Integrated Mobility and Resource Management for Cross-Network Resource Sharing in Heterogeneous Wireless Networks using Traffic Offload Policies

Dmitry Sivchenko[12], Veselin Rakocevic[2], Joachim Habermann[3]
[1]Deutsche Telekom Laboratories, Darmstadt, Germany
[2]City University London, London EC1V 0HB, United Kingdom
[3]University of Applied Science Mittelhessen, Friedberg, Germany

Abstract: The problem of efficient use of resources in wireless access networks becomes critical today with users expecting continuous high-speed network access. While access network capacity continues to increase, simultaneous operation of multiple wireless access networks presents an opportunity to increase the data rates available to end-users even further using intelligent cross-network resource sharing. This paper introduces a new Integrated Mobility and Resource Management (IMRM) framework for automatic execution of policies for cross-network resource sharing using traffic offload and pre-emptive resource reservation algorithms. The presented framework enables both mobile-initiated and network-initiated resource sharing policies to be executed. This paper presents the framework in detail and analyses its performance using extensive ns-2 simulations of the operation of a set of static policies based on measured signal strength, and dynamic pre-emptive network-initiated policies in a WiFi/WiMAX scenario. The detailed evaluation of the static policies clearly shows that the quality of voice applications shows large deviation, mostly due to very different levels of delay in access networks. Based on these conclusions, this paper presents a design of two new dynamic policies and shows that such policies, when efficiently implemented using the new IMRM framework can greatly improve the capacity of the network to serve voice traffic with a minimal impact on the data traffic and with a very low signalling overhead.

Keywords: Mobility management, Resource management, Handovers, VoIP, Heterogeneous Networks, Quality of Service

1. Introduction

Providing continuous increased level of quality of service for next-generation network applications introduces numerous technical challenges. One of the most challenging problems is the optimal use of multi-interface terminals, which are typically able to simultaneously support connections to more than one wireless network. Optimal selection of the network by these terminals is not only an optimization problem - it requires a comprehensive, integrated solution for the management of terminal mobility, network resources and vertical handovers. In addition to the simple increase in the bandwidth when several access networks are present, it is possible to improve the overall capacity of the network by developing intelligent ways to share and use the access capacity.

A typical architecture of a converged access network consists of a number of heterogeneous Access Nodes (AN) connected to the core network via an Edge Router (ER). In general, terminal devices equipped with multiple Network Interface Cards (NICs) are used by today's subscribers. Using such devices makes it possible to use the most suited access network in every particular location of the MN. If multiple NICs available on the mobile node can be used simultaneously, a much better quality will be provided by sharing the resources of different ANs available for the communication. The possibility for seamless handovers of MN's IP flows between available NICs can be used to increase the overall capacity in the network and to offer a better performance for subscribers by unloading selected access nodes as shown in a simple reference scenario in Figure 1.

Assuming AN_A provides a better service than AN_B, network operator can manage the network configuration so that IP packets of a prioritised application (e.g. VoIP) are transmitted using AN_A while the best-effort traffic is sent using AN_B. In this way, the intelligent management of IP flows in the network can be leveraged by network operators firstly to provide a better service for prioritised applications or for prioritised MN in order to increase attractiveness of own networks for customers. Secondly, by using additional AN in the regions with higher traffic demand network operators can increase the overall capacity of their networks in a fast way.

Re-distribution of user traffic between available access nodes can be implemented by using vertical handovers enabling handover of established IP flows belonging to different applications between different NICs of the mobile node. The following three issues must be solved for this purpose: (1) IP transport: flow based multi-homing support enabling simultaneous usage of multiple NIC on MN for both downlink (DL) and uplink (UL) IP flows; (2) Logic for vertical handovers: algorithms for taking handover decisions which are used to initiate re-configuration of the flow paths through the network for diverse applications; (3) Implementation scheme: a framework comprising a set of additional entities and functions which can be implemented in the network in order to implement and to execute different handover algorithms.

This paper provides an integrated solution for these three problems. The paper presents a new framework which we call the Integrated Mobility and Resource Management (IMRM) framework. The main goal of the framework is to support any general user-initiated network selection, and at the same time to support any network-initiated vertical handover. The handover is vertical when the network node changes the type of connectivity it communicates to the network with. For example, if a two-interface terminal with a WiFi and a WiMAX interface changes from sending packets using a WiFi interface to sending packets using a WiMAX interface, we consider that handover to be vertical. Horizontal handover is the change between communicating interfaces when these belong to the same networking standard. For example, a handover between two communicating base stations because of user mobility is considered to be a horizontal handover.

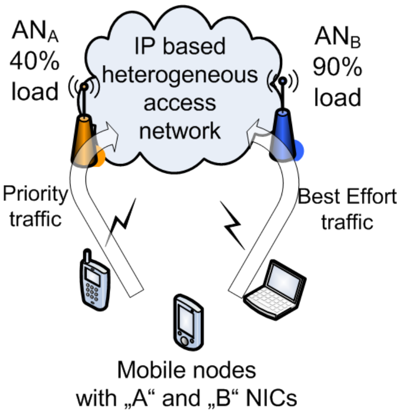
AN$_A$
40%
load
IP based heterogeneous access network
AN$_B$
90%
load
Priority traffic
Best Effort traffic
Mobile nodes with „A" and „B" NICs

Figure 1. Exploitation of heterogeneous Access Nodes to increase network capacity and to avoid overloads

Network selection in heterogeneous wireless networks can be categorized into two general approaches: network-driven and user-driven. The network-driven approach is typical for a tightly integrated environment in which a central management entity is responsible for the distribution of traffic among the access networks. In the user-driven selection the user terminals are expected to make network selection decisions, explicitly intitiated by the user. The network-initiated handover is typically defined as action taken in the network to initiate the handover based on either: (1) radio link quality (e.gl signal strength, signal-to-noise ratio), (2) introduction to the network of new network devices; or new service requirements; (3) events generated in the network (e.g. resource management, location-based optimization).

The main contribution of this research is the development of a system that can provide full support (both logical design and layer-2 implementation) for intelligent, QoS-aware network selection and vertical handovers for terminal with multiple NICs. The framework supports a new layer 2 transport mobility solution which is designed to enable usage of more than one of the network interfaces simultaneously for the same destination IP address, which is not possible using today's mechanisms of IP routing.

The IMRM framework consists of a set of functions needed to take handover decisions. As it will be explained in more detail in the paper, these functions perform flow selection, network (access node) selection and the timing of the potential handover.

There is a significant body of research work dedicated to vertical handovers and mobility management (see e.g. [1] and references within). Most often the focus is on the user-driven network selection process, and on the development of multi-attribute decision making algorithms, which use different models of cost or utility to develop optimal resource management algorithms across different networks. Examples of such analyses can be found in [2-7]. While [2] and [3] consider static environments, [4] and [5] analyse the dynamics of the networks and the dynamics of the obtained user utility, and [6] and [7] present examples of the studies of the user preference related metrics. In addition, [5] formulates the network selection algorithm as a Bayesian game, considering users to be of bounded rationality. Recent research also addresses the problem of vertical handover. In [8] the authors address a tightly coupled interworking architecture and develop a seamless and proactive vertical handoff scheme. Their scheme executes a proactive handoff (a user-driven handoff). This work is similar to ours, as the authors develop network condition detection algorithms for stations to estimate the available bandwidth and the packet delay of WiMAX and WLAN networks. They also develop algorithms to estimate the available bandwidth and packet delay in WiMAX and WLAN. However, the paper provides no architectural solutions and neglects a number of implementation issues which are described in detail in our paper. While several large EU research projects develop network architecture frameworks for vertical handovers, there is a large body of research work which formulates vertical handover decision mechanisms as optimisation problems. Paper [9] compares the performance of four such vertical handover decision algorithms, which allow different performance attributes (e.g. bandwidth, delay, packet loss, cost) to be included in the vertical handover decision process. In this and similar comparison analyses, algorithms are typically evaluated using numerical methods, with algorithms and games showing different levels of convergence and performance. While this work is essential for the development of intelligent network selection algorithms, it typically abstracts completely from the implementation details, ignoring often the origins of network dynamics and its impact on application performance. Authors of [27] address this lack of implementation details, investigating the potential 'inetgration gain' that can be achieved by intelligently sharing the resources of co-located WiMAX and WiFi access networks. Their work is similar to ours, although the provide an over-simplification of the layer-2 implementation details.

A significant amount of work has been done to standardise the network architecture frameworks for quality-based vertical handover support. One of the best known and important systems designed to enable QoS-aware handover decisions is the IEEE802.21 standard [10]. The main idea of IEEE802.21 is to provide media-independent framework and associated services to enable seamless handovers between heterogeneous access networks. IEEE802.21 does not specify any mobility management mechanisms which should be used to execute transport mobility, however. Rather, it provides tools to exchange events, information and commands in order to initiate handover execution that can be done using some handover execution protocol. On the basis of this, solutions for vertical handover management can be developed (e.g [11][12]).

One of the main problems of IEEE802.21 specification is the lack of clarity about the initiation point of the vertical handover. The lack of the required level of QoS support or low available capacity in a candidate access network may lead the network selecting entity to prevent a planned handover. The IEEE802.21 standard provides methods for continuous network performance monitoring. The standard does not, however, specify any methods for collecting the dynamic information about the network performance at the link layer. It specifies essential higher-layer mechanisms to gather all necessary information required for an affiliation with a new access point before breaking up the currently used connection. For this reason, a number of related works simply use the received signal strength

[12] as indication for an imminent handover. Therefore, the 802.21 standard specifies the use of the network-related information but does not specify which link-layer information will be collected and how. One of the main contributions of the framework presented in this paper will be to cover this missing part of the network infrastructure, as our framework specifies in detail how the quality of a network connections can be controlled.

This paper addresses some of the shortcomings of the current research work by providing the following contribution: (1) Design of the management framework for adaptive, self-organized mobility and resource management system, capable of: (a) solving the challenges of traffic distribution presented above, and (b) supporting both network-initiated and user-initiated vertical handovers; (2) Detailed simulation analysis to identify critical quality parameters which should be used for efficient network selection in WiFi/WiMAX scenario; (3) Development of custom-made dynamic network selection policy for WiFi/WiMAX scenario to justify the use of the new framework.

