Project Number:	**CELTIC / CP7-011**
Project Title:	Mobile Networks Evolution for Individual Communications Experience – MEVICO
Document Type:	P (Public)

Document Identifier:	D2.3
Document Title:	**Final Evaluation Report**
Source Activity:	WP2
Main Editor:	Zoltán FAIGL (BME-MIK)
Authors:	Jörgen ANDERSSON (Ericsson AB), Erick BIZOUARN (ALBLF), Michael BOC (CEA), Khadija Daoud (Orange), Çağatay EDEMEN (AVEA), Salih ERGÜT (Turk Telekom), Zoltán FAIGL (BME-MIK), Jean-Luc LAFRAGETTE (ALBLF), Conny LARSSON (Ericsson AB), Rashmi PURUSHOTHAMA (Ericsson AB), Ahmet Serdar TAN (Turk Telekom), Engin ZEYDAN (AVEA), Laszlo BOKOR (BME-MIK)
Status / Version:	0.1
Date Last changes:	24.08.2012
File Name:	D2.3 Final Evaluation Report.doc

Abstract:

This document describes the validation results of WP2. The document describes for each validation topic the objective of the validation, the selected test scenarios, the applied tools and test environment, the expected results, the results and the involved partners.

Keywords:

WP2 validations, test scenarios, validation tools and environments, expected results, results

Document History:	
30.11.2011	Document with draft TOC created
01.06.2012	BME-MIK added the results part in Section 2.6.1 (HIP DEX-AKA)
05.06.2012	BME-MIK added the results part in Section 2.6.2 (suitability analysis)
18.06.2012	France Telecom added the UFA-SIP validation results
19.06.2012	NSN Finland added the results for DMA with GTP
25.06.2012	CEA added results on PMIP-RO
25.06.2012	France Telecom updated validation resultsfor UFA-SIP
28.06.2012	Ericsson AB added results for SCTP-based mobility and Wi-Fi offloading
03.07.2012	CEA added results for PMIP-NEMO
09.07.2012	TT/AVEA added results for Section 2.7 on mobile relaying
10.07.2012	BME-MIK added conclusions section and made some editorial work
14.08.2012	Partners included changes based on internal reviewer comments Final editorial work has been done.

Table of Contents

Authors

Partner	Name	Phone / Fax / e-mail
BME-MIK	Zoltán FAIGL	
		Phone: +36 70 943 9862
		e-mail: zfaigl@mik.bme.hu
BME-MIK	Laszlo BOKOR	
		Phone: +3614633420
		e-mail: bokorl@hit.bme.hu
CEA	Michael BOC	
		Phone: +33 16 908 3976
		e-mail: michael.boc@cea.fr
ALBLF	Jean-Luc LAFRAGETTE	
	Phone:	+33 13 077 2738
	e-mail:	jean-luc.lafragette@alcatel-lucent.com
ALBLF	Erick BIZOUARN	
		Phone: +33 13 077 2724
		e-mail: erick.bizouarn@alcatel-lucent.com
AVEA	Engin ZEYDAN	
		Phone: +90 216 987 6386
		e-mail: engin.zeydan@avea.com.tr
AVEA	Çağatay EDEMEN	
		Phone: +90 216 987 6386
		e-mail: cagatay.edemen@avea.com.tr
TURK TELEKOM	Salih ERGÜT	
		Phone: +90 212 309 9976
		e-mail: salih.ergut@turktelekom.com.tr
TURK TELEKOM	Ahmet Serdar TAN	
	Phone:	+90 212 309 9975
	e-mail:	ahmetserdar.tan@turktelekom.com.tr
Ericsson AB	Rashmi PURUSHOTHAMA	
	Phone:	+46 10 715 5964
	e-mail:	rashmi.purushothama@ericsson.com
Ericsson AB	Jörgen ANDERSSON	
	Phone:	+46 10 719 7013
	e-mail:	jorgen.andersson@ericsson.com

Ericsson AB	Conny LARSSON	
	Phone:	+46 10 714 8458
	e-mail:	conny.larsson@ericsson.com

Executive Summary

The general aim of validations is to evaluate how a particular technology satisfies the claimed challenge(s).

