A Cross-layer Fast Handover Scheme For Mobile WiMAX

Ling Chen*, Xuejun Cai**, Rute Sofia***, Zhen Huang ****

Abstract—**The Mobile WiMAX standard (IEEE 802.16e-2005) brings wireless broadband to a new level due to the support of nomadism. Still, handover latency in Mobile WiMAX is an issue that may affect real-time continuity of application sessions. This is partially due to the Layer 2 scanning/ranging, as well as the network re-entry procedure, which may result in a latency of hundreds of milliseconds, far exceeding the requirement of typical real-time services (e.g., 150 ms for Voice over IP). In this paper, we describe a mechanism which incorporates information from several OSI Layers to speed up the Layer 2 handover. We show by means of simulations that this new mechanism can decrease the handover latency significantly, to less than 100 ms in most cases.**

Index Terms—**Cross-layer, Handover, WiMAX, Ranging**

I. Introduction

THE IEEE 802.16 standards which are the core of the *Worldwide Interoperability for Microwave Access (WiMAX)* define a system where *Mobile Stations (MSs)* correspond to end-user equipment and *Base Stations (BSs)* control activity within a *cell* range. A cell therefore includes one BS and several MSs which together form a point-to-multipoint infrastructure controlled by the BS. It should also be noticed that while the WiMAX MAC layer has been designed specifically for point-to-multipoint infrastructures, WiMAX can also be configured to support point-to-point and mesh topologies.

The most common WiMAX version in use is the so-called *fixed WiMAX* (IEEE 802.16-2004), which simply provides fixed wireless broadband access, thus enabling coverage in places where fixed broadband technologies are not accessible. The fixed WiMAX standard theoretically mentions rates of 70 Mbps (shared) or 2-10 Mbps per user with coverage of around 10 Km^2 [5].

The mobile WiMAX standard (IEEE 802.16e-2005) brings WiMAX to a new stage by incorporating *nomadic roaming at vehicular speeds*, i.e., it provides the means for users to roam between different WiMAX networks while keeping their application sessions active. But Mobile WiMAX advantages go beyond the mobility support. Mobile WiMAX also provides superior performance due to relying upon the upgraded *Orthogonal Frequency Division Multiplexing Access (OFDMA)* which, as multiplexing technique, supports multi-path environments, thus resulting in the ability to generate higher throughput and improved coverage.

The mobile WiMAX standard theoretically supports *peak* data rates of around 30 Mbps and average data rates between 1 Mbps and 4 Mbps [6]. Transmission distances range from a few hundred meters to up to 5 km, and mobility is offered to speeds up and beyond 60 km per hour with relatively low-cost universal hardware. Because of these promising characteristics in mobile environments, mobile WiMAX starts to be considered a powerful competitor for emerging beyond 3G and 4G wireless communication specifications.

As background for the integrated mobility process, IEEE 802.16e-2005 defines *handovers (HO)* between different mobile WiMAX networks. A HO is defined as the migration of a MS between air-interfaces of different BSs. The BS associated to the MS before the HO is called *serving* BS, while the BS associated to the MS after the HO is named *target* BS.

The HO process introduces a large number of interactions between the MS and adjacent (*neighboring*) BSs for the purpose of scanning, ranging, parameter negotiation and information exchanging. In this process, which may take hundreds of milliseconds, the MS leaves its previous channel thus temporally suspending running services.

Solutions [1] proposed to improve the mobile WiMAX HOs fail to address the latency issue. While the scheme in [1] is limited to the *downstream* (from BS to MS) direction, the algorithms proposed in [2] oversimplify the network topology acquisition procedure and thus impair the accuracy of HO.

In this paper, we provide an alternative, cross-layer solution which reduces the mobile WiMAX HO latency. Specifically, we use Layer 3 to transmit MAC control messages between the MS and the BS during the HO. We show, by means of simulations, that our solution can effectively reduce the Mobile WiMAX HO latency.

The remainder of this paper is organized as follows. Section 2 introduces related work. Section 3 provides the description of our concept, while section 4 covers the performance evaluation. We finally conclude the paper in section 5.