The rest of this paper is organised as follows. Section 2 defines the architecture and the operation of the IMRM framework in detail. Section 3 explains the operation of IMRM framework for network-initiated handovers. Section 4 introduces the observed network scenario, the simulation environment and the performance parameters used for performance evaluation. Section 5 presents the discussion of the performance of the IMRM framework. Firstly, the operation of static network selection policies are discussed. The results show interesting deviation in the VoIP performance due to significant differences in the experienced delay levels. Based on this, two new network selection policies are designed and implemented in the simulator. Simulation results show significant improvement in terms of the performance of both voice and data traffic when dynamic policies are used. The paper is concluded in Section 6.

2. Integrated Resource and Mobility Management Framework

It is important to stress that the presented solution does not concentrate on horizontal handovers which can be performed by the mobile nodes using embedded mechanisms of particular access technologies. A function executing horizontal handovers is, however, assumed in the defined framework. For the handover execution, PMIPv6 based approach is assumed to be implemented in the network. Alternatively, other solutions can be used instead of PMIPv6. For the support of multi interface mobile nodes, PMIPv6 extensions like [13] or [14] can be applied.

The framework is presented in detail in this section. The performance of particular policies is then analysed in detail in the remainder of the paper.

For the successful integration of resource and mobility management in heterogeneous networks, the following information and functionality is required: (1) location of mobile nodes (MN); (2) monitoring of the quality of service (QoS); (3) flow selection for vertical handovers; (4) vertical handover execution mechanisms. This section describes how the new IMRM framework deals with each of the four functionalities.

The architecture of the IMRM framework is presented in figure 2. The architecture consists of a number of functions which are specifically designed to perform actions required for intelligent execution of vertical handovers. The Connectivity Observation Function (COF) and the Location Tracking Function (LTF) are used to determine the location of an MN within the network. This is necessary to identify which AN may be used to communicate with an MN. The COF observes the connectivity status of particular NIC of an MN. Additionally, the COF sends to the LTF the information about the quality of the access link between the AN and the MN. This is necessary to predict possible quality of service on IP level for a particular MN. The COF informs the LTF about the link quality upon its modification. Since the AN is the first network element learning about connection of a NIC to the network and the AN knows about the link quality established with every NIC, the COF is proposed to be located in AN as shown in Figure 2. The LTF is a database containing information about connectivity status of all online NIC connected to the network. The LTF gets information from COF on all ANs in the network. Thus, it is feasible to implement the LTF in a central network entity able to communicate with every AN in the network. A signalling message ICI is introduced to transmit the required information from the COF to the LTF.

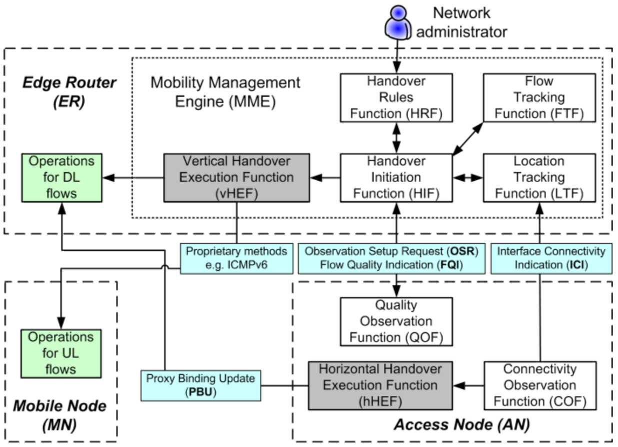
Network administrator
Edge Router (ER)
Mobility Management Engine (MME)
Handover Rules Function (HRF)
Flow Tracking Function (FTF)
Operations for DL flows
Vertical Handover Execution Function (vHEF)
Handover Initiation Function (HIF)
Location Tracking Function (LTF)
Proprietary methods e.g. ICMPv6
Observation Setup Request (OSR) Flow Quality Indication (FQI)
Interface Connectivity Indication (ICI)
Quality Observation Function (QOF)
Operations for UL flows
Proxy Binding Update (PBU)
Horizontal Handover Execution Function (hHEF)
Connectivity Observation Function (COF)
Mobile Node (MN)
Access Node (AN)

Figure 2 Architecture of the IMRM framework

For the network-controlled PMIPv6-based horizontal handovers no modifications should be performed for outgoing uplink flows handled by a mobile node. This is because the destination IP address for the outgoing packets must not be changed. Indeed, additional operations must be performed for downlink flows targeted to the mobile node because the access node used for data transfer changes. A PBU message is used for this purpose as specified in [15]. A PBU message is generated by an AN when it detects a new connected NIC. We introduce the function generating PBU messages as Horizontal Handover Execution Function (hHEF). As the COF detects connection of new NIC to AN, the COF is also used for triggering the execution of a horizontal handover using hHEF.

To select a flow for a vertical handover, the network must be able to know about all established IP flows in the network. We introduce the Flow Tracking Function (FTF) that can track all IP flows handled by MN in the network. The FTF represents a dynamic database where the information about all flows of a particular MN as well as their assignment to particular NIC and AN is stored. Having the information about every flow, the network can take handover decisions for each of them. As every flow in the considered access network architecture passes through the Edge Router (ER), it is beneficial to implement the FTF in the ER. To define the moment of time when a vertical handover must be initiated, we propose to observe the QoS level offered for prioritised applications or MN.

The Quality Observation Function (QOF) is responsible for the observation of the quality provided for different IP flows. The QOF is able to detect the reduction of the quality offered either for a prioritised network application like VoIP, IPTV or for a prioritised MN. As an AN or its RNC is only able to determine whether it can support the demanded quality for a MN, the QOF is proposed to be located in the AN.

To define the moment of time when a vertical handover must be initiated, we propose to observe the QoS level offered for prioritised applications or MN. The QOF is responsible for the observation of the quality provided for different IP flows. QOF is able to detect the reduction of the quality offered either for a prioritised network application like VoIP, IPTV or for a prioritised MN. The focus of the QoS awareness considered in this paper is on the access link between the network and the MN. As an AN or its RNC is only able to determine whether it can support the demanded quality for a MN, the QOF is proposed to be located in AN.

The Handover Initiation Function (HIF) is the central element in the IMRM framework that performs decisions about vertical handovers upon triggering messages from QOF. To initiate an observation process on the AN, OSR message is used. Using this message the HIF informs the QOF which QoS related parameter Q' must be observed on the appropriate AN as well as threshold values Q_t defining when the QOF has to inform the HIF about quality modification. A new signalling message FQI is introduced to be used for sending the observed quality Q' to the HIF. Thereby a unique identifier of an IP flow or of an application (a group of IP flows belonging to the same application) whose quality has been changed is sent to the HIF. To be able to perform handover decisions, the information about the flows handled in the network must firstly be known to the HIF. For this purpose the HIF communicates with the FTF. Furthermore, to be able to select the target AN for the handover, the HIF has to know, to which AN every flow in the network may be handed off. The HIF communicates with the LTF for this purpose. A handover algorithm defines how exactly a flow and a new AN must be selected for the handover. It also defines the type of the QoS parameter for the observation Q' as well as the threshold values Q_t. The Handover Rules Function (HRF) is introduced to store different handover rules. It is important to stress here that there is no limitation in terms of the handover policy that can be applied, i.e. the function can handle any general rule, and the framework can therefore execute any general rule for vertical handover and/or network selection. Different handover rules can then be selected or modified by the network administrator. To avoid additional overhead for the information exchange between HIF, HRF, LTF and FTF, it is beneficial to install all of them in the same network element. As the FTF is installed in the ER, all these functions are also implemented in the ER. The set of all functions located in the ER is called MME as shown in Figure 2.

Applying a handover rule the HIF can initiate a vertical handover following the receipt of a triggering message FQI from the QOF. Such handovers can be considered as **rescue** handovers. Regarding the QoS issues, they have only to be performed if the quality provided for a prioritised application of MN decreases below a defined value Q_t. Otherwise such handovers are not necessary. For a vertical handover some operations either for downlink or for uplink flow or for the flows in both directions must be performed. We introduce the function executing vertical handovers as vertical Handover Execution Function (vHEF). As the ER is the crossover point for downlink flows in the considered network architecture, handover execution for downlink flows is limited to the modification of forwarding rules for downlink IP packets in the ER. vHEF is therefore also located on ER. Since it is triggered by the HIF, the vHEF is also installed on the ER. The operations for the downlink flows performed within the ER are internal system commands. For uplink flows, currently proprietary solutions like [10] can be used, e.g. using an ICMPv6 message. Upon reception of ICMPv6 message the flow based routing table on MN will be modified and another NIC will then be used for IP packets after handover.

4. Operation of IMRM framework for network-initiated handovers

The IMRM framework presented in this paper is designed to manage both user-initiated and network-initiated handovers. To investigate the performance of the framework in more detail, in the reminder of the paper the focus will be on the network-initiated handovers. One of the reasons for this is the need to investigate a range of network selection rules. We are interested in investigating different network selection rules to understand in more detail the bottlenecks for QoS delivery for the multiservice voice-video-data traffic which is typical for the next generation wireless networks. With this in mind, in the remainder of the paper we will focus on two main types of handover rules which can be stored inside the Handover Rules Function (HRF). As described in section 2, these are the rules that can initiate rescue handovers upon triggers from the QOF. The two main types of handover rules can be defined as:

- Static rules mean that only the QoS related information from the COF (e.g. RSS of an access link) is considered for the selection of the AN for MN. It means, the dynamically changing information about the QoS level on AN is not considered thereafter. Static rules are therefore used as static selection policies to choose the NIC used on MN by default for all IP flows if multiple NIC on MN are online.
- Dynamic rules mean that the dynamically changing QoS information from the QOF is considered in order to initiate rescue handovers. Dynamic handover rules have to be applied in the network to implement intelligent load balancing. Two new dynamic rules are elaborated in section 5 to improve the performance provided for VoIP and HTTP applications in the network. The dynamic rules also include the traffic offloading mechanisms, which are used to redirect traffic belonging to a particular flow (or

‘class’ – voice / video / data), in order to maximise the performance gain for the integrated network.

In the remainder of the paper we will investigate in detail the performance of static and dynamic resource management rules. We will first use the static rules to understand better the main performance parameters that need to be monitored and measured in order to develop the optimal dynamic policies. Following this, we will define the dynamic policies and the traffic offload policy to implement these and show detailed performance results.