This document presents the validation results for WP2. For each validation there is a description of objectives, validation scenarios, validation tools, expected results, applicability of the results, results and involved partners. The sections on validation objectives collect the claimed challenges resolved by the particular technology, and the key performance indicators (KPIs). Validation scenarios describe the test cases where the technologies show their benefits, furthermore any important issues on the planned test cases. The sections on validation tools describe the validation environments, but we must note that Section 6 in deliverable D2.1 [1] contains more detailed information on a few of them. The description of the technologies can be found in deliverable D2.2 [2].

List of acronyms and abbreviations

3GPP	Third Generation Partnership Project
AAA	Authentication, Authorization and Accounting
ADSL	Asyncronous Digital Subscriber Line
A-GW	Access GW
AKA	Authentication and Key Agreement
ANDSF	Automatic Network Decision Selection Function
BEX	Base Exchange
BNG	Broadband Network Gateway
BS	Base Station
CERT	Certificate
CMAC	Cipher-based Message Authentication Code
CN	Correspondent Node
CQI	Channel Quality Indicator
CTS	Clear to Send
DeNB	Donor Evolved NodeB
DEX	Diet Exchange
DHCP	Dynamic Host Control Protocol
DoS	Denial of Service
DUID	DHCP Unique IDentifier
EAP	Extensible Authentication Protocol
EC	Elliptic Curve
E-E	End-to-end
eNB	Evolved NodeB
eNodeB	Evolved NodeB
EPC	Evolved Packet Core
ePDG	Evolved Packet Data Gateway
ePGW	Evolved Packet Data Network Gateway
GBR	Guaranteed Bit Rate
GGSN	Gateway GPRS Support Node
GTP	GPRS Tunnelling Protocol
GW	Gateway
HA	Home Agent
HeNB	Home eNodeB
HIP	Host Identity Protocol
HIT	Host Identity Tag
HO	Handover
HSDPA	High-Speed Downlink Packet Access
HSS	Home Subscriber Server
IA	Intermediate Anchor
IKEv2	Internet Key Exchange Protocol version 2
IMS	IP Multimedia Service

KPI	Key Performance Indicator
L2	Layer-2
L3	Layer-3
LFN	Local Fixed Node
L-GW	Local Gateway
LMA	Local Mobility Anchor
LRA	Localized Routing Acknowledgment
LRI	Localized Routing Initiation
LTE	Long Term Evolution
MAG	Mobility Anchor Gateway
MIIS	Media Independent Information Service
MIMO	Multiple-Input and Multiple-Output
MIPv6	Mobile IPv6
MME	Mobility Management Entity
MN	Mobile Node
MNID	Mobile Node Identifier (PMIPv6)
MPTCP	MultiPath Transport Control Protocol
MR	Mobile Router
NAT	Network Address Translation
NEMO	Network Mobility Protocol
NEMO BS	NEMO Basic Support
NMIP	Non Mobile IP
OFDMA	Orthogonal Frequency Division Multiplexing
OPEX	Operational Expenditure
P-CSCF	Proxy-Call Service Control Function
PBA	Proxy Binding Acknowledgment
PBU	Proxy Binding Update
PCEF	Policy and Charging Enforcement Function
PGW	Packet Data Network Gateway
PMIPv6	Proxy Mobile IPv6
PoA	Point of Access
POP	Point of Presence
PSK	Preshared Key
QoS	Quality of Service
RADIUS	Remote Authentication Dial In User Service
RAN	Radio Access Network
RN	Relay Node
RO	Routing Optimization
RTS	Request to Send
RTT	Round-trip time
SA	Security Association
SCTP	Stream Control Transmission Protocol

SEG	Security Gateway
SGW	Serving Gateway
SINR	Signal to Interference and Noise Ratio
TCP	Transmission Control Protocol
TLS	Transport Layer Security
UDP	User Datagram Protocol
UE	User Equipment
UFA	Ultra Flat Architecture
UFA-GW	Ultra Flat Architecture Gateway
UFA-HIP	HIP-based UFA
UFA-SIP	SIP-based UFA
UMTS	Universal Mobile Telecommunications System
UPnP	Universal Plug and Play
USIM	Universal Subscriber Identity Module