II. Related Work

[1] proposes an enhanced link-layer handover algorithm where the serving BS forwards downstream data to the neighboring BS being ranged; therefore the MS can receive data downstream as soon as it becomes synchronized with the neighboring BS. An obvious inefficiency of this scheme is that it cannot reduce HO latency in the *upstream* (from MS to BS)

*Ling Chen (chenling@siemens.com) is with Nokia Siemens Networks, RTP NT, ZPark, 1, Beijing 100094, P.R.China (+86-10-64766507; fax: 86-10-64764717)

**Xuejun Cai (xuejun.cai@siemens.com) is with Nokia Siemens Networks, RTP NT, ZPark, 1, Beijing 100094, P.R.China, P.R.China

***Rute Sofia (rute.sofia@siemens.com) is with Nokia Siemens Networks, RTP NT NCT, Otto-Hahn-Ring, 6, Munich, Germany.

**** Zhen Huang (hzbeyond@gmail.com) is with Beijing University of Posters and Telecommunications

1-4244-0264-6/07/$25.00 ©2007 IEEE

direction, which is sensitive to interactive applications (e.g., Voice over IP).

[2] suggests using *Carrier-to-Interference plus Noise Ratio* (*CINR*) and *Arrival Time Difference (ATD)* to predicate the "best" target BS. This scheme therefore prunes "unnecessary" interactions with neighboring BSs other than that with the target BS. This method is highly effective in reducing the number of required interactions. However, it also prevents the MS from acquiring more precise information which would normally be obtained from complete ranging and could be decisive for the final BS selection.

[3] presents a cross-layer, fast-scanning concept for the IEEE 802.11 network. Their method relies on the use of an extended MAC-layer probe-request which contains the IP address of the MS and which is sent on every channel to be scanned. After this process the MS returns to its original channel and waits for responses (based on UDP) sent by candidate *Access Points (APs).* While waiting, the MS therefore can proceed with communication. Therefore, their scheme reduces latency by reducing the waiting period associated with the channel scanning process. The scheme has been specifically designed for IEEE 802.11 which is an asynchronous and connectionless technology. To adapt the scheme to mobile WiMAX (which can be seen as connection-oriented and asymmetric) careful synchronization and accessing considerations must be made. Furthermore, the IP probe-response messages are sent from the scanned AP(s) to the MS directly. This may pose security problems, given that a malicious user may act as a fake AP and obtain the MS IP address. A third drawback is that the scanned APs buffer the probe response for a fixed delay before sending it out so that the MS has time to reach its original channel; this introduces unnecessary latency.

Our work presents a fast HO scheme which addresses HO latencies in both upstream and downstream in mobile WiMAX scenarios. Because the MS still executes scanning/ranging with every neighboring BS, our scheme does not lose HO accuracy as [2]. By successfully removing the IP route between MS and ranged BSs and the fix delay of ranging response, our scheme has better security and performance than [3] in scanning/ranging. Finally, our scheme eliminates the network re-entry latency which has not been addressed previously.

The next section describes the issues with current mobile WiMAX HO in detail.

III. Mobile WiMAX Background and Handover Analysis

The Mobile WiMAX HO procedure includes several phases, namely, network topology acquisition and advertisement, association procedure, HO decision and initiation, target BS scanning and network re-entry. We provide details about each of these stages next, explaining what is their role in the overall MAC-layer HO latency.

A. Network Topology Advertisement

The BSs periodically broadcast *Mobile Neighbor Advertisement (MOB_NBR_ADV)* control messages (Fig. 1). These messages contain both physical layer (i.e., radio channel) and link layer (e.g., MAC address) information.

By means of such broadcasts, the MS becomes aware of the neighboring BSs. The MS then triggers the second-phase.

B. Scanning/ranging Procedure

In the second phase of HO, the MS scans and synchronizes with the neighboring BSs based on channel information from the neighbor advertisement. If the synchronization successes, it then starts the ranging procedure.

The scanning and ranging processes are shown in Fig. 1.

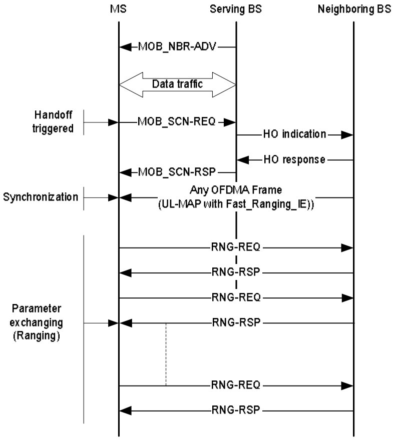

Fig. 1. Non-contention-based scanning/ranging process.