3.1. Static Network Selection

Static network selection refers to the traditional process of network selection. In today’s IP networks and operating systems the selection of the network interface for the default communication is done based either on the interface priority or using a static configuration manually performed by the user. The type of the network interface and its capabilities are usually not considered by the operating system. Such network selection process takes no account of the dynamics of the network environment. We can identify two traditional static network selection processes: random network selection and network selection based on the received signal strength (RSS).

In the **random selection policy**, the default network interface is chosen at random. Once chosen, a single NIC on MN is used as long as the NIC stays online. If an MN has *k* online NICs the i^{th} interface will be selected to be used by all IP flows of the MN, whereby $i = uniform[1 \dots k]$.

If the network selection is performed based on the **received signal strength (RSS)**, the choice of the network interface depends on the signal quality received at individual NICs. Such concept can be efficient especially when the network interfaces support very different communication ranges. Management of vertical handovers using signal strength has been analysed in great detail in the literature (see e.g. [23,24,25]). To demonstrate the use of the RSS-based selection, in our analysis we set an RSS threshold for the choice of WiMAX network interface. Based on the operation of the PHY layer of WiMAX networks, we choose the RSS resulting in 16QAM1/2 modulation as the threshold. In other words, considering network scenario in Figure 4, if $modulation(WiMAX) \geq @MOD$, the WiMAX NIC is used by the MN and the Wi-Fi access is used in all other cases. For the simulation analysis presented in this paper, the modulation scheme 16QAM1/2 is used for $@MOD$ because approx. 50% of MNs with more efficient modulation techniques then use the WiMAX AN while the MNs with worse WiMAX channel conditions use Wi-Fi for access.

3.2. Simulation Environment

The Integrated Resource and Mobility Management framework presented in the paper so far is capable of applying generic device-initiated or network-initiated policies and rules. To illustrate the application of this concept, we will observe the performance of the VoIP, HTTP, and FTP traffic in heterogeneous network scenario presented in Figure 3. We consider VoIP traffic as QoS-critical and in this paper will present resource management policies aimed at increasing the capacity of a heterogeneous network to satisfy QoS requirements for an increased number of VoIP calls, with minimum effect on the data-based HTTP and FTP connections. This section will give detailed explanation of the network scenario, traffic modelling and simulation used to test the operation of the policies and the framework. All simulations have been performed using the ns-2 network simulator.

3.2.1. Network Scenario

The reference network scenario used in the performed simulations is shown in Figure 4. We consider a scenario where mobile terminals have a choice between a WiMAX access network and a WiFi access network. A WiMAX access node working in the Point-To-Multipoint (PMP) mode. A smaller coverage area of the Wi-Fi AN is overlapped by a bigger coverage area of the WiMAX AN providing different modulation techniques for MN depending on the quality of wireless channel.

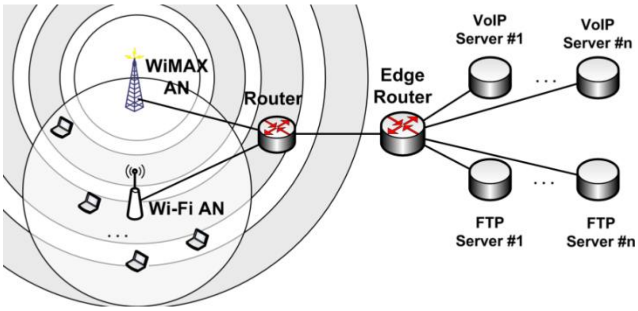

Figure 3 Reference network scenario

Table 1 Simulation Parameters of WiMAX and Wi-Fi access technologies [16]

Parameter	IEEE802.16	IEEE802.11
Physical Layer	OFDM	OFDM
Channel Bandwidth [MHz]	7	20
Operation frequency *f* [GHz]	3.5	2.4
Frame length T_{frame}	10ms	
Useful symbol time T_b	32μs	
Cyclic prefix *G*	1/16	
DL:UL subframe ratio $R_{.16}$	1:1	
Fragmentation	ON	
PLCP rate / Data rate [Mbps]		6/54
Modulation technique {64QAM3/4, 64QAM2/3, 16QAM3/2, 16QAM1/2, QPSK3/4, QPSK1/2, BPSK1/2}	**Coverage radius [m]** {352, 396, 527, 538, 602, 770, 867, 1054}	

As stated above, the main objective of the simulation analysis is to evaluate the ability of the IMRM framework and resource management policies to increase the capacity of the heterogeneous network to satisfy QoS-critical VoIP calls. In order to do this, two sets of simulations have been performed. The first set focuses on understanding the network performance parameter which is critical for the initiation of IMRM functions (primarily rescue handovers). To do this, detailed analysis of traditional, static access network selection policies has been performed. The random selection and selection based on the signal strength have been analysed to identify critical network performance parameters. These parameters will then be used to define more complex and more intelligent dynamic policies.

The main IEEE802.16 and IEEE802.11 related parameters used in performed simulations are summarised in Table 1. The number of MNs connected to the network N_{MN} is variable:

$$N_{MN} = \{16, 100, 120, 140, 160, 180, 200\}.$$

Each MN has a WiMAX and a Wi-Fi NIC with multi-homing support so that a MN can communicate using either of NIC or both. All MN are randomly distributed over the coverage area of the Wi-Fi AN so that they can use both accesses simultaneously. If MN move, additional effects of horizontal handovers (for instance, scanning for new AN) may negatively impact the network performance. Then it is more difficult to distinguish between the impact of vertical handovers initiated in a congested network and the influence of effects related to horizontal handovers. As the potential of the IMRM framework using dynamic handover rules is the goal for the following evaluations, horizontal handovers are eliminated by using static MN.

Both ANs are connected via an intermediate router to the ER. For PMIPv6-based mobility approach, the ER acts also as LMA. The ER is connected to a number of VoIP, FTP and HTTP servers. All wired links in the core part of the access network have the capacity of B = 10Gbit/s and the buffers of the ANs are set to be very large. The bottleneck of the whole network is therefore the capacity of access links of ANs.

The performance provided for typical services is investigated in this paper: VoIP, HTTP and FTP. [17,18] have been used to derive parameters for different traffic types. The VoIP service is used as an example of interactive multimedia traffic with high demands on packet delay. VoIP traffic has the highest priority and dynamic handover rules are designed to keep the VoIP quality at the satisfactory level for the most users in the network. We use G.711 codec without silence suppression to emulate the voice quality typical for ISDN networks. HTTP based web browsing is characterised by short sessions while a web page is being downloaded to the user initiating sessions by HTTP requests. FTP based file downloading is another TCP based application that typically has a lower priority in the network. FTP is the best effort traffic in our evaluations.

3.2.2. Performance metrics

For the evaluation of the VoIP performance, the well-known Mean Opinion Score (MOS) parameter [19] has been used. The work of Cole and Rosenbluth in [19] presents a very good summary of [21,22]. Therein they specify how MOS can be estimated using measurable network performance parameters which are packet delay and packet error rate. Hence, we calculate MOS for each particular VoIP flow using the method described in [20]. The average MOS of N VoIP flows $\overline{MOS}$ and the standard deviation between the qualities of these flows MOS_{std} are also calculated. The standard deviation of MOS is evaluated to assess the fairness of the VoIP service in the network. A high MOS standard deviation indicates that the VoIP quality is highly varying between different VoIP flows. Then even if $\overline{MOS}$ is high, some VoIP flows in the network may be served with a bad quality. Thereby we consider the value MOS as the lowest quality provided for a VoIP flow:

$$MOS^* = \overline{MOS} - MOS_{std} \qquad (1)$$

Additionally, to evaluate the VoIP quality for real Internet traffic, the methodology described in [18] is used. [22] defines that the VoIP quality in the network is acceptable if at least 98% of VoIP users are served with a satisfactory quality. The MOS quality corresponding to the *Best* ($MOS \geq 4.34$) or *High* ($MOS \geq 4.03$) ranking is considered as satisfactory in our evaluations. It means a user is considered unsatisfied if the VoIP quality is provided to it with $MOS < 4.03$ as defined in [20].

TCP goodput is the useful bitrate produced by TCP data packets accepted by the TCP logic only. Duplicated TCP packets are not considered by the goodput. TCP goodput is calculated as $T_{FTP} = data/\tau$ where $data$ is the number of useful bytes extracted from all TCP data packets received within the time interval τ. The average TCP goodput of N FTP flows in the network is used to assess the average quality of the network FTP service.

The average time needed to load an HTTP page $\overline{t_{HTTP}}$ is the main parameter impacting the HTTP performance experienced by network users. After the user requests a particular HTTP page the user wants to get its content as fast as possible.

To compare the performance of any policy a to any policy b, we use relation factors. We denote relation factors $\gamma^{a\rightarrow b}$ and define them for different services as follows:

$$\gamma_{VoIP}^{a\rightarrow b} = \frac{MOS^{*a}}{MOS^{*b}}; \gamma_{FTP}^{a\rightarrow b} = \frac{\overline{T_{FTP}^{a}}}{\overline{T_{FTP}^{b}}}; \gamma_{HTTP}^{a\rightarrow b} = \left[\frac{\overline{t_{HTTP}^{a}}}{\overline{t_{HTTP}^{b}}}\right]^{-1} \qquad (2)$$

The relation factor for the capacity can be defined as:

$$\gamma_{C_k}^{a\rightarrow b} = \frac{C_k^a}{C_k^b} \qquad (3)$$

The VoIP relation factor $\gamma_{VoIP}^{a\rightarrow b}$ expresses the relation between the lowest MOS values provided to users with both policies. The higher $\gamma_{VoIP}^{a\rightarrow b}$, the better the policy *a* performs in comparison to the policy *b*. Similarly, the FTP relation factor $\gamma_{FTP}^{a\rightarrow b}$shows the factor by which the average FTP goodput provided with policy *b* can be improved by using policy *a*. The HTTP relation factor $\gamma_{HTTP}^{a\rightarrow b}$ shows the factor by which the HTTP opening time is reduced when policy *a* is used instead of policy *b*.

Network capacity C_k for the traffic type *k* is used to define how many flows of the traffic type *k* can be supported in the network with a given quality. For example, C_{VoIP} defines how many VoIP calls may be established in the network with the satisfactory quality $\overline{MOS} \geq 4.03$. The capacity relation factor $\gamma_{C_k}^{a\rightarrow b}$ defines the factor by which policy *a* supports more flows of the type *k* than policy *b*.