1. Introduction

Work Package 2 works on the terminal and core network function aspects of an advanced Third Generation Partnership Project (3GPP) Evolved Packet Core (EPC) architecture that should scale well with the increasing data traffic demand, growing number of connected terminals, increasing mobile internet usage foreseen for the next decade. Considering these expectations, the main challenge is to find the best alternatives for smart traffic steering, investigate multipath communication and offloading possibilities in order to balance and minimize the load in the EPC. This Work Package (WP) is also responsible to work out appropriate mobility management strategy to maintain session continuity for mobile multi-access terminals across heterogeneous access networks and multiple, overlapping IP domains advertised by distributed gateways such as Packet Data Network Gateways (PGWs), local breakout and third-party gateways. The detailed description of challenges and technologies in focus of WP2 can be found in D2.1 Advanced EPC Architecture [1]. The technologies are described in D2.2 Architectural EPC extensions for supporting heterogeneous mobility schemes [2].

This document describes the validation plans of WP2. The goal of validations is to demonstrate with simulations and/or proof-of-concept systems the feasibility and the significance of the achieved results. The document describes for each validation topic the objective of the validation, the selected test scenarios, the applied tools and test environment, the expected results, and the involved partners.

2. Validations of WP2

2.1 Offloading

This validation topic demonstrates the offloading capabilities of operator-managed Wi-Fi access networks. Wi-Fi access saves 3GPP radio resources, increases indoor coverage for UEs, and facilitates offload of broadband traffic.

2.1.1 Improving UE's multiple network access capability and load balancing through Wi-Fi offloading

2.1.1.1 Objectives

Operator managed Wi-Fi

- Facilitate off-load of broadband traffic from wide area radio network to Wi-Fi.
- Provide better indoor coverage for Wi-Fi enabled devices.
- Provide operator services tied to mobile subscription also over Wi-Fi

2.1.1.2 Validation scenarios

Figure 1 depicts the validation setup.

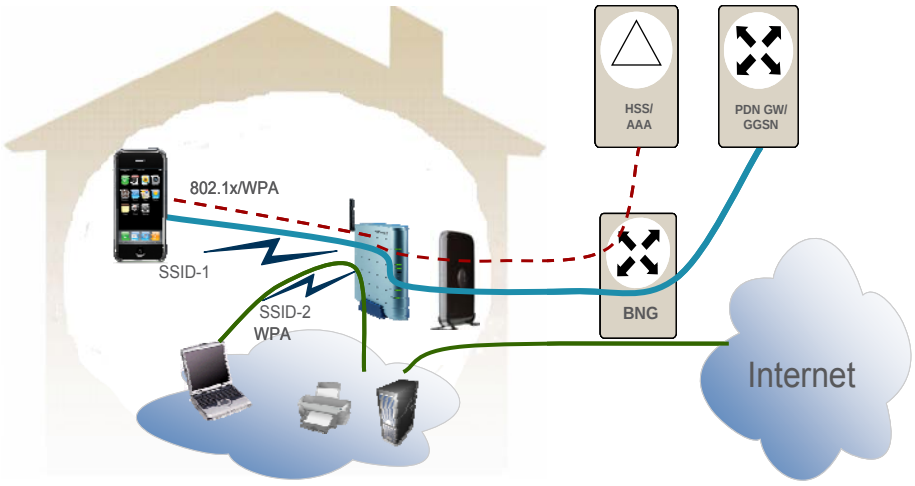


Figure 1 – Wi-Fi offload scenario.

The components used in this setup are as follows:

1. UE
   - Any device capable of EAP-SIM authentication with Wi-Fi interface.
2. Wi-Fi access point
   - Proxy function UPnP (Universal Plug and Play)
   - NAT to reach the local network
   - DHCP proxy
   - Capable to set up a tunnel towards the BNG
3. BNG
   - RADIUS proxy
   - DHCP server
   - Tunnelling protocol implemented (e.g. GTP) to set up a tunnel to GGSN/PGW

2.1.1.3 Validation tools

Since the implementation of operator managed Wi-Fi does not include any measurements, we are not using any validation tools.

2.1.1.4 Expected results

The user shall be able to use the same services regardless of access type, and the Wi-Fi access is expected to provide better indoor coverage.

2.1.1.5 Results

We have been able to demonstrate off-load of broadband traffic from wide area radio network to Wi-Fi, provide better indoor coverage for Wi-Fi enabled devices and provide operator services tied to mobile subscription also over Wi-Fi.

2.1.1.6 Applicability of the results

Facilitate off-load of broadband traffic and provide operator services tied to mobile subscription also over Wi-Fi.

