$YC' = YS + SC' = \frac{r-d}{2} + \frac{2\sqrt{3}}{3}L$$

$$\therefore QC' = 2YC' = r - d + \frac{4\sqrt{3}}{3}L \quad \text{and}$$

$$ZC' = d + SC' = d + \frac{2\sqrt{3}}{3}L$$

Using the data from the Appendix A,

For r = 25m and d = 1m,

YC' = 13.89m, ZC' = 2.89m and YZ = 11m

For r = 25m and d = 2m,

YC' = 15.1m, ZC' = 5.6m and YZ = 9.5m

A distance of 9.5m for YZ is short compared to the distance that the marginal coverage can provide. Marginal coverage is at least equal to that of the effective radius, namely, 25m, when assuming SS exhibits negative log function behavior.

C. Minimal Period for Movement Pattern Detection

Problem.

Find out the relationship between the minimal overlapping distance, $A'B'$ and the coverage radius, r.

Solution.

From the solution in Appendix A, we can see that the distance $A'B'$ is in terms of r, and d for $d < r/2$. Also, we know that the signal attenuation due to human body is 6.4dB and due to orientation of mobile device receiver is 9dB on average [16]. These equates to a difference in distance greater than 5m. Hence, the radius difference of the good coverage circle (internal circle in Figure 9) and the effective coverage circle (external circle in Figure 9) for a single AR must be greater than 2m (i.e. $d > 2$m). However, in order to maximize the efficiency of the AR, it is necessary to make the coverage area as large as possible. This means that, we want to minimize the radius difference, d. Therefore, it is reasonable to keep d as a constant and find out the relationship between $A'B'$ and r.

Now suppose d is a constant and let d = 2, therefore $r >$ 4m, as $d < r/2$. Figure 10 shows the relationship of r and $A'B'$ varies from r = 4m to r = 200m. As we can see from the curve, when r is increased up to about 50m, the distance of $A'B'$ tends to a constant level which is about 7m. And the difference of $A'B'$ between r = 25m and r = 50m is only 0.3m. Therefore, it is reasonable to assume that the distance $A'B'$, in our context, is about 7m. The distance $A'B'$ converges to a constant when r is greater than 20m. This computation shows that even through typical 802.11 coverage radius, r, varies from 25m to 50m, the minimal overlapping distance, $A'B'$, does not change significantly, i.e., within 1m. Therefore, the assumption of 7m for $A'B'$ is valid. The movement speed in the indoor environment is relatively slow and we assumed a constant speed of 1m/s. In this case, if the location sampling period is 1 sample/s, then after 3 samples, the mobile device is still located near the center of the two overlapping zone, i.e. at most half way in between $A'B'$. Strategically, this is the best point in making the handoff decision. Hence, we show that in the worst case scenario, i.e. linear movement across the minimum overlapping distance path, 3 samples as a minimal are still sufficient for the movement pattern detection.

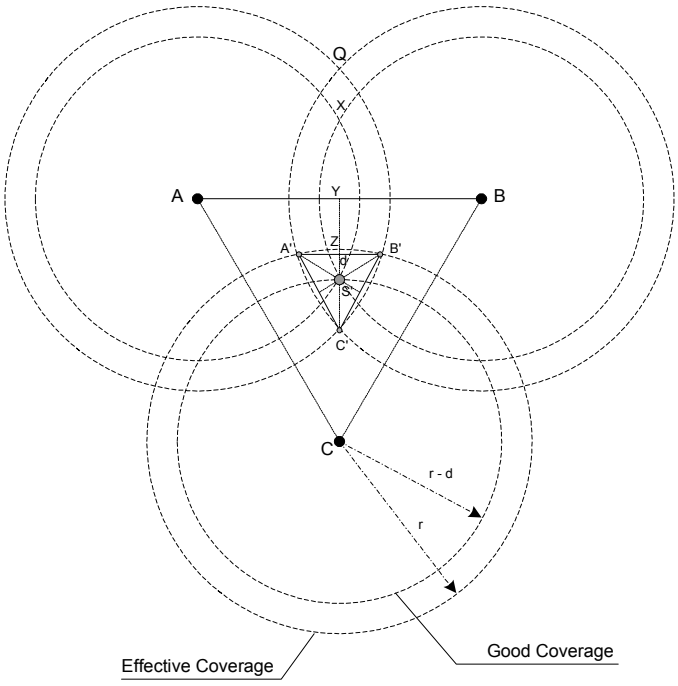

Figure 9 Overlapping Distance of Marginal Coverage

D. S-MIP Architectural Validity

Problem.

Show that packet out of order must occur within the hierarchical MIPv6 architecture at the nAR.

Assumption.