3.2.3. Performance Analysis of Static Policies

In order to define new network selection policies, the parameter Q' for QoS observation as well as the threshold values Q_t must be defined. Thereafter the QOF can be implemented in AN and a strategy for the selection of IP flows for load balancing can be defined. For this purpose we firstly investigate the VoIP performance using typical static network selection policies.

In the analysis presented in this paper, we will focus on analysing the performance of static and dynamic traffic offload policies, applied within the IMRM framework. We will not compare the performance of the network deploying the IMRM framework with any other solutions (e.g. [8]), because the integrated frameworks have different implementation solutions, which makes any cross-framework comparison very subjective. The analysis presented here shows the performance benefits that are possible to gain when a defined policy is implemented, with traffic given strict priorities and traffic offload policies strictly applied.

For the performance analysis of the static network policies, a mix of VoIP and CBR flows in the network has been used to evaluate the performance of the defined static policies. The goal of this traffic mix is to investigate how the quality of the prioritised VoIP traffic reduces with slowly increasing background traffic leading to the overload of the entire access network. One VoIP call is assigned to every of $N_{MN} = 16\ MN$ every 5 simulation seconds. At the simulation time $t_{start} = 40s$ two CBR flows with random rates $\left(L_{CBR_i} = uniform[64,128,256,384,512]\ kbit/s\right)$, one in the downlink and another one in the uplink direction, are assigned to two random MNs every $t_{int} = 5s$.

To assess the expected advantage of the RSS based selection policy, Figures 4 and 5 show the influence of the offered AN load L on the processed load M'. The results for the WiMAX AN are predictable due to the fixed separation of the WiMAX transport resources - OFDM symbols - available for the downlink and for the uplink transmission. The WiMAX Access Node stays in the non-overloaded state as long as it has sufficient OFDM symbols to serve the whole incoming load L. An interesting observation is that $M'_{ULWiMAX}$ does not stay constant when the WiMAX AN becomes overloaded, but it slightly decreases as we can see from Figure 4. This is due to the particularities of the MAC layer of the WiMAX access technology. One synchronisation symbol is always used during data transmission from an MN. Consequently, the more MN transmit in uplink, the more symbols are required for the overhead due to the synchronisation and the fewer OFDM symbols are available for the useful data transmission. The available throughput in the uplink direction then is reduced. Obviously, with an increasing load generated by every MN more MN transmit in uplink and more symbols are used for the synchronisation purpose. Additional OFDM symbols are used in uplink as contention slots for ranging and bandwidth requests.

It can clearly be seen in Figure 4 that a higher load is needed to congest the WiMAX AN using the RSS based selection policy, i.e. $M'^{RSS}_{WiMAX} > M'^{r}_{WiMAX}$. This is due to the increased average throughput in the WiMAX cell as a result of the *@MOD* usage.

The results for the Wi-Fi AN shown in Fig. 6 show clearly the differences between the coordinated TDMA based access in WiMAX and the uncoordinated CSMA/CA access in Wi-Fi. When the Wi-Fi AN is in the non-overloaded state (i.e. when $M'_{WiFi} = L^{i}_{WiFi}$), the Wi-Fi AN can serve the whole downlink load and each MN gets sufficient time to transmit its packets in uplink. This is because of the random access to the resources of the Wi-Fi AN with the same priority for the AN and for the MN. However, when the saturation point is achieved, the amount of time used for downlink and uplink transmissions is re-distributed due to the random access to the channel. As AN and MN have the same priority for channel access so that every MN still gets sufficient time to transmit its packets in uplink while the time for DL packets becomes insufficient on AN. That is why $M'_{DL.WiFi} < L_{DL.WiFi}$ in Figure 6 while $M'_{UL.WiFi} = L_{UL.WiFi}$ even with a higher incoming load.

Since both AN become overloaded in downlink with increasing L_{DL} while the WiMAX AN becomes overloaded also in uplink with increasing L_{UL}, it is expected that the VoIP quality in downlink is smaller for all flows in the network. Figures 6 and 7 show $\overline{MOS}$ and MOS_{std} depending on the overall incoming load into the network ($L = L_{WiMAX} + L_{WiFi}$). In general, the VoIP quality decreases with the increasing incoming load causing the overload of both AN. As the RSS based selection policy can support a higher incoming load in the network while AN are not overloaded, the VoIP quality using the RSS based selection policy can be kept above the defined threshold also with a higher incoming load. To understand what the exact causes for the reduced VoIP quality are, the network performance parameters are analysed.

The measured transfer delay t_d (that is the sum of the queuing and transmission delays) of VoIP packets on both ANs is shown in Figures 8 and 9. Thresholds for the transfer delay $t_{dBest} = 0.137s$ and $t_{dHigh} = 0.2103s$ corresponding to the Best and High VoIP rankings are also shown in these figures. There are no packet losses in the used network scenario as result of very large buffers on ANs and high capacity backbone links. Consequently, the degradation of the VoIP quality is impacted by packet delay only as follows from [20].

The increasing incoming load leads to the increase of the transfer delay on both access nodes. The increasing transfer delay causes the decrease in the VoIP quality. The transfer delay for the VoIP packets over the WiMAX AN with a higher load is lower in both directions using the RSS based selection policy. The same effect can be observed for the transfer delay over the Wi-Fi AN in the downlink direction. This is due to the fact that a lower load must be processed by the Wi-Fi AN because fewer MNs use this AN using the RSS based selection policy. The uplink transfer delay over the Wi-Fi AN stays very low and is approximately the same using both policies. This is because the MNs always get sufficient time on air to transmit their packets as discussed above

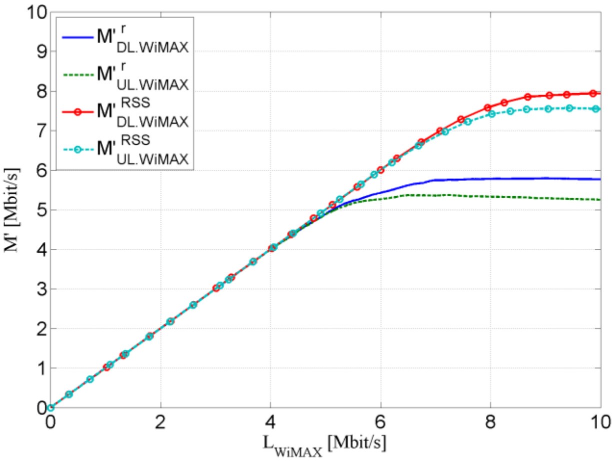

Figure 4 Effective service rate in Access Nodes when static selection policies are used WiMAX AN

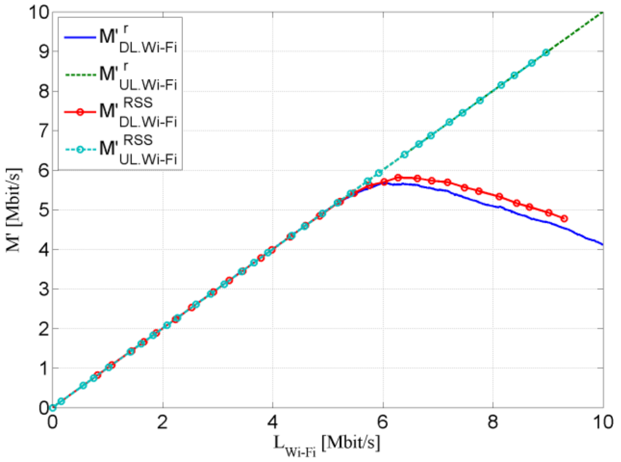

Figure 5 Effective service rate in Access Nodes when static selection policies are used WiFi AN

The analysis of MOS_{std} in Figure 6 explains two further important issues in reference to the VoIP quality in a heterogeneous network environment. Based on the results shown in Figure 6, it can be concluded that the transfer delay over the WiMAX AN increases over the threshold for the maximal network delay corresponding to the High VoIP quality ($t_d = 0.2013s$) with a lower load than the transfer delay over the Wi-Fi AN. Consequently, there are two groups of VoIP flows in the network at this time. The quality of the downlink VoIP flows using the WiMAX AN is lower than the quality provided by the Wi-Fi AN. The difference in the qualities of these VoIP groups causes an increasing standard deviation of MOS. With a further increase of the transfer delay for both AN the VoIP quality reduces to 1, i.e. $\overline{MOS}_{WiMAX} = \overline{MOS}_{WiFi} = 1$. MOS_{std} over all VoIP flows in the network is then reduced. The explanation for the increased and constant MOS standard deviation in uplink is the same. The quality provided for uplink VoIP flows via the WiMAX AN becomes very bad with a higher load while uplink VoIP flows using the Wi-Fi AN experience a very low transfer delay so that their quality is much better.

Another important fact becomes clear analysing results for the RSS-based selection policy in Figure 8. The transfer delay over both ANs increases over $t_d = 0.2103s$ with approximately equal incoming load. However, it can be seen from the result on Figure 7 that the MOS standard deviation using the RSS based selection policy is higher than in the case of the random selection policy. This is due to the properties of the IEEE 802.16 MAC. When the WiMAX AN approaches its overload, MNs using inefficient modulation techniques get insufficient OFDM symbols for the transmission of their packets within a single MAC frame. However, all packets of MN using efficient modulations can still be transmitted within a single frame. Consequently, the packets of the MN using, e.g. BPSK modulation are transmitted using multiple MAC frames using the fragmentation option of the IEEE 802.16 MAC. As the frame length used in simulations is $T_{frame} = 10ms$, the transfer delay for such MN increases significantly and the VoIP performance provided for such MN is much lower than the performance experienced by MN with efficient modulation techniques.