The MS first sends a *Mobile Scanning Request (MOB_SCN-REQ)* message to the neighboring BS with a potential target BS list (selected in the previous phase). The serving BS replies a *Mobile Scanning Response (MOB_SCN-RSP)* message to the MS to allocate a scanning duration. The serving BS may negotiate directly with the listed BSs the allocation of a unicast ranging opportunity. If successful, the ranging procedure can be non-contention-based as shown in Fig. 1. Else, the MS starts a contention-based CDMA procedure to be allocated a ranging slot by the neighboring BS. For the sake of simplicity, in analysis we always assume non-contention-based ranging.

Then the MS starts a hand-shake ranging procedure with the neighboring BS for the OFDMA uplink synchronization and parameter (e.g., transmission power) adjustment. This process may contain multiple message (*Ranging Request* (RNG-REQ) and *Ranging Response* (RNG-RSP)) transmission and parameter adjustment transactions. This procedure ends after the MS has completed ranging with all its neighbors,

In the ranging phase, a MS may switch to a new channel, thus temporally loosing connection with the serving BS. Although the MOB_NBR_ADV provides the MS channel information which accelerates its synchronization speed with neighboring BSs, ranging transactions incur a penalty in terms of latency which is analyzed later in section V.

C. HO Decision and Initiation

The HO trigger decision and initiation can be originated by both the MS and the BS using a *MS HO Request* message *(MOB_MSHO-REQ)* or a *BS HO Request* message *(MOB_BSHO-REQ)* respectively. Here we use the HO started by the MS as an example as illustrated in Fig. 2.

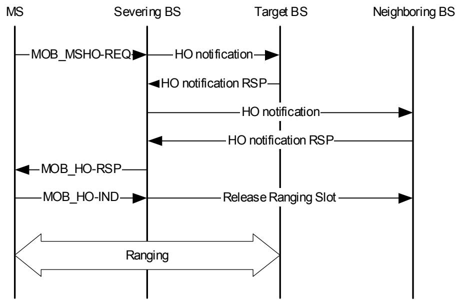

Fig. 3. HO Decision and Initiation.

The MS makes a decision about which BS(s) is (are) its target(s). A HO begins with when the MS sends a MOB_MSHO-REQ message to its serving BS indicating one or more possible target BSs. The serving BS may obtain directly from potential target BSs the expected MS performance at the target BSs through the exchange of HO indication and response messages.

After receiving a response from a target BS (MOB_BSHO-RSP), the MS notifies the serving BS about its decision to perform a HO by means of a *HO Indication (MOB_HO-IND)* message. The MS can also ask the serving BS to negotiate with the target BS the allocation of a ranging opportunity. If necessary, the MS may start ranging after HO initiation.

The HO decision and initiation process does not provoke connectivity break-up nor does it add latency. However, the possible ranging procedure after does introduce additional latency.

D. Network Re-entry

After all the physical parameter adjustments have been completed successfully, the network re-entry process is initiated to establish connectivity between the MS and the target BS. As defined in IEEE 802.16e, this procedure may include capability negotiation, authentication and registration transactions (cf. Fig. 3)

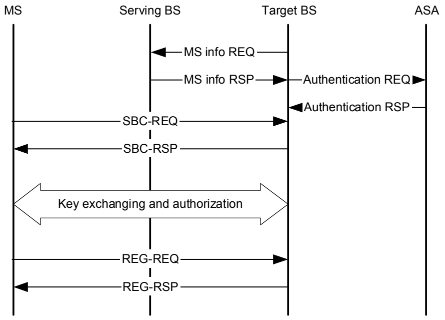

Fig. 4. Network re-entry procedure example.

Because the MS must wait until the re-entry procedure is completed successfully before it can restore communication, duration of this phase should be taken account into the entire HO latency.

The next section explains our cross-layer concept to speed up the HOs.

IV. Cross-layer fast handover concept

The goal of our fast handover concept is to reduce or eliminate the HO latency introduced by MAC layer management message exchanging (e.g., ranging, capability negotiation and registration) between the MS and the target BS. The key idea behind our concept is the use of Layer 3 to redirect and relay MAC-layer messages used during the HO. This is achieved by means of two tunnels. We establish a Layer 2 tunnel between the MS and the serving BS. A Layer 3 tunnel is established between the serving and the target BSs.

Compared with directly exchanging messages with the target BS, the benefit of our scheme is that it minimizes direct message transportation between the MS and neighboring BSs, a major source for latency in the entire HO procedure.

