RA Beaconing Optimization for Fast Handoff in IPv6 Wireless Networks

G. De Marco*, K.Sugita, L.Barolli
Department of Information and Communication Engineering
Fukuoka Institute of Technology (FIT)
3-30-1 Wajiro-Higashi-ku, Fukuoka 811-0295, Japan
tel: +81-92-606-4970
e-mail: {demarco,sugita,barolli}@fit.ac.jp

***Abstract*— In this paper we present an original approach to speed up the auto-configuration of mobile nodes in Mobile IPv6. The handoff process associated with movement of mobile nodes in IP networks has to face the delays arising to guarantee the reachability of the node, while preserving packets to be lost because of the movement. Here we deal with the reduction of address auto-configuration delay. This could be reached by means of a modulation (beaconing) of the frequency of Router Advertisement (RA) messages sent by the Access Router. The algorithm is based on a classical LMSE prediction. Simulations confirm that the beaconing of RA messages can decrease by 50% the handoff setup time due to address configuration. Moreover, the tradeoff between bandwidth-occupation/handoff-delay is acceptable, with respect to other schemes based on constant rate of RA messages. Finally, we give directions of further improvements for cases not considered in the present work.**

I. INTRODUCTION

In cellular IP networks, a mobile node (MN) traveling towards a new subnet needs to correctly configure the addresses of its Network Interface Card (NIC) in order to establish and/or maintain IP layer communications with its peers. To this aim, the MN needs information by Access Router (AR) carried in ICMP messages called RA. Usually the MN initiates handoff process by soliciting an RA message to the new AR. This is the case of *mobile-initiated handoff*. On the other hand, if the ARs (i.e. the network) knew the arrival time of the next handoff instance, the configuration delay of the IP address of the MN could be reduced. This is the case of *network-initiated* handoff.

Many researchers proposed different solutions to reduce the time an MN take to complete the handoff process. For example, in [1] the authors suggest to use a learning algorithm to predict future location of MN. The algorithm is based on a Hidden Markov Model and could be used to predict the time arrival of MN to the next Access Point (AP) as well. However, the authors did not address consider the current protocols.

In this paper, we propose to use the classical Least Mean-Square Error (LMSE) filtering to predict the time instants of handoff instances. Consequently, we realized that a proper modulation of the frequency of the RA messages could speed up the address auto-configuration process. Moreover, simulations confirm that the cost function of handoff setup time tends to the minimum value 0.5, which means an halving of the address autoconfiguration part due to the reception of an RA message. The trade-off between bandwidth-occupation/handoff-delay is acceptable, with respect to other schemes based on constant rate of RA messages.

The paper is organized as following. In Section II, we briefly review specifications of MIPv6 and fast handoff. In Section III, we illustrate the model used to test the performance of the beaconing technique and in Section IV-V we present some simulation results. Finally, in Section V, we give directions of further improvements for cases not considered in the present work.

*On leave from Department of Information Engineering and Electrical Engineering, University of Salerno, Italy

II. HANDOFF PROCEDURES IN MIPV6

Generally, wireless IP network components are grouped into ARs and APs. The APs are L2-bridges connecting wireless MNs to the wired IP network. The area within MNs are provided with IP wireless connection is divided into subnets, which in turn are managed by a dedicated AR. The duty of AR is to furnish connection towards peers outside the subnets. Basic functionalities of ARs are routing information updates by means of dedicated signaling messages.

We can distinguish two instances of handoff.

- Link-layer handoff

The link-layer (L2-handoff) takes place every time the MN changes parameters of the link layer connection, e.g. when it moves from one AP to another. The typical delays are due to the following phases[1]: detection of transmission failures over the current link, searching for new channels by means of either active or passive probing, and, eventually, executing re-authentication procedures, which in total could come up to $2s$. Measurements on commercial IEEE 802.11b cards as well as proposals to speed up these phases have been done in [2], [3], [4].

- Network handoff

The aim of network layer handoff (L3-handoff) is guaranteeing the global *visibility* of MN, i.e. the global routability of packets. This goal is accomplished by sending binding update messages of new Care of Address (CoA) to routing agents.