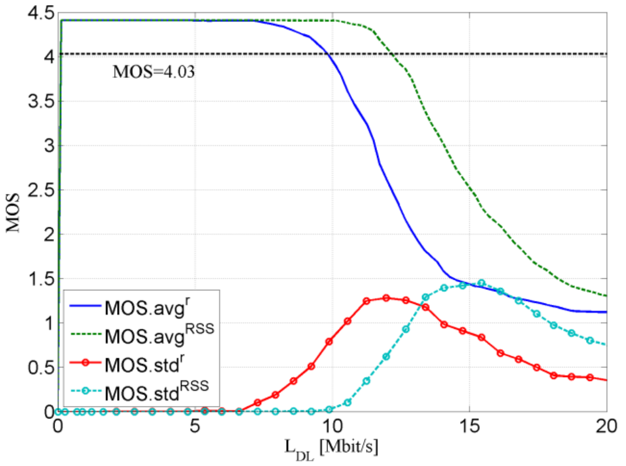

Figure 6 VoIP quality when static selection policies are used, VoIP quality downlink

Using (2) for results in Figures 6-9, the average VoIP relation factors can be calculated as $\overline{\gamma_{VoIP_{DL}}^{RSS\to r}} = 1.344$ and $\overline{\gamma_{VoIP_{UL}}^{RSS\to r}} = 1.18$. This means that the RSS-based selection policy in average improves the VoIP quality compared to the random selection policy for 34:4% in downlink and for 18% in uplink. However, the improvement of the VoIP quality available with the RSS-based selection policy does not mean that the VoIP service is provided for all MNs with a satisfactory quality MOS > 4:03; this can clearly be observed in Figures 6-9.

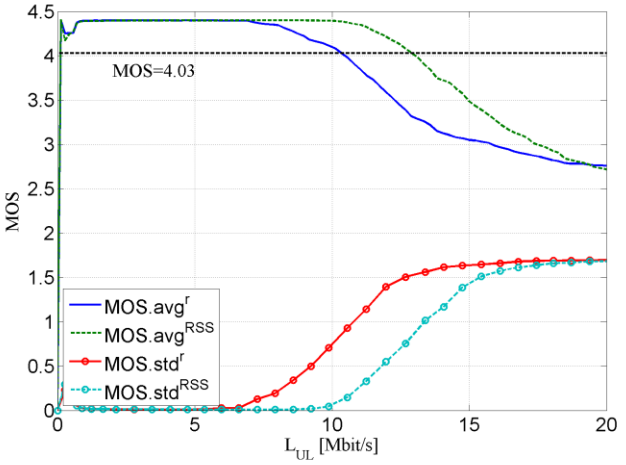

Figure 7 VoIP quality when static selection policies are used, VoIP Quality uplink

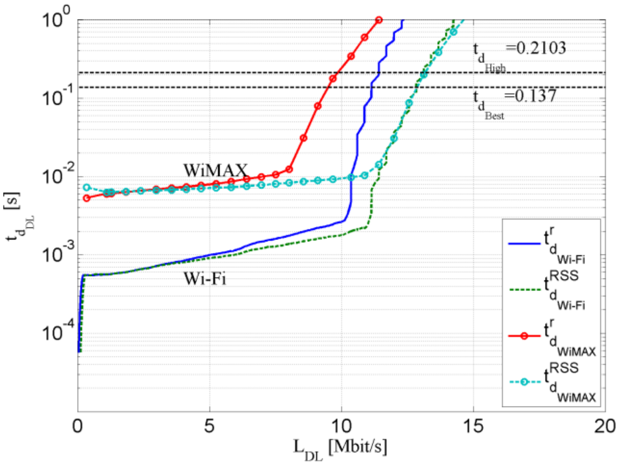

Figure 8 Analysis of delay when static selection policies are used Delay downlink

To summarise, the following three main conclusions can be made based on the evaluation of the static selection policies.

- Firstly, the exact reason of the decreasing VoIP quality has been identified. This is the increasing transfer delay caused by the incoming network load.
- Secondly, it has been defined that different groups of VoIP users with different VoIP qualities may exist in the network so that the combined *MOS** value must always be analysed. A higher MOS_{std} is due to the usage of multiple ANs which may become overloaded at different times. The transfer delay of AN can be used as an indicator for AN overload and insufficient VoIP quality.
- Finally, it has been identified that the VoIP quality provided to MN using the same AN can also highly differ depending on the QoS-related parameters of the MNs, e.g. modulation techniques. The standard deviation of the transfer delay over an AN can be used to identify such situations in the network.

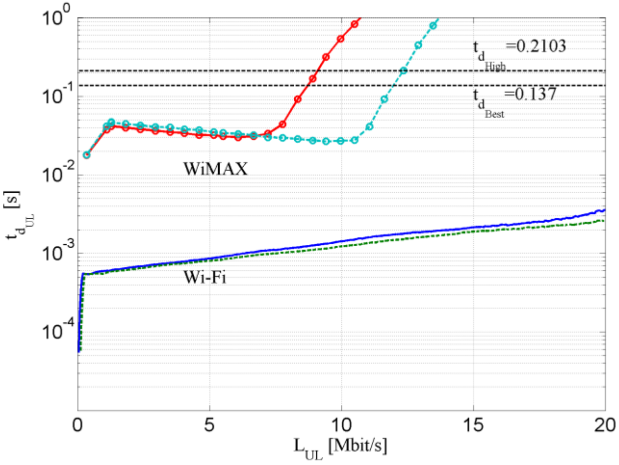

Figure 9 Analysis of delay when static selection policies are used delay uplink

3.3. Dynamic Network Selection

Detailed performance evaluation of traditional, static network selection procedures presented in the previous section can be used to design efficient network selection policies which are able to address efficiently the dynamics of the available networks. We have seen in the analysis of static network selection that the average transfer delay $\overline{t_d}$ is the main factor impacting the average VoIP quality in the considered network environment. In this section, we follow this by arguing that transfer delay should be used as the observation parameter to be monitored by the Quality Observation Function QOF of the IMRM framework to initiate rescue vertical handovers.

More precisely, we can say that the standard deviation of the transfer delay $t_{d_{std}}$ defines the variation of the quality offered for different VoIP flows served by an AN. Thereby we consider the sum $\overline{t_d} + t_{d_{std}}$ as the highest transfer delay experienced by a VoIP flow assigned to an AN. The basis of our dynamic network selection algorithms is:

If the quality provided for this flow is above the considered satisfactory threshold *MOS* = 4.03, the quality of all other VoIP flows over this AN is also acceptable. Consequently, we use the defined sum as the observation parameter for the monitoring by QOF, i.e. $Q' = \overline{t_d} + t_{d_{std}}$. This means that the QOF informs the MME in the case $\overline{t_d} + t_{d_{std}} > \xi$ where ξ is the maximal allowed transfer delay over an AN, i.e. $Q_t \equiv \xi$. As the transfer delay on AN is very dynamic, a dwell timer, or, the reaction time σ is additionally introduced for the observation. The idea of the reaction time σ is to avoid the FQI messages sent to the HIF due to accidental increases of the observed transfer delay. The observation process performed by the QOF can finally be formulated as:

$$If\ \overline{t_d} + t_{d_{std}} > \xi\ during\ t \geq \sigma\ then\ inform\ the\ MME$$

This defines the moment of time when a vertical handover can be initiated. The MME has then to define the flow that must be handed off in the case a handover decision is taken.

The interference, the changing modulation techniques of MNs, and the number of the demanding MNs connected to an AN all influence the average transfer delay of data packets transmitted by the AN. Upon reception of the FQI message on the AN as a result of the observation process, the incoming load on the indicating AN has to be reduced. Vertical handovers used for this purpose are therefore rescue handovers, initiated by the MME to keep the transfer delay over the indicating AN low.

3.3.1. Single threshold handover rule

The dynamic network selection process, then, should be controlled by monitoring the packet delay and by initiating the rescue handovers when necessary. The next two sections introduce two methods for the observation process, show how Integrated Resource and Mobility Management framework would manage network selection and present detailed simulation analysis.

The first method is based on a single threshold for packet delay. The general idea of the single threshold handover rule is depicted in Figure 10. Threshold values for the observation in downlink ξ_{DL} and in uplink ξ_{UL} are pre-defined for the observation and stored in the HRF. The QOF of each AN permanently observes Q' in each direction *r* (*r* = DL or *r* =UL). As soon as Q'_r increases over the defined threshold ξ_r for the duration $t \geq \sigma$, a FQI message is generated by the QOF to the MME.

The type of the flow selected for the handover by the MME depends on the type of the AN. If the overall network load Λ is higher than the overall service rate of the network *M*, at least one AN must be selected to be kept in the non-overloaded state to serve VoIP traffic with a satisfactory quality. The transfer delay over this AN can then be kept low. All VoIP flows must then be handed off to the non-overloaded AN in order to provide a satisfactory VoIP quality in the network. We call this the **traffic offload**. With regard to the considered network scenario, the WiMAX AN has a larger coverage area and its transport resources are managed centrally by the WiMAX AN. The WiMAX AN is thus used to serve the VoIP flows if both ANs are going to become overloaded. If the WiMAX AN becomes overloaded, firstly the flows of other traffic types will be handed off to the Wi-Fi AN. The actual definition of the priority of traffic types can be open to interpretation and can be defined within the IMRM framework. For the purpose of simulation analysis presented in this paper, we consider VoIP traffic to be of high priority and HTTP and FTP traffic to be of low priority.

3.3.2. Threshold matrix handover rule

Using the single threshold handover rule, the observed Q' on different AN may be very different. For the example of two ANs, it can happen that $Q_1' \approx \xi$ and $Q_2' \approx 0$. The VoIP flows using the first AN are then served with a worse quality of service than the VoIP flows using the second AN. To reduce the unfairness of the VoIP service offered in the network using different ANs for such cases, a threshold matrix handover rule is introduced.

The general idea of the threshold matrix handover rule is to keep Q' of all ANs used for the prioritised VoIP traffic as close to each other as possible. For this purpose a threshold matrix Ξ with a number of threshold values ξ_i is defined, $\Xi = \{\xi_1, \xi_2, \xi_3, \dots, \xi_n\}$ whereby $\xi_1 < \xi_2 < \xi_3 < \cdots < \xi_n$. Different reaction times σ_i can also be defined for each threshold value within Ξ. The goal of the MME is to keep Q' on all ANs within the same interval $[\xi_i, \xi_{i+1}]$ whereby *i = 0…n-1* and $\xi_0 = 0$. Figure 11 explains the idea of the threshold matrix handover rule for a network consisting of two ANs. The increasing load in the network, Q' on AN_1 increases over the first threshold ξ_1 earlier than on AN_2 (1 in Figure 11). To keep the increased Q_1' within the first interval $[0, \xi_1]$ (where Q_2' is in), some flows must be handed from the AN_1 off to the AN_2 having a lower Q_2'. Frequent modifications of the transfer delay for a VoIP flow may cause additional packet drops due to the de-jitter buffer functionality [20]. Consequently, a lower priority flow then is handed from the AN_1 off to the AN_2. The decreased incoming load on AN_1 then leads to the reduction of Q_1' as shown in Figure 11 after 1. The quality of the VoIP flows using AN_1 becomes similar to the VoIP quality provided for the VoIP flows using AN_2.