2.1.1.7 Partners involved

Ericsson

2.2 Dynamic Mobility Anchoring

Dynamic Mobility Anchoring aims to switch on and off of mobility management based on the mobility of the UEs. Hence the traffic of non-moving UEs will not go through mobility anchors in case of anchor-based mobility management solutions. The goal of this section is to show the performance improvements of DMA.

2.2.1 DMA principles applied to GTP based mobility

2.2.1.1 Objectives

The purpose is to compare the effect of GW distribution when anchoring is done as it is defined in 3GPP and when dynamic mobility anchoring (DMA) is used. Comparisons are done by using simple enough models in order to get some estimates of GW changes due to mobility.

2.2.1.2 Validation scenarios

Comparison is done by creating different cases by changing n = the number of gateways as follows:

- n = the number of eNodeBs, a flat architecture,
- 1 < n < the number of eNodeBs, a distributed architecture.
- n = 1, current centralized architecture

as illustrated in Figure 2.

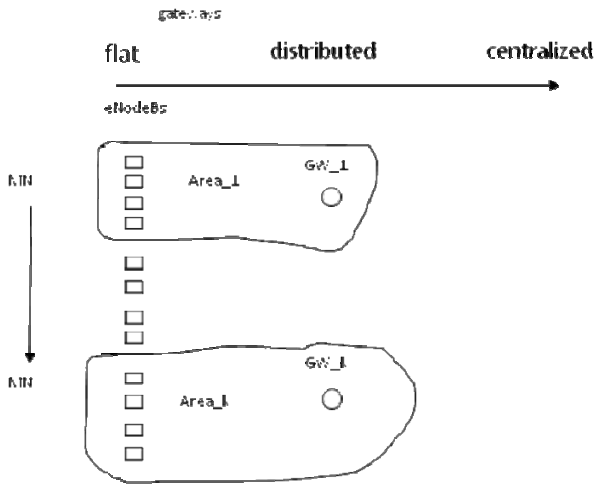


Figure 2 - Distributed vs. centralized architecture.

It is assumed that eNodeBs are located at intersection points of a grid. The "area" of a GW is defined by the coverage of the corresponding eNodeBs.

Mobility model is a simplified model: It is assumed that cell and area residence times are exponentially distributed measured in seconds (Example: If a mobile moves 4,5 km/h (=1,25m/s) and the diameter of cell area is 50m, then one estimate for average cell residence time is 40s). It is also assumed that the density of the mobiles is constant, that they are uniformly distributed and that they move with a constant velocity in a random direction maintaining their direction.

Mobility assumptions (expressed by average values of a single mobile)

- medium: 0,02 cell crossing per sec: 72 per h
- fast: 0,2 cell border crossing: 720 per h

Corresponds to Cell Residential Time

- medium: 50 sec
- fast: 5 sec

Examples for velocity depending on cell size

- V [km/h]=#HO/h x d [km] x 0,7 (random changing direction…)
- d=500m, v medium = 25km/h, v fast = 250km/h
- d=100m, v medium = 5km/h, v fast = 50km/h

Traffic is modelled as a session of several flows (Figure 3).

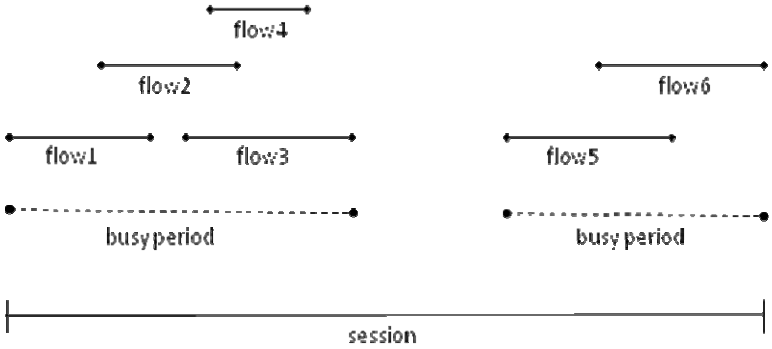


Figure 3 - Session comprising several flows.