In following sections we introduce a fast scanning/ranging algorithm and thereafter, discuss how to tunnel transactions in the network re-entry procedure during HOs.

A. Fast Scanning/Ranging Algorithm

The fast scanning/ranging procedure is shown in Fig. 4:

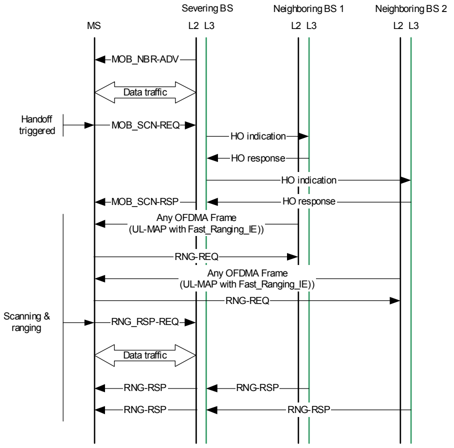

Fig. 2. Fast scanning/ranging operation example

As in the standard operation, the MS first requests a scanning interval to look for possible target BSs. Both the serving BS and the MS start to cache packets to be sent after the allocation. The serving BS then communicates with neighboring BSs to allocate a fast ranging opportunity.

In the fast scanning/ranging phase, the MS switches to the channels to be scanned one by one and tries to get synchronized. If successful, the MS sends Ranging Requests to the neighboring BS. If the MS fails to synchronize, it should consider the neighboring BS as invalid and no further actions are taken.

After sending ranging requests on all channels, the MS returns to its serving BS, sends a *Ranging Response Request (RNG_RSP-REQ)* message and restores the uplink data transmission. The RNG_RSP-REQ is used to inform the serving BS that the MS has returned and is able to receive Ranging Responses. The serving BS should then restore the downlink communication immediately.

The MS should start a timer to wait for RNG-RSP. If the MS receives no response until the timer times out after *minResponseTime*, it should conclude the ranging fails. While if the MS has not received all the expected responses before timeout, it should wait for *maxResponseTime*.

When the neighboring BS receives RNG-REQ, it sends a RNG-RSP to the serving BS through the backbone between them. The RNG-RSP message should be encapsulated in UDP/IP packet whose destination is the IP address of the serving BS. The serving BS should de-capsulate and buffer the payload (RNG-RSP). If it finds that it has received RNG_RSP-REQ, the serving BS should forward them to the MS in the MAC layer immediately. Otherwise, the serving BS should keep buffering until the RNG_RSP-REQ comes.

The fast ranging transaction may be repeated multiple times until both the MS and the neighboring BSs get consensus with the parameter setting as in the standard operation.

B. *Fast Network Re-entry*

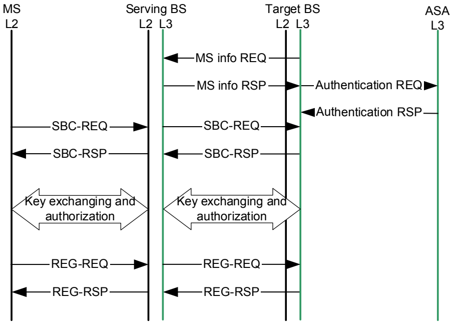

Fig. 5. Fast network re-entry

The fast network re-entry procedure is shown in Fig. 5:

Instead of directly communicating with the target BS, the MS sends all the messages to its serving BS which then relays them through the IP backbone to the target BS. In the counter-direction, the serving BS also relays responses to the MS; therefore, the serving BS works as a relay agent between the MS and the target BS. Here we use the capability negotiation as an example to show how the fast network re-entry works:

Firstly, the MS sends an *SS Basic Capability Request (SBC-REQ)* message which is extended with a target *BS identification (BSID)* item in the *Type-Length-Value (TLV)* field to the serving BS. The serving BS looks up in its neighboring table with the target BSID for the target IP address of the target BS. The serving BS then encapsulates SBC-REQ in a UDP message and sends it to the target BS.

Secondly, the target BS handles SBC-REQ in the UDP message as it is received directly from the MS. The target BSID item is ignored. The target BS then sends back *SS Basic Capability response (SBC-RSP)* to the serving BS in UDP.

Finally, the serving BS decapsulates the UDP message and forwards the SBC-RSP to the MS by the MAC address in it.

As in standard capability negotiation, the MS should also maintain a timer to monitor responses. We suggest the upper-limit of this timer should be larger than standard value since the L3 round trip delay between the serving BS and the target BS exists.