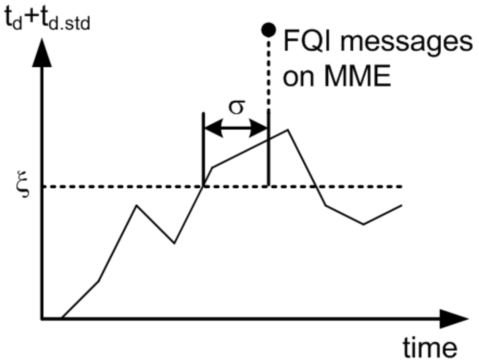

Figure 10 Idea of the single threshold handover policy

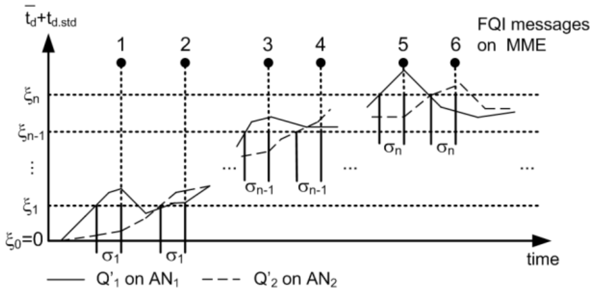

Figure 11 Idea of the threshold matrix handover rule for two ANs

When the incoming load in the network further increases, the observed Q' on both ANs overcomes ξ_1 (2 in Figure 11). Receiving FQI messages from both ANs the MME gets to know that the incoming load of the whole network is too high to keep Q' of both ANs below ξ_1. As ξ_1 is lower than the maximal allowed threshold ξ_n, the MME increases the allowed threshold for Q' to the next observation interval. Both ANs then monitor Q' that must be kept within the second observation interval $[\xi_1, \xi_2]$.

The process of such load re-distribution and the switching between different observation intervals is performed by the MME as long as the allowed threshold for Q'stays below the maximally allowed threshold ξ_n. Hence, after receiving FQI messages 3 and 4 shown in Figure 11, the MME performs the same operations as during the reception of the FQI messages 1 and 2. When the observed Q'overcomes the maximally allowed ξ_n, the whole network becomes overloaded and the same method as with the single threshold handover rule is used to unload an AN to still provide a satisfactory VoIP quality in the network. The threshold matrix handover rule is summarised in Figure 12. Using the threshold matrix rule, the MME tries to avoid handovers of VoIP flows while keeping their quality at the acceptable level.

It can be concluded that the single threshold handover rule is a simplified version of the threshold matrix rule with only one observation interval, i.e. n = 1. In this case the observed Q' value on both AN fluctuates within the interval $[0, \xi_1]$.

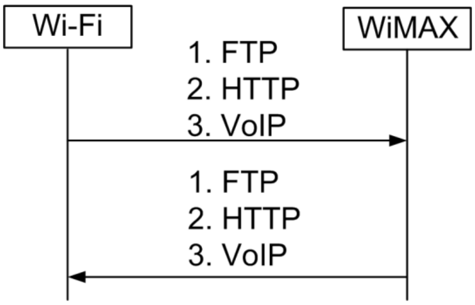

Figure 12 Selection of the flow for handover using threshold matrix rule

3.3.3. Implementation of Dynamic Network Policies

Once the logic of the network selection policies has been designed, it is critical to have a framework capable of implementing these policies. After connection of all NIC of an MN to the network the Interface Connectivity Indication ICI messages are sent to the LTF by the COF every time the quality of its access link changes. For example, in IEEE802.16 networks the actual modulation technique of a MN's NIC can be updated to the LTF upon modification. For per-flow or per-MN quality observation the HIF has to send explicit OSR messages to the QOF on appropriate AN. This must happen after detection of a new flow in the network identified by the FTF. Using per-application observation, OSR messages may be sent to AN once during the network initialisation. The MSC in Figure 3 shows an example of rescue handover initiated by the IMRM framework for the case of per-application QoS observation on AN where QOF on AN starts QoS observation right after initialisation. The modification of the used modulation technique influences the data rate that can be offered for MN. If less efficient modulation schemes are used by more MN, the average data rate provided for the MN decreases. This may result in insufficient resources for demanding MN so that the observed quality for a prioritised application will decrease below the defined threshold, i.e. $Q' < Q_t$. The QOF then informs the HIF about this event using a FQI message as shown in Figure 3 for AN$_1$. The main parameters needed to be contained within the FQI message are the identifier of the impacted application and the current value of the observed quality Q'. To know how to react to the received notification, the HIF retrieves from the HRF the appropriate handover rule. For example, the retrieved rule may define that a flow of the indicated application type assigned to the informing AN has to be rescued by handing off to any other AN available for the communication. Then the HIF retrieves from the FTF all flows of the indicated application type which are assigned to the triggering AN. To know which AN can also be used for the impacted flows the HIF retrieves connectivity information for MN maintaining these flows from the LTF. After that a handover decision can be made according to the handover rule. A rescue vertical handover for the selected flow is then initiated by the HIF. The execution is performed by means of the mobility protocol deployed in the network that is triggered by the vHEF.

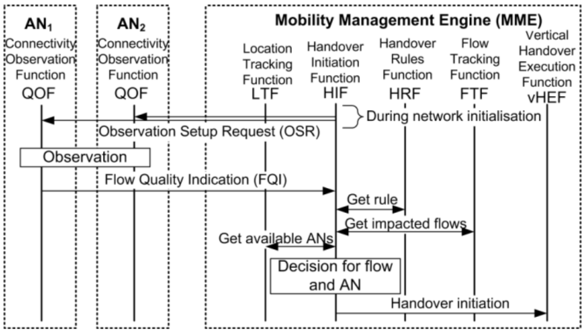
AN$_1$
Connectivity Observation Function
QOF
AN$_2$
Connectivity Observation Function
QOF
Mobility Management Engine (MME)
Location Tracking Function
LTF
Handover Initiation Function
HIF
Handover Rules Function
HRF
Flow Tracking Function
FTF
Vertical Handover Execution Function
vHEF
During network initialisation
Observation Setup Request (OSR)
Observation
Flow Quality Indication (FQI)
Get rule
Get impacted flows
Get available ANs
Decision for flow and AN
Handover initiation

Figure 13 Initiation of a rescue handover using IMRM framework

4. Performance Comaprison of network selection policies

This section presents a detailed simulation analysis of all presented policies. The thresholds for "best" and "high" quality for Q' have been defined in section 3.2.3. The analysis presented in section 3.2.3. lead us to refine the value for ξ_{Best} to $0.120s$, while $\xi_{High} = 0.210s$. The performance of the single threshold handover rule is evaluated using both of these threshold values, i.e. $\xi_1 = 0.120s$ and $\xi_2 = 0.210s$. For the threshold matrix handover rule, the matrix Ξ with n=4 steps has been defined. Thereby two additional thresholds, one lower than ξ_{Best} and another one in the mid of the interval $[\xi_{Best}; \xi_{High}]$, have been used, i.e. $\Xi = \{0.090, 0.120, 0.160, 0.210\}s$. The reaction times σ_i for all n steps are the same as for the single threshold rule. The empirically found value of the reaction time $\sigma = 0.4s$ gives a good tradeoff between the delay required to wait till a FQI message is generated by the QOF and the dynamic fluctuation of the observed Q' values.

To assess the gain achievable with the IMRM framework deploying defined handover rules, the Internet traffic performance is evaluated. The model for Internet traffic from [18][26] has been used. The methodology described in [18] is used to derive the mix of users with different traffic types. The percentage of users managing different applications is: FTP - 17%, HTTP - 33% and VoIP - 50%. The duration of the performed simulations is t = 3700s, average values of 10 simulations are presented. The overall load in the network is emulated by different number of users, N_{MN} = 100…200. The threshold matrix handover rule has been slightly adapted for the varying load in the network. When the observed Q' values over both ANs reduce below the lower threshold of the current observation interval *i* (i.e. below ξ_{i-1}), the admitted interval for Q' is decremented. The interval is decreased to the first one ([0, ξ_1]) when the observed Q' values decrease below ξ_1. The switching of the observation thresholds on the ANs is implemented using OSR messages as described in section 2.

4.1. VoIP performance

According to [18], the system capacity is defined as the number of users in the network when more than 98% of them are satisfied. To determine the number of users in the network when the VoIP quality becomes unsatisfactory, the CDF graphs have been used. Using CDF of the VoIP quality, the probability *P(MOS < 4.03)* can be determined for any N_{MN}. Figure 14 summarises the estimated probabilities *P(MOS < 4.03)* for the VoIP traffic in both directions for different N_{MN}. The simulation results presented in Figure 14 highlight the fact that the dynamic policies result in much higher probability of VoIP flows generating MOS above 4.03, thus greatly improving the Quality of Experience for the VoIP users.

The first set of simulation results is given in Table 2. This table contains the numbers of users in the network C_{VoIP} while 98% of MNs handling VoIP service are satisfied. As expected, the VoIP performance provided by the random selection policy is the worst. A network with the deployed IMRM framework using dynamic handover rules can support a higher number of MN with the satisfactory VoIP quality. The highest number of users is supported in the network using either the threshold matrix handover rule or the single threshold rule with ξ =0.210s. The VoIP capacity relation factor $\gamma_{C_{VoIP}}$ calculated using (2) is used to numerically express the performance gain enabled by the defined handover rules. Table 3 contains calculated $\gamma_{C_{VoIP}}$ for both directions. The calculated VoIP capacity relation factors indicate that the threshold matrix handover rule (as well as the single threshold rule with ξ = 0.210s) enhances the VoIP capacity of the network for 29% and 80% in comparison to particular static selection policies.