[1]If we consider WLAN-like technologies

The L3-handoff is not performed if MN moves among APs belonging to the same subnet. To detect the change of subnet, the MN periodically receives RA messages from the current AR, responsible of the visibility of MNs within a certain subnet. These messages contain several information, among which there is the link address prefix of the subnet which the message is referred to. However, MIPv6 furnishes advanced mechanisms for the address auto-configuration process which is simpler than in MIPv4 and can be completely taken automatically [5]. The fast handover mechanisms of MIPv6 also have procedures to 1) signal imminent handoff between two ARs and 2) reserve buffer space at both old AR (oAR) and new AR (nAR), by means of tunneling and fast binding updates [6]. The fast handover generally relies on the availability of L2 trigger signals.

Depending on the technology, the handoff can be smooth, i.e. with low latency perceived by higher layers, and/or seamless, i.e. without disconnection from the oAR and with temporary joint connection to both oAR and nAR, as shown in Fig. 1-b. The IEEE 802.11 WLAN uses hard hanoffs, i.e. the MN is allowed to communicate with one AP at time [7] [8].

From the above discussion, we stem that the overall handoff delay is:

$$T_{handoff} = T_{L2} + T_{L3} = \\ (T_d + T_s + T_e) + (T_A + T_{BU} + T_S) \ , \quad (1)$$

where the L2-handoff (T_{L2}) parts are T_d, T_s, T_e i.e. the delays of the detection phase, searching phase and execution phase, respectively. The other terms, T_A, T_{BU}, T_S account for the delays of auto-configuration phase, address bindings updating and security re-association phase, respectively[2].

It is well known that, as one adopts current technologies like IEEE 802.11b and the MIPv4/v6 protocol, the overall handoff delay $T_{handoff}$ reaches values not tolerated for any real-time applications. In any case, both changes to routing and bridging components of the network, i.e. the ARs and/or APs, and to the infrastructure are needed if multimedia and delay sensitive data are to be carried in such networks, as argued also in [10].

A. Movement Detection and Fast handoff

After the reception of an RA message, the MN can check if it is moved to a different subnet. Simply, the MN compares the link address prefix information of RA with that previously stored in its cache. If the comparison fails, a mobile-initiated handoff takes place: The MN sends a multicast message, the Router Solicitation for Proxy Advertisement (RtrSolPr)[3] or simply an RS message. If the MN can receive L2 frames from the nAP before it disconnects from the current AP, the RS message can be sent to the oAR in order to request AR information (AR-Info) associated to the nAP identity (AP-Id).

[2]We are aware of the additional delay caused by the Detection of Address Resolution protocol, which could add up to $2s$ (see for example [9]). Moreover, other additional delays are irrelevant to the main focus of the paper, i.e. the propagation delays due to the reception of advertisement messages from AR.

[3]It can be also a unicast message.

The AR receiving the RtSolPr sends back a unicast Proxy Router Advertisement (PrRtrAdv), or simply an RA, with the requested information in the form (AP-Id,AR-Info). This RA is delayed of random value between 0 and $500ms$[11]. If this 2-way handshakes is being done before the disconnection from the current AR, a real fast handoff takes place. The fast handoff is possible whenever the non ideal shape of radio coverage of APs allows the MN in the overlapping area to hear signals from different APs.

In the case of network initiated handoff, the AR sends *unicast* and *unsolicited* RA messages to the MN whose an imminent handoff from the current subnetwork is predicted. Network initiated handoff should track all mobile nodes in a subnet and AR should implement some prediction technique to estimate the future (immanent) location of MN as well as the time instant of the handoff.

Generally, RA messages are sent to multicast address with a constant frequency ranging from `MinRtrAdvFreq` to `MaxRtrAdvFreq`. The last Neighbor Discovery protocol draft [12] reduces this values to $0.03s$ and $0.07s$, respectively.