V. PERFORMANCE ANALYSIS

In this section, we provide a performance analysis for the concept described in section 4. The performance evaluation here provided was performed by means of simulations carried out with Matlab ^{TM}.

A. *Simulation Model*

In order to emphasize the theme of the paper, we build our simulations on a simplified but typical scenario: for the physical layer specification, we assume OFDMA/TDD is used and the OFDMA frame duration (T_F) is 5ms. We rely on a network topology which contains a serving BS and 4 neighboring BSs. The MS is aware of the neighboring BS list from neighbor advertisements before the HO. During scanning, the MS gets synchronized with 1~4 neighbors (following the discrete even distribution) and finally chooses one after several ranging (transmission power adjustment) transactions.

Both the conventional and the optimized HO introduce synchronization latencies. As defined in IEEE 802.16e, the MS must try at least $2T_F$ to get synchronized before ranging. We therefore assume it take the MS T_{SYNC} (a random valude following even distribution ($\mu(0,\ T_{SYNC\_MAX})$ and T_{SYNC_MAX} is $2T_F$) for initial synchronization and downlink quality estimation. If the synchronization is not successful after T_{SYNC_MAX}, the neighboring BS is assumed invalid and no further actions are taken.

The second part is the ranging transaction duration. In order to simulate the transmission power adjustment process, we establish a random model based on a normal distribution: the MS must estimate an initial ranging request transmission power (P). We assume the distribution of this estimation follows a normal distribution $N(P_{min},\ \delta)$. δ is the variance which is assumed as one power adjustment step. P_{min} is the lower limit for the successful ranging request transmission power. In this model, if and only if $P_{min} > P$, do the neighboring BS receive ranging requests. The MS must increase P by 1 step and retry until $P \geq P_{min}$.

In the conventional HO, the MS directly ranges with valid BSs. The total ranging transaction latency can be calculated as $\sum_{i}^{N_1}\sum_{j}^{M_i} T_{RNG_{(i,j)}}$, while N_1 is the number of valid neighboring BSs and M_i represents the number of ranging retrials for the i_{th} neighboring BS. $T_{RNG_{(i,j)}}$ corresponds to the ranging delay caused by the j_{th} ranging retrial for the i_{th} neighboring BS. $T_{RNG_{(i,j)}}$ is assumed 15~30ms in modest load condition ($2T_F$ roundtrip delay plus a random handling duration) and 30~50ms when the load is heavy (considering about longer transmission and handling durations).

In the optimized HO, the MS need not wait for ranging responses from the neighboring BS. Therefore, the ranging transaction latency is $M_i{*}T_{RNG\_REQ}$, while T_{RNG_REQ} is the ranging request transmission delay (T_F). The RNG_RSP transpoting latency through the backbone does not affect the ranging transaction latency since the MS is still served by the serving BS during ranging transactions.

The network re-entry delay in the conventional scheme (T_{REEN}, which is $6T_F$ for roundtrip dealy in capability negotiation, authentication and registration plus a random handling duation) is assumed to be 50ms in modest load and 100ms in heavy load. This latency is eliminated in the optimized scheme since the MS still keeps the communication with the severing BS during the network re-entry process.

As the result, global HO latencies of the conventional and optimiaed schemes can be calculated as:

$$T = \sum_{i}^{N_1}(T_{SYNCi} + \sum_{j}^{M_i} T_{RNG(i,j)}) + N_2 * T_{SYNC\_MAX} + T_{REEN}$$

$$T' = \sum_{i}^{N_1}(T_{SYNCi} + M_i * T_{RNG-REQ}) + N_2 * T_{SYNC\_MAX}$$

N_2 is the number of invalid neighboring BSs.

B. Simulation Results and Analysis

We run the simulation 1000 times to analysis effects of different parameters on the HO latency.

Fig. 6 shows a comparison of the global HO latency between the conventional solution and the optimized solution in the modest load mode. Fig. 7 shows the comparison in the heavy load mode.

The global HO latency is greatly reduced (more than 60% in modest load and 80% when load is heavy) in the optimized solution. The improvement comes from elimination of network re-entry delay and waiting time for RNG-REQ. Another advantage of our cross-layer scheme is that the performance does not degrade when the load increases.