A very important fact that can be concluded from Figure 14 and from Table 2 is that the VoIP quality provided by the single threshold handover rule with ξ =0.120s is the lowest among other dynamic handover rules. The reason for that is a low threshold for Q'. Using the maximal threshold ξ = 0.120s for Q' the VoIP traffic generated by a higher number of network users (N_{MN}> 160) cannot be served by the WiMAX AN only while Q'_{WiMAX}< 0.120s. A part of VoIP flows then is assigned to the Wi-Fi AN that is overloaded when the whole network is congested. The quality provided for these VoIP flows then is very low so that *P(MOS < 4.03)^r* increases.

On the other hand, when the maximal threshold $\xi = 0.210s$ is used for Q', the whole VoIP load can be served by the WiMAX AN also for N_{MN} = 180 while Q' < ξ. Consequently, Q' of all VoIP flows can be kept below the defined threshold and the fraction of the VoIP flows served with the unsatisfactory quality *MOS* < *4.03* is below 2%.

The benefit of the threshold matrix-based policy can be identified when we observe an increasing heterogeneous load, i.e. when greater amount of FTP-based traffic is introduced to the network. Using the real traffic mix of the VoIP and the FTP traffic, the load in the network may fluctuate very much. To adapt the threshold matrix selection policy to the varying load in the network, the functionality of the QOF must slightly be extended. The QOF has to observe not only the growth of Q', but also its possible reduction. This is required to adapt the upper observation threshold to the current load in the network. For the ith observation interval Q' should be kept within the interval $[\xi_i, \xi_{i+1}]$ as defined in the paper. Additionally, the QOF informs the MME in the case $Q' < \xi_{i-1}$ for the duration $t > \sigma_{down}$. When the observed Q' on all ANs comprising the access network reduce below ξ_{i-1}, the observation interval is switched from i down to $i-1$, i.e. the MME adapts the observation interval to the decreasing network load. The lower threshold for the observation of the load reduction ξ_{i-1} is selected to guarantee that the network load decreases enough to be considered as a trend. For $i = 1$ the lower threshold is ξ_1 as Q' is always positive. The value of $\sigma_{down} = 1.0s$ is used in our simulations of the real traffic mix to ensure that the observation interval is reduced only if the load in the network trends to decrease for a longer time.

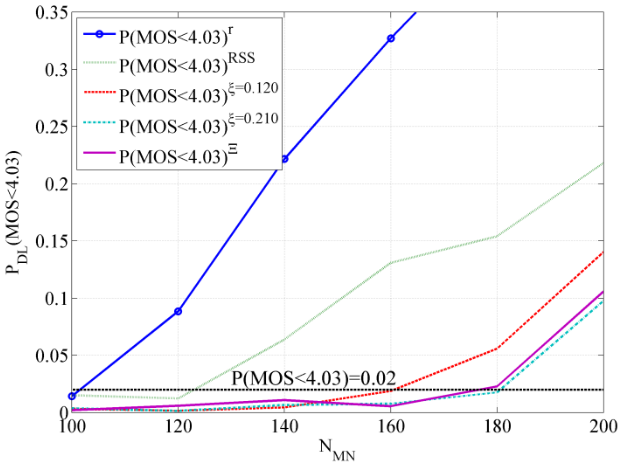

Figure 14 Estimated probabilities P(MOS<4.03) for the VoIP traffic (downlink)

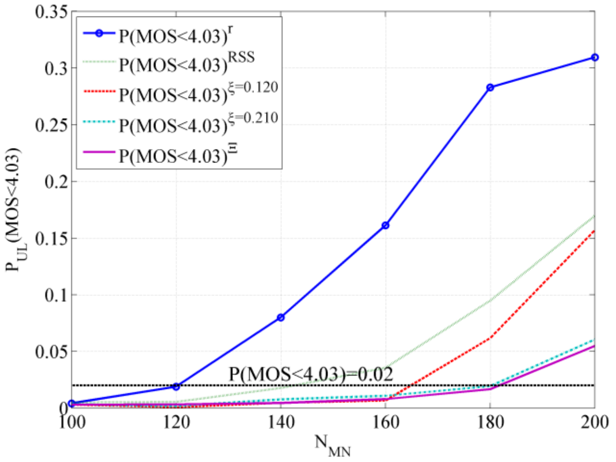

Figure 15 Estimated probabilities P(MOS<4.03) for the VoIP traffic (uplink)

Figure 16 shows average download times of HTTP pages $\overline{t_{HTTP}}$ for different N_{MN}. As expected, $\overline{t_{HTTP}}$ increases with increasing N_{MN} due to the increasing traffic in the network. The threshold matrix handover rule enables lowest average download times of HTTP pages when the network load is high ($N_{MN} \geq 180$). This is a very important fact indicating that the threshold matrix rule combines advantages from usage of both, a lower (ξ_1=0.090s) and a higher (ξ_4=0.210s) thresholds for Q'.

Table 2 Number of users in the network supported with a satisfactory VoIP quality

Policy	VoIP capacity C_{VoIP}	
	Downlink	Uplink
Random	100	120
RSS based	125	140

Single threshold $\xi = 0.120s$	160	165
Single threshold $\xi = 0.210s$	180	180
Threshold matrix Ξ	180	180

Table 3 Calculate VoIP capacity relation factor with Internet traffic

Direction	Policy *b*	
	Random	RSS-based
Downlink	1.8	1.5
Uplink	1.44	1.29

Table 4 contains average HTTP relation factors $\overline{\gamma_{HTTP}}$ calculated using (2) for the results in Figure 16. Summarising, it can be stated that the threshold matrix handover rule enables the reduction of HTTP download times by factor $\overline{\gamma_{HTTP}^{\Xi\rightarrow r}} = 4.4$ in comparison to the random selection policy and by factor $\overline{\gamma_{HTTP}^{\Xi\rightarrow RSS}} = 2.39$ in comparison to the RSS based selection policy.

Figures 17 and 18 show the dependence of the average FTP goodput in downlink and uplink directions on the number of network users. The FTP goodput in both directions is higher by using dynamic handover rules. It decreases with increasing number of MN because more users then share the available network resources. An important observation from the results in Figures 17 and 18 is the sharply decreasing FTP goodput when $N_{MN} \geq 140$ using the threshold matrix rule. This is due to the adaptive threshold value for Q'. Since the VoIP traffic is the prioritised traffic type in the network, the threshold matrix rule aims to keep Q' for the VoIP traffic within the actual observation interval. To achieve this, FTP load firstly is redistributed between both AN. This leads to frequent handovers of FTP flows causing an increase of packet reordering. Packet reordering then leads to the decreasing FTP goodput. However, the performance of the prioritised VoIP service can thereby be significantly improved as can be seen from the presented evaluations of VoIP services.

Having all this in mind, we can observe the results at Figure 16 and Table 4 which show that for high network loads the average file transfer time for HTPP traffic is smaller for the threshold matrix compared to the single threshold matrix. We can see similar performance for the uplink FTP traffic in Figure 18, where the threshold matrix based policy provides much better file transfer delay for high traffic volume.

To understand the main reason for better performance of the data traffic types for the threshold-matrix-based policy we need to explain the traffic offload process used in the signal threshold policies and the matrix-based policy. The rules for the traffic offload are kept in the Handover Rules Function (HRF) in the IMRM. The traffic offload for the experiments presented in the paper is explained in Figure 12

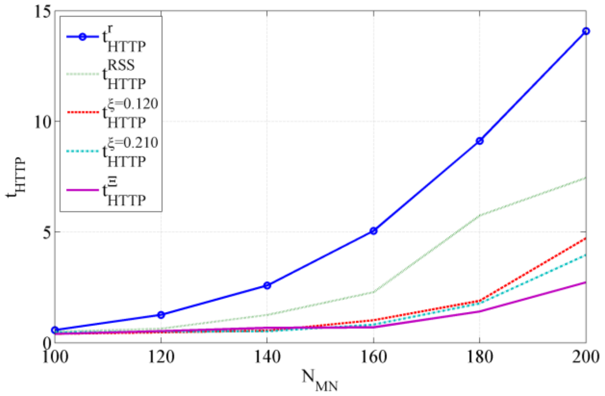

Figure 16 Average open time of HTTP pages

Table 4 Average HTTP relation factors

Policy *a*	Policy *b*	
	Random	RSS-based
Single threshold $\xi = 0.120s$	3.58	1.95
Single threshold $\xi = 0.210s$	3.91	2.10
Threshold matrix Ξ	4.40	2.39

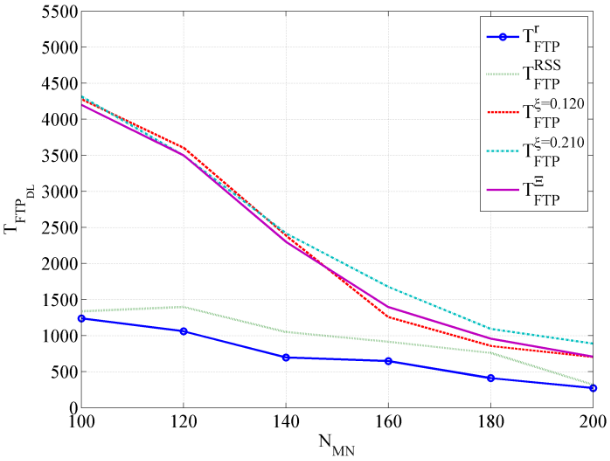

Figure 17 Average FTP goodput in downlink

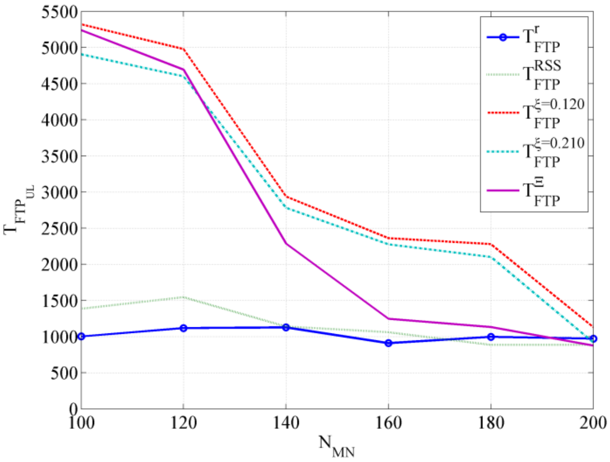

Figure 18 Average FTP goodput in uplink

Figure 20 shows the number of signalling messages in the network using the IMRM framework. The number of signalling messages using the single threshold handover rule with

$\xi = 0.120s$ is maximal for $N_{MN} > 170$. As it has already been discussed above in section 5.3, this is due to the lower threshold for Q'. When Q' cannot be kept below $\xi = 0.120s$, the QOF on both ANs generates FQI messages destined to the MME every σ seconds that explodes the signalling load in the network. The single threshold rule with $\xi = 0.210s$ causes the smallest number of the signalling messages. This is because no signalling is generated in a low to average loaded network when Q'. < 0.210s, which is the most frequent situation in the network with Internet traffic.