1. As Assume that $T_{in} < D_h$ where T_{in} is the inter-arrival time between 2 packets, and D_h is the hop delay for *Path1* and *Path2*. See Figure 11.
2. Since the packet manipulation processing time $T_{process} << D_h$, we ignore $T_{process}$.
3. Since the bandwidth inside the core network is greater than that of at the edge network, we assume that the data rate in *Path3* is not less than *Path1* or *Path2*.

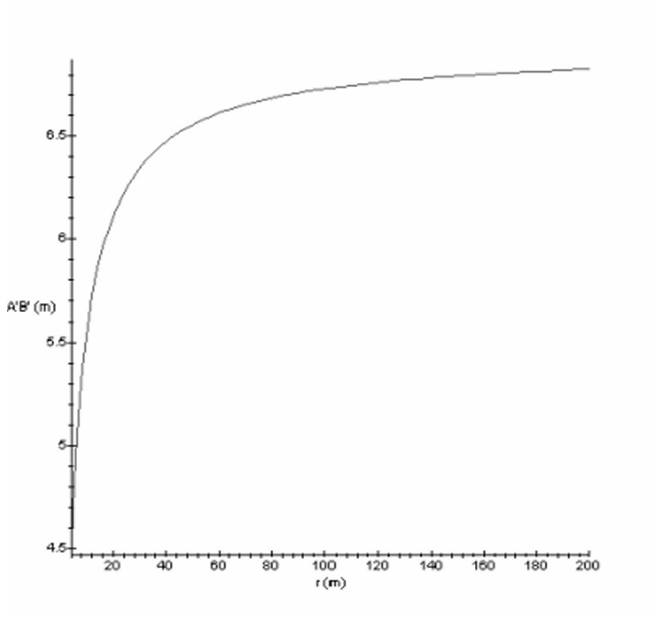

Figure 10 Minimum Overlapping Distance (A'B') versus Radius (r)

Solution.

Let packet with the sequence number *n* be the first forward packet from oAR back to MAP originated from MAP. Let T be the MAP arrival time for the packet originated from MAP destined for oAR, but forward from oAR to nAR through MAP. For an end-to-end transmission using transport protocol with sequencing mechanism, i.e. TCP,

$\forall$ i,j $\in$ S : i < j $\Leftrightarrow$ $t_i < t_j$
where S is the set of sequence number and t is the packet arrival time

$\because$ $T_{in} < D_h$ and the time interval for n traveling from MAP to oAR and return to MAP is $2D_h$

$\therefore$ $\exists$ *m* packets in Path1 and Path2 for some $m \geq 2$

Case 1: *n* and *n+m+1* arriving at MAP simultaneously

If MAP process packet *n* first

$\because$ $m \geq 2$

$\therefore$ some packets must exist in either Path1 or Path2, so that, $t_{n+m+1} < T_{n+m}$

$\Rightarrow T_n < t_{n+m+1} < T_{n+m}$ i.e. packet with sequence number *n+m+1* is in between *n* and *n+m*, showing out of sequence

If MAP process packet *n+m+1* first

$\therefore$ $t_{n+m+1} < T_n$

We need to prove that $t_{n+m+1} < T_n < T_{n+1} < \ldots < T_{n+m} < t_{n+m+2}$ can not exist.

Suppose *n+m+2* has the longest delay which is D_h^- but packet *n+m* is still in Path1. Since $T_{in} < D_h$ and packet

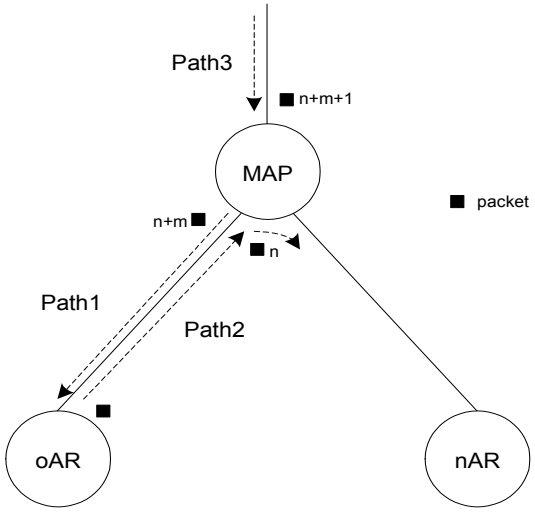

Figure 11 Packet out of order Model

n+m+1 just arrived at MAP so that packet *n+m* does not reach oAR yet.

$\Rightarrow T_n < t_{n+m+2} < T_{n+m}$

Case 2: *n* arrived at MAP but *n+m+1* still in the distance

Similar to second scenario in Case 1, *n+m* is still in Path1,

$\therefore$ $T_n < t_{n+m+1} < T_{n+m}$