Fig. 8 and Fig. 9 show the effect of the number of valid neighbors (N1) on the global HO latency. HO latency increases with the number of valid neighbors because the MS needs more time to range with neighboring BSs. Fig. 10 and Fig. 11 show the effect of ranging retrials on the global HO latency. HO latency increases with the number of ranging retails because every ranging retrial takes tens of milliseconds in conventional HO and $1T_F$ in optimized HO.

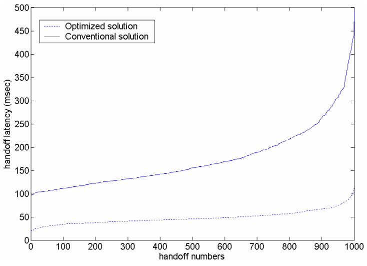

Fig. 6. Global Handover Latency (modest load)

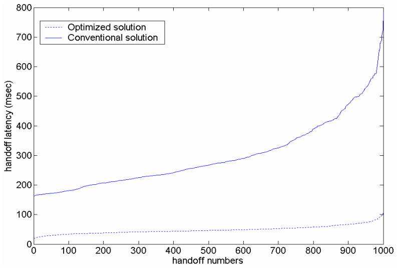

Fig. 7. Global Handover Latency (heavy load)

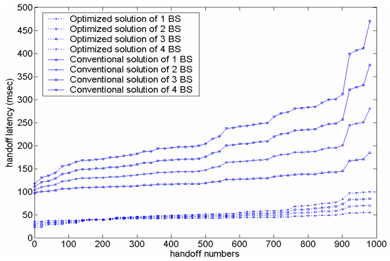

Fig. 8. Valid BS number vs. HO Latency (modest load)

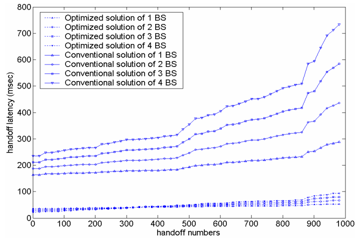

Fig. 9. Valid BS number vs. HO Latency (heavy load)

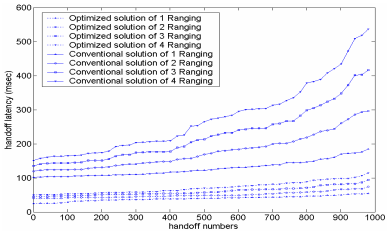

Fig. 10. Ranging retails vs. HO Latency (modest load)

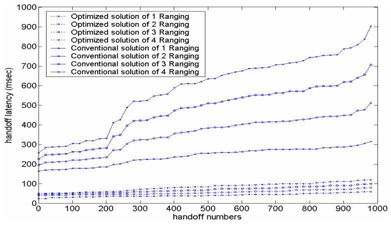

Fig. 11. Ranging retails vs. HO Latency (heavy load)

VI. Conclusions

In a conventional mobile WiMAX HO, a MS has to spend hundreds of milliseconds for the purpose of scanning, ranging, parameter negotiation, and information exchanging. And in this period, the MS leaves its previous channel thus temporally suspending running services. This phenomenon results in impairing the service level of time-sensitive applications.

Our cross layer solution speeds up the Layer 2 HO significantly by reducing the ranging latency and eliminating the network re-entry latency through cross-layer tunneling.

References

[1] Sik Choi, Gyung-Ho Hwang, Taesoo Kwon, Ae-Ri Lim, and Dong-Ho Cho, "Fast Handover Scheme for Real-Time Downlink Services in IEEE 802.16e BWA System", *Vehicular Technology Conference (2005 IEEE 61st),* June 2005, Volume 3, pp. 2028 - 2032

[2] Doo Hwan Lee, Kyandoghere Kyamakya and Jean Paul Umondi, "Fast Handover Algorithm for IEEE 802.16e Broadband Wireless Access System", *Wireless Pervasive Computing, 2006 1st International Symposium on*, 16-18 Jan. 2006, pp. 1- 6

[3] Sebastian Speicher and Christian Bünnig, "Fast MAC-Layer Scanning in IEEE 802.11 Fixed Relay Radio Access Networks", *ICN/ICONS/MCL*, 2006. 23-29 April 2006

[4] IEEE, "IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed Broadband Wireless Access Systems", *IEEE Standard 802.11d*, 2005

[5] J. Bienaimé, "IMT-2000 versus Fixed Wireless Access (FWA) systems", UMTS Forum, May 2005.

[6] IEEE, "IEEE Standard for Local and metropolitan area networks, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, Amendment 2:Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands", *IEEE Standard 802.16e*, approved 7 December 2005