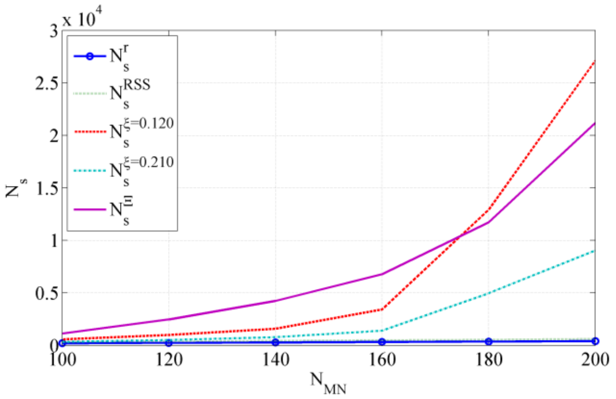

Figure 19 The number of signalling messages in the network with IMRM framework

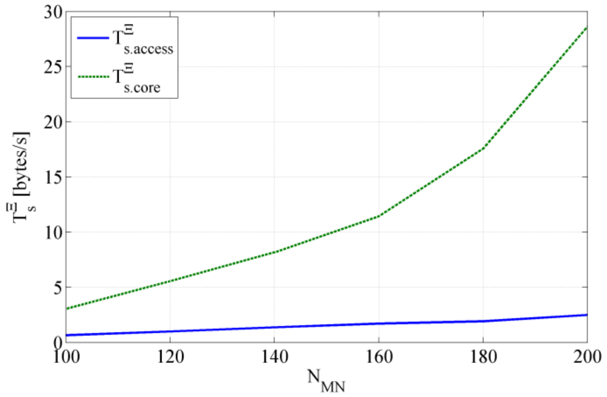

Figure 20 The signalling load per user in the network with IMRM framework

As the threshold matrix handover rule enables the highest performance for the prioritised traffic while it provides the same quality for the BE traffic as using other dynamic handover rules, the number of signalling messages caused by the threshold matrix rule N_s^{Ξ} is taken as a reference value for the calculation of the signalling load in the network. Figure 20 presents the the number of signalling messages in the network N_s^{Ξ} for each N_{MN}. However, only the signalling for handover of uplink flows (N_{UL}) is sent to MNs over capacity limited wireless links. Indeed, all N_s^{Ξ} messages are sent through the core part of the network. The signalling load per-user T_s in core and access can then be calculated as

$$T_{s.access} = \frac{f(N_{UL}) \cdot N_s^{\Xi} \cdot s_{UL}}{N_{MN} \cdot t_{stop}} T_{s.core} = \frac{N_s^{\Xi} \cdot \bar{s}}{N_{MN} \cdot t_{stop}} \qquad (5)$$

where $f(N_{UL})$ is the fraction of messages sent to the MN, s_{UL} is the size of a signalling message sent to the MN for handover execution and $\bar{s}$ is the average size of signalling messages exchanged in the core part of the network.

To retrieve numerical values for T_s^{Ξ} , an assumption has been done for the size of signalling messages. It is assumed $s_{UL} = \bar{s}$ = 1000bytes that should be sufficient to transmit the logical information contained in different signalling messages. Figure 21 presents the calculated per-user signalling load in the network using the IMRM framework. $T_{s.access}^{\Xi}$ slowly increases with the increasing number of MNs. However, $T_{s.access}^{\Xi}$ < 5bytes/s also for larger numbers of users in the network. The signalling load in the core part of the network increases more rapidly, but $T_{s.access}^{\Xi}$ < 30bytes/s also for N_{MN} = 200.

5. Conclusion

This paper presented the network performance improvements that can be achieved by implementing intelligent network selection algorithms in heterogeneous wireless network scenarios. The paper addresses the integration of resource and mobility management as an important problem in current research and development of wireless networks. The new IMRM framework is presented in detail. This framework consists of a number of functions which are able to track the location of a mobile node and its distance from the Access Points, to track the delivery of IP flows to mobile stations and to execute vertical handovers and traffic offload mechanisms to dynamically adapt to changes in the network performance. Analysis of the operation of static network selection policies is provided. The results show interesting deviation in the VoIP performance due to significant differences in the experienced delay levels. Based on this, two new network selection policies are designed and implemented in the simulator. Simulation results show significant improvement in terms of the performance of both voice and data traffic when dynamic policies are used.

References

1. Fernandes S, Karmouch A, (2012), Vertical Mobility Management Architectures in Wireless Networks: A Comprehensive Survey and Future Directions, *IEEE Communications Surveys and Tutorials*, 14(1), 45-63

2. Bari F, Leung V C M, (2007), Automated Network Selection in a Heterogeneous Wireless Network Environment", *IEEE Networks*, 21(1), 33-40

3. Shen W, Zheng Q A, (2008), Cost-function-based Network Selection Strategy in Integrated Wireless and Mobile Networks, *IEEE Transactions on Vehicular Technology*, 57(6), 3778-3788

4. Niyato D, Hossain E, Dynamics of Networks Selection in Heterogeneous Wireless Networks: An Evolutionary Game Approach (2009), *IEEE Transactions on Vehicular Technology*, 58(4), 2008- 2017

5. Zhu K, Niyato D, Wang P (2010), Networks Selection in Heterogeneous Wireless Networks: Evolution with Complete Information, *In Proc. IEEE WCNC 2010*

6. Youngkyu C and Sunghyun C, (2007), Service Charge and Energy-Aware Vertical Handoff in Integrated IEEE 802.16e/802.11 Networks, *In Proc. IEEE INFOCOM 2007, 589-597*

7. Yang C. C., Tsai C.S, Hu J.Y, and Chuang T.C, (2007), On the Design of Mobility Management Scheme for 802.16e-Based Network Environment, *Computer Networks*, 51(8), 2049-2066

8. Ma D, Ma M, (2012), A QoS Oriented Vertical Handoff Scheme for WiMAX/WLAN Overlay Networks, *IEEE Transactions on Parallel and Distributed Systems*, 23 (4), 598-606

9. Stevens-Navarro E, Wong V W S (2006), Comparison Between Vertical Handoff Decision Algorithms for Heterogeneous Wireless Networks", *In proc. IEEE VTC Spring 2006*

10. IEEE Std for Local and Metropolitan Area networks - Part 21: Media Independent Handover (2009), IEEE Std 802-21-2008, December 2009

11. Magadula L A, Chan H A, (2008) IEEE802.21 Optimized handover Delay for Proxy Mobile IPv6", *In Proc. IEEE MILCOM 2008*

12. Melia T, de la Oliva A, Vidal A, Soto I, Corujo D, and Aguiar R, (2007) Toward IP converged heterogeneous mobility: A network controlled approach, *Computer Networks*, 51(17), 4849-4866

13. Larsson C, Eriksson M, and Arvidsson P, (2009), Simultaneous Multi- Access and Flow Mobility Support for PMIPv6, *Internet-Draft, NetExtWorking Group, March 2009, draft-larsson-netext-pmipv6- smaflow-mobility-00.*

14. Bernardos C, Melia T, Seite P, and Korhonen J (2009), Multihoming extensions for Proxy Mobile IPv6, *Internet-Draft, NETEXTWorking Group, October 2009, draft-bernardos-mif-pmip-01*.

15. Gundavelli S, Leung K, Devarapalli V, Chowdhury K, and Patil B (2008), Proxy Mobile IPv6, *IETF RFC 5213, August 2008, http://www.ietf.org/rfc/rfc5213.txt*.

16. IEEE Std for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems (2004), IEEE Std 802.16-2004 (Revision of IEEE Std 802.16-2001), 2004.

17. Staehle D, Leibnitz K, and Tran-Gia P (2000), Source Traffic Modelling of Wireless Applications, Lehrstuhl fur Informatik III, Universit at W urzburg, Tech. Rep. 261, June 2000.

18. Next Generation Mobile Networks: Radio Access Performance Evaluation Methodology (2007), NGMN Alliance, White paper, June 20th 2007, version 1.2.

19. Mean Opinion Score (MOS) terminology, ITU-T, Tech. Rep., July 2006, recommendation P.800.1.

20. Cole R and Rosenbluth J (2001), Voice over IP Performance Monitoring, *ACM SIGCOMM Computer Communication Review*, 31(2)

21. The E-Model, a computational model for use in transmission planning (1998), ITU-T, Tech. Rep., December 1998, recommendation G.107.

22. General Characteristics of General Telephone Connections and Telephone Circuits - Transmission Impairments (1996), ITU-T, Tech. Rep., February 1996, recommendation G.113.

23. Yang K, Gondal I, Qiu B, and Dooley L (2007). Combined SINR Based Vertical Handoff Algorithm for Next Generation Heterogeneous Wireless Networks, *in Proc IEEE Global Telecommunications Conference, GLOBECOM 07*, 4483-4487,.

24. Ylianttila M, Pande M, Makela J, and Mahonen P, Optimization scheme for mobile users performing vertical handoffs between IEEE 802.11 and GPRS/EDGE networks (2001), *in Proc. IEEE Global Telecommunications Conference, GLOBECOM 01,* vol. 6, 3439-3443.

25 Lee S, Sriram K, Kim K, Kim Y. H, and Golmie N (2009), Vertical Handoff Decision Algorithms for Providing Optimized Performance in Heterogeneous Wireless Networks, *IEEE Transactions on Vehicular Technology*, 58(2), 865-881

26. The 3rd Generation Project Partnership 2 (3GPP2), cdma2000 Evaluation Methodology, Revision 0, Tech. Rep., December 10 2004, C.R1002-0, ver. 1.0.

27. Wang W, Liu X, Vicente J, Mohapatra P, Integration Gain of Heterogeneous WiFi/WiMAX Networks, IEEE Transactions on Mobile Computing, 10(8), August 2011