- **Flow:** A (e.g. TCP) connection where a given amount of data is to be transferred
- Flows arrive independently of each other, they can be partly concurrent.
  - For simplicity: Flow duration is assumed to be exponentially distributed, they arrive according to a Poisson process
- A **session** consists of a **number of flows** (which is assumed to be geometrically distributed)

- In calculations, the **inputs** are: **arrival rate** of flows, the average **length of a flow** in seconds and the average **amount of flows in session (**Figure 4**).**

By using the above model, the mean number of GW passages of a flow is calculated in the following cases:

- When flows are always anchored at the first GW (3GPP case)
- When a flow is anchored at the current GW (DMA case)

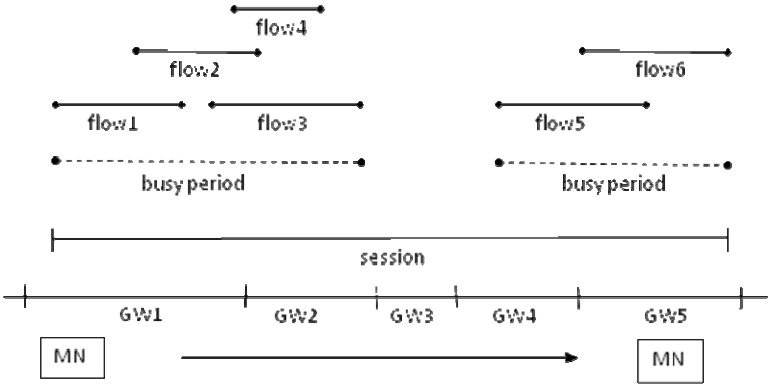


Figure 4 - Assignment between flows and GWs.

2.2.1.3 Validation tools

The estimations are based upon the above simple enough models in order to get some, even rough, formulas for estimations.

2.2.1.4 Expected results

Quantitative understanding on the benefits and drawbacks of distributing GWs.

2.2.1.5 Results

From the simple mobility model it can be concluded by using the border crossing formula (the rate of mobiles leaving an area depends on its perimeter):

”GW” HOs / all HOs = 1/sqrt(number of eNodeBs in the area)

- 100 eNodeB GW => In 10% of all HOs GW changes
- 2000 eNodeB GW => In 2,2% of all HOs GW changes
- 10000 eNodeBs GW => In 1% of all HOs GW changes

Using the traffic model the results for a low mobility scenario are illustrated in Figure 5

1) Mean cell residence time 40 sec, mean flow duration 20 sec

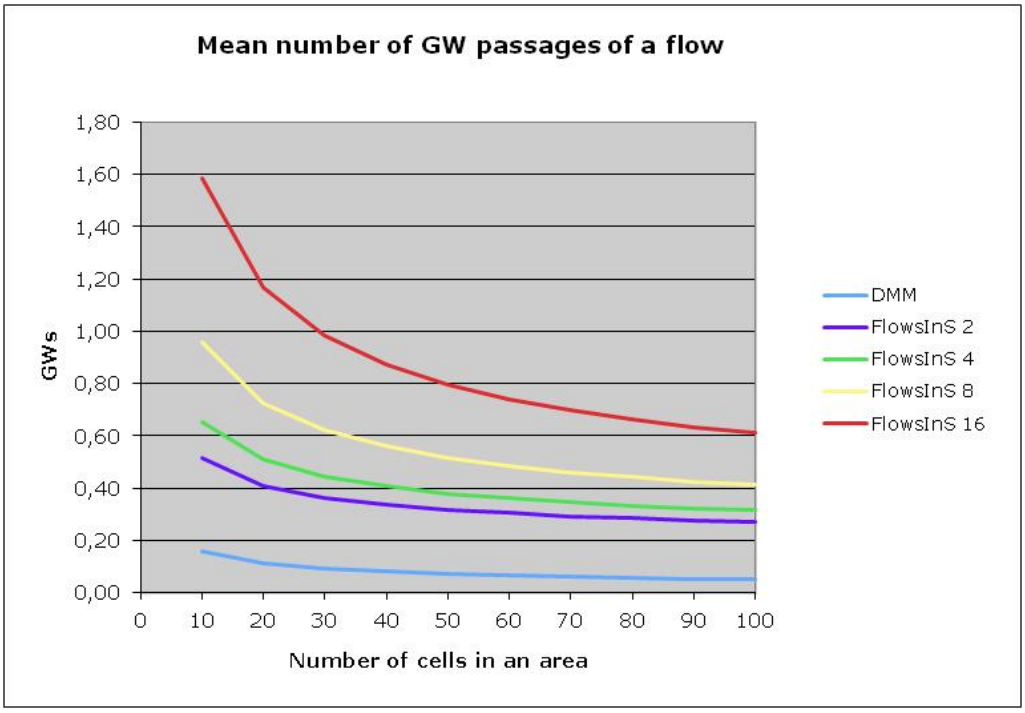


Figure 5 - Low mobility scenario.