If the ARs can estimate, with some tolerance, the arrival time of the next handoff event, they could send multicast RA messages freeing the MN from sending RS messages. If the estimaton error is small “enough”, this would result in a decrease of 50% of the address auto-configuration delay T_A. We show in the next Section that for simplicistic assumptions on the handoff arrivals, LSME suffices to this aim. In the Conclusion we give some directions for more complicated scenarios.

III. Proposed technique

With IPv6 in mind, we show how to reduce the address auto-configuration part T_A in Eq. 1. Our objective is to refrain the MN from sending RS message and then waiting for a (solicited) RA message. For this reason, our scheme is based on unsolicited and *multicast* RA messages.

A. Model assumptions

A conceptual model of the network architecture is given in Fig. 1, where a generic AR, labeled with (i, j), is responsible of a certain group of cells. As example, in this figure we divide the subnet in four APs and we suppose that overlapping areas extend around all the boundaries of the wireless subnet. In Fig. 1, we skecthed a simplified timing diagram of fast handover procedures. As comparison, in Fig. 1-a, we report the operations of fast handover, which generally requires L2 triggers at MN to detect the presence of nAP. Potentially, in the overlapping area, the MNs can receive data from both adjacent APs, both belonging to the same subnet or not, as shown in Fig. 1-b. If the MN is allowed to jointly communicate with adjacent APs, our scheme will result in a seamless handover scheme, otherwise it is a smooth handoff scheme only. We suppose that the joint communication with APs is possible, although the following concepts could be applied to hard handoff schema too. Further, we assume that when the re-association procedure is accomplished, the APs are capable

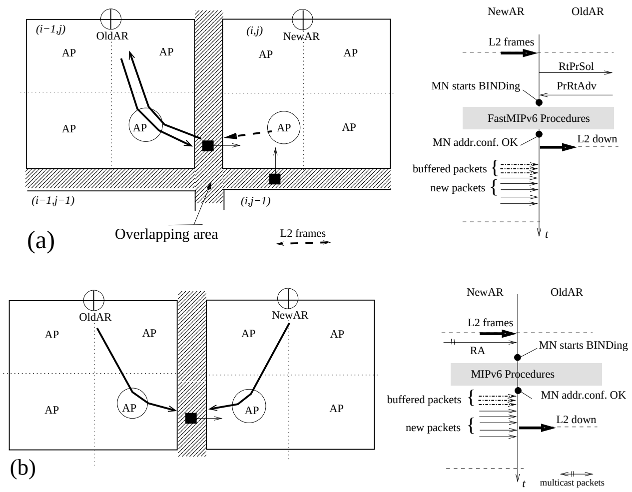

Fig. 1. Conceptual Example of communication when MN traverses the overlapping area, in a) it is supposed that MN can use L2 triggers to detect a nAP, while in b) the MN can jointly communicate with both adjacent APs.

to inform AR of this event. We call this event as a *handoff arrival* or *handoff request* to the nAR.

The mobile nodes handing out from adjacent subnets, e.g. $(i-1, j)$ and $(i, j-1)$, will generate handoff arrivals in the (i, j) AR. The frequency of these arrivals inside the APs depends on the radio coverage of APs and service deployment inside them. But, we assume that the handoff request intensities are the same among all the APs, i.e. the handoff intensities are uniformly distributed among the APs. Moreover, we consider handoff requests are being generated by MNs which *are entering* the (i, j) subnet.

B. Prediction of Handoff Arrivals

Let be $\{t_k\}_{k\in\mathbb{N}}$ the process of handoff requests of MNs coming from the adjacent subnets . The overall handoff request process $\{x(t)\}_{t\in\mathbb{R}}$ for the (i, j) subnet can be modeled as:

$$x(t) = \sum_k \delta(t - t_k)\,. \tag{2}$$

where $\delta()$ is the Dirac function. From this process we consider the discrete-time signal $X[n]$, representing the arrival time of the next handoff request:

$$X[n] = X[n-1] + T[n]\,, \tag{3}$$

where $T[n]$ is the discrete-time signal associated with the interarrival time between two handoff requests, i.e. $T[n] = t_{n+1} - t_n$. Given a sequence of values of handoff arrival times up to $X[n]$, we shall predict the value of $X[n+1]$. Let us note that in general the signal $X[n]$ is *not predictable,* which means that the variance of the prediction error is not zero. Moreover, from an implementation point of view, it is difficult to store always increasing values of $X[n]$. We suggest to predict $T[n]$ instead of $X[n]$.

The statistical properties of the signal $T[n]$ depend on the behavior of MNs inside a cell, i.e. the residence time within a cell. There are many proposed statistical models to capture the impact of cell residence time on resource management algorithms [13]. However, in this study, we assume that $T[n]$ be a stationary process with exponential distribution.

The prediction of $T[n]$ is based on the classical Least Mean-Square Estimation (LMSE). That is, the estimate of $T[n]$ given its P observations up to time $n-1$ is[4]:

$$\hat{T}[n] = \sum_{k=1}^{P} a_k T[n-k]\,. \tag{4}$$

By applying the orthoganility principle, the a_k coefficients of the prediction filter in the Eq. 4 are given by solving the Normal Equation for the correlation matrix[14].

The frequency of the RA messages is derived from the interarrival time of RA messages, $\hat{X}_{RA}[n]$:

$$\hat{X}_{RA}[n] = X[n-1] + \hat{T}[n]\,. \tag{5}$$

This equation represents the beaconing of the RA messages. If the RA message arrives *too late*, the MN will send an RS message, otherwise it could halves part of the address configuration delay T_A in Eq. 1. The Fast MIPv6 draft [6] does not specify when it is time to send RS message, i.e. we do not know what the expected delay between the end of the L2 re-association procedure with nAP and the sending time of RS message is. In this work, we suppose that this delay is constant and known, ϵ_{RS}. Thus, the RA beaconing scheme is effective if the prediction error $\epsilon = \hat{T}[n] - T[n] < \epsilon_{RS}$[5]. Otherwise, we consider that an *RA fail* has occured.

C. Metrics

The beaconing of RA messages can reduce handoff setup time at the cost of bandwidth occupation in the subnet. On the other hand, if we used a constant frequency for RA messages, as suggested by the IETF draft documents, we obtain an increase of the bandwidth occupation of the signaling messages along with a diminuition of the handoff setup time. To measure this trade-off, we use two cost functions, computed for both schema. Let us define F the bernoullian random variable (r.v.) representing the RA fails accordingly to the RA scheme used. That is, for every n , the outcomes of F are $F_n = 1$ if an RA fail occured, otherwise $F_n = 0$. The cost of handoff setup time is computed as the ratio between the handoff delay weighted by the probability of an RA fail :

$$C_H = \frac{\sum_{n=1}^{n_{tot}} F_n}{n_{total}} + \frac{(n_{tot} - \sum_{n=1}^{n_{tot}} F_n)}{2n_{tot}} = \frac{\left(1 + \hat{E}[F]\right)}{2}\,, \tag{6}$$

where n_{tot} is the total number of handoff requests along a simulation run, and $\hat{E}[F]$ is the sample mean of F. The worst

[4]We suppose that observations from an infinite past are available. We do not deal here with prediction with increasing order of the filter in the Eq. 4.

[5]Although the reasoning is the same, we do not contemplate propagation delays from the MN nodes and the AR.

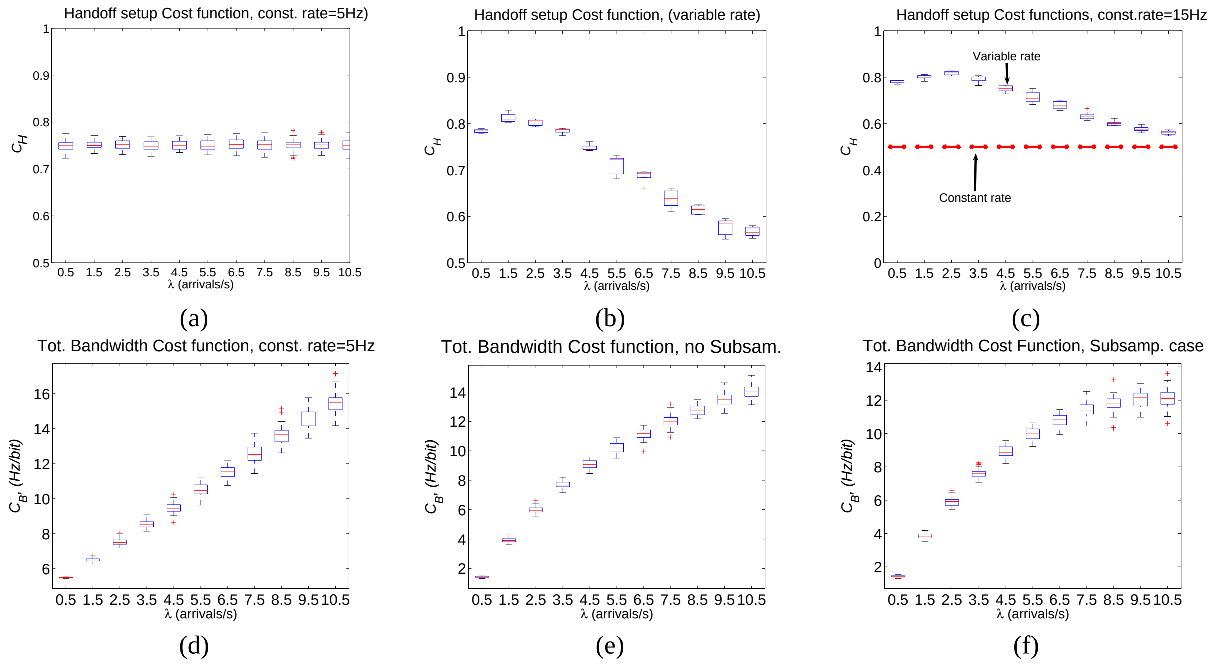

Fig. 2. Comparison of Cost functions, in (b)(c)(e) the scheme with the beaconing of RA messages (with $P = N$) and in (a) with constant frequency ($5Hz$). The subsampling of RA messages is in (f)

case is when $C_H = 1$, because there is no improvement of the technique with respect to the RS-RA handshake mechanism (cfr. II-A). On the other hand, the lower the number of RA fails is, the more $C_H \to 0.5$ does, and, consequently, the handoff setup part, T_A, is halved. Let note that $C_H = \frac{(1+P\{F=1\})}{2}$ .We could have been defined the range of C_H between 0 and 1, by simply defining $C_H = P\{F = 1\}$, but we use Eq. 6 to emphasize the halving of the handoff delay with respect to the standard scheme. The cost function related to the bandwidth occupation is composed of two terms, $C^U$and C^D, i.e. the costs of uplink and downlink bandwidth occupation, respectively. We name C_{RS} the cost due to the transmission of RS messages, wich in turn affect the uplink bandwidth. Conversely, C_{RA} is the measure of the downlink bandwidth occupation due to RA message. Let us note that whenever an RA fail occurs, we have to send one more (solicited) RA message, because the MN sends an RS message. Thus, the overall cost function of the bandwidth occupation is:

$$C_B = C^U + C^D = 2C_{RS} + C_{RA} \,, \tag{7}$$

If we form the variable F'_k, which represent the time instant of the $k-$th RA fail, and the variable $\Delta F'_n = F'_n - F'_{n-1}$, we have:

$$C_{RS} = \left\{ \begin{array}{cc} \frac{1}{\hat{E}[\Delta F']}, & \hat{E}[\Delta F'] > 0 \\ 0, & otherwise \end{array} \right\}, \qquad C_{RA} = \left\{ \begin{array}{cc} \frac{1}{\hat{E}[\Delta X_{RA}]}, & \hat{E}[\Delta X_{RA}] > 0 \\ 0, & otherwise \end{array} \right\}, \tag{8}$$

where ΔX_{RA} is the variable representing the interarrival time of two consecutive RA messages. The cost functions of Eqs. 8 and 7 are normalized with respect to network bandwidth and packet size of the messages.

IV. Simulation results

We performed some simulations of the proposed model in MATLAB. Given the order of the prediction filter and the value of ϵ_{RS}, a simulation run consists in the generation of a vector of handoff arrival times with intensity λ and the computation of cost functions. The size of the vector of handoff arrival times is $N = 500$. From the vector of handoff arrivals, we compute the vector of predicted ones, according to Eq. 5 and 3. For every handoff arrival time, we compute the vector of RA fails, F_n, with respect to the both constant frequency and RA beaconing scheme. Accordingly, we compute the vectors $\Delta F'_n$, ΔX_{RA_n}, and costs of Eqs. 8 and 6. The procedure is repeated over all the simulation runs, wich we set to 50. We present the statistical properties of cost functions by means of box plots[6].

For the case without beaconing, the RA frequency is set to $5Hz$ ($200ms$), which is lower than the values proposed in the newer standard. We assume $\epsilon_{RS} = 100ms$.

In the case of a filter with order $P = N$, we obtain the results shown in Fig. 2. For the range of simulation parameters we chosen, the scheme with beaconing of RA messages gives a minimum $C_H \simeq 0.6$ for $\lambda = 10.5 arrivals/s$ (Fig. 2-b-c). In other words, from the definition of C_H, we see that the reduction of C_H is 100% with respect to the worst case (full 2-way handshake), and it increases with respect to the

[6]For a set of data, the box plot shows inside the box the median value, the quantile intervals as edges of the box and the extents of the rest of the data.

constant rate scheme as λ increases[7]. Consequenlty, the delay component of the handoff setup time due to the reception of an RA message is halved as the handoff request intensity λ increases.

In the constant rate scheme, we have $P\{F = 1\} \approx P\{F = 0\}$ and then the C_H is roughly a constant (~ 0.75). The same is for the bandwidth occupation, as shown in Fig. 2-e, even if the magnitude of the reduction is lower than that found in C_H. As expected, in the constant rate scheme, the total bandwidth occupation increases with λ, because of the increase of C_{RS} in Eq. 7. In other words, in the constant rate scheme, the bandwidth occupation due to RA messages is constant, while the bandwidth of RS messages increases as the number of RA fails. Let remember, in fact, that after an RA fail occured the MN sends an RS message. The higher the number of fails is, the higher the number of the RS messages. We have a linear increase of the number of RS message with λ. In the case of variable rate of RA messages, we obtain a complementary exponential like increase for C_B.

The cost function C_B can be further decreased if we observe that not all the RA messages are necessary. In fact, the $n-$th RA message is unnecessary if its sending time $\hat{X}_{RA}[n]$ differs from the previous ones by up ϵ_{RS}. This is a sort of sub-sampling of the signal $\hat{X}_{RA}[n]$, i.e. we are selecting some of the original samples of $\hat{X}_{RA}[n]$. With this trick, the bandwidth cost function C_B behaves as in Fig. 2-f. Although it is not shown for the sake of space, but the same results hold for a $2-$order prediction filter, because of the autoregressive modeling of $X[n]$.

V. Discussion

The performance gains shown in the previous Section are obviously connected with the values of the simulation parameters. For example, if we increase the value of RA frequency in the constant rate scheme, we obtain the result shown in Fig. 2-c. The RA beaconing scheme performs worse than the constant rate scheme, because of the large error in the estimated handoff arrival times. It is worth noting that this does happen in general. Usually, the *lower* λ is, the higher the prediction error variance is. We are currently studying an alternative method based on the transmission of bursts of RA messages centered at the predicted handoff arrival times. Moreover, we are studying the performance of our technique for handoff arrival intensities not uniformly distributed among the cells of a subnet.

VI. Conclusion

In this paper, we presented a technique to reduce part of the handoff setup delay which is composed by two different parts. We focused on the contribute concerning the address auto-configuration delay, i.e. the time required to formulate and setup a new CoA. In particular, we proposed a dynamic variation of the sending time of RA messages in order to promptly inform MNs of the change in the subnet identity. The technique could be coupled with other optimization mechanisms to reduce the overall handoff setup time $T_{handoff}$. Besides the benefits we presented by the simulations, some problems are yet open. For example, in wireless networks which do not permit the concurrent transmission over multiple wireless paths (read as multiple APs), the technique should be implemented inside an AR tracking MNs handing out the current subnet. In this case, the knowledge of the ϵ_{RS} parameter is not required. The same happens if we assume the APs can embed AR information, as subnet prefix and AR IP address, inside the beacon frames. In that case, we would be able to differentiate the handoff request among APs. Architectural options are being currently studied around the CAR protocol [15].

In any case, the analysis of fast movements and ping-pong effects should be addressed. In the present paper, we have skipped possible additional delays and packet loss arising in the communication between MNs and ARs, and among ARs. The linear prediction technique could be re-written as a problem of estimation of a signal embedded in noise, which could be account for the erroneous (or absent) movement information transmitted to AR. Implemenation of the technique inside commercial and/or open devices is under investigation.

[7]In fact, the higher λ is, the more probable the closeness between an handoff event and an RA message is.

References

[1] J.-M. François and G.Leduc, "Learning movement patterns in mobile networks: a generic approach," in *Proceeding of European Wireless Conference'04*, February 2004, pp. 128–134.

[2] A.Mishra, M.Shin, and W.A.Arbaugh, "Context caching using neighbor graphs for fast handoffs in wireless network," in *Proceedings of IEEE INFOCOM'04*, vol. 23, March 2004, pp. 351–361.

[3] M. S. an A.Mishraand W.A. Arbaugh, "Improving the latency of 802.11 hand-offs using neighbor graphs," in *Proccedings of ACM MobiSys'04*, June 2004, pp. 70–82.

[4] H.Velayos and G.Karlsson, "Techniques to reduce the ieee 802.11b hadoff time," in *Proceedings of IEEE Conference on Communications*, June 2004, pp. 3844–3848.

[5] S.Thomson and T.Narten, *IPv6 Stateless Address Autoconfiguration*, 1998.

[6] R. Koodli, *Fast Handovers for Mobile IPv6*, IETF draft, 2004. [Online]. Available: http://www.ietf.org/internet-drafts/draft-ietf-mipshop-fast-mipv6-03.txt

[7] S.Sharma, N.Zhu, and T.-C. Chiueh, "Low-latency mobile ip handoff for infrastructure-mode wireless lans," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 4, pp. 643–652, May 2004.

[8] P. D. Cleyn, N. V. den Wijngaert, L. Cerdà, and C. Blondia, "A smooth handoff scheme using ieee802.11 triggers - design and implementation," *Elsevier Computer Networks*, vol. 45, no. 3, pp. 345–361, 2004.

[9] S.-H. Hwang, Y. Han, S.-G. Min, and C.-S. Hwang, "An address confi guration and confi rmations scheme for seamless mobiliy support in ipv6 network," in *Proceeding of WWIC'04*. Springer-Verlag, 2004, pp. 74–86.

[10] G. Daley, B.Pentland, and R. Nelson, "Movement detection optimizations in mobilie ipv6," in *Proceedings of ICON'03*, 2003, pp. 687–692.

[11] T. N. et al., *Neighbor Discovery for IP Version 6 (IPv6)*, IETF RFC 2461, Dec 1998.

[12] D.Johnson, C.Perkins, and J.Arkko, *Mobility Support in IPv6*, IETF RFC 3775, 2004. [Online]. Available: http://www.ietf.org/rfc/rfc3775.txt

[13] C.-H. Wu, H.-P. Lin, and L.-S. Lan, "A new analytic framework for dynamic mobility management of pcs networks," *IEEE Transactions on Mobile Computing*, vol. 1, no. 3, July 2002.

[14] C.W.Therrien, *Discrete Random Signals and Statistical Signal Processing*, P.-H. I. Editions, Ed. Prentice-Hall International Editions, 1992.

[15] M. L. et al., *Candidate Access Router Discovery*, IETF draft - work in progres, 2004. [Online]. Available: http://www.ietf.org/internet-drafts/draft-ietf-seamoby-card-protocol-08.txt