And for a high mobility (fast moving) scenario illustrated in Figure 6:
Mean Cell residence time is 5 sec, mean flow duration is 20 sec.

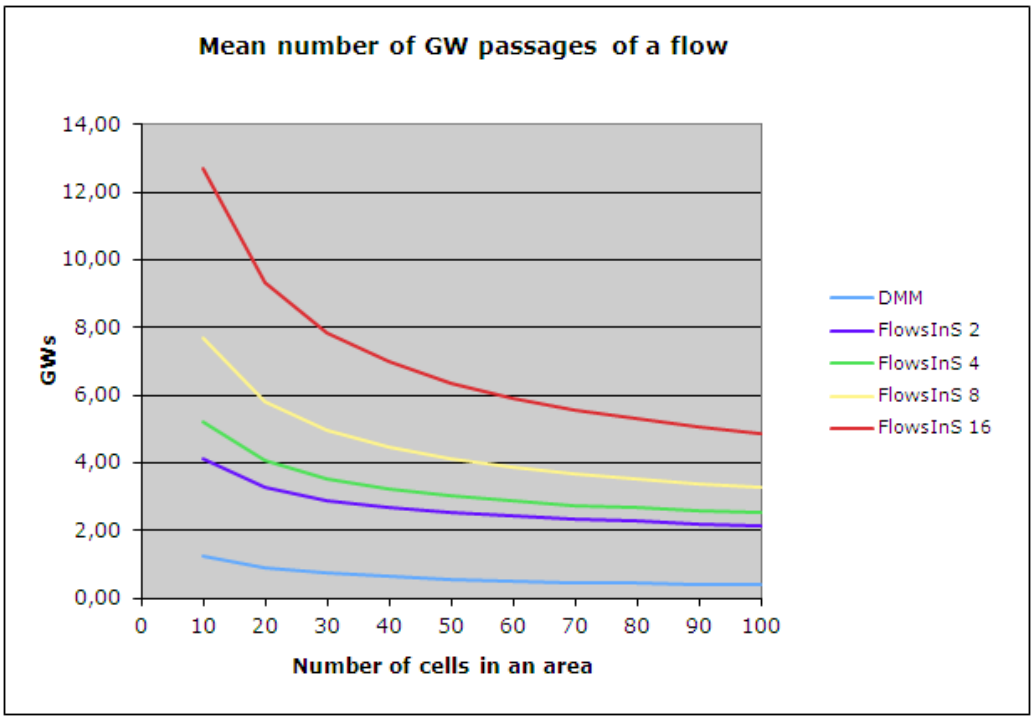


Figure 6 - High mobility scenario.

2.2.1.6 Applicability of the results

The conclusion from the simple mobility model is: At least in cases where a GW serves a "small" amount of cells, optimization of local GW change procedures should be considered.

The estimations obtained by using the traffic model is that for Dynamic Mobility Anchoring almost only one GW is used per flow whereas for the 3GPP case a fast moving UE would pass 3 to 8 GWs (depending on the number of cells per GW). This leads to the following conclusion:

Only for fast moving terminals the problem of GW changes/routing optimization need to be considered (e.g. for transport systems). For these highly mobile scenarios it is worth investigating how dynamic mobility anchoring principles can be applied to the EPC.

2.2.1.7 Partners involved

NSNF

2.3 Terminal-based mobility management

This section validates the performance and viability of terminal-based mobility protocols which do not need mobility management functions deployed in the core network. These protcols provide optimized routing and flow mobility for the transport layer, i.e., for TCP or SCTP protocols.

2.3.1 Functional and performance validation of NMIP, SCTP and MPTCP

2.3.1.1 Objectives

The main objective of this validation is to test session continuity during mobility events. Session continuity is achieved by introducing session layer for SCTP protocol. These validations can be split into two parts.

From the first part from Ericsson AB, the objectives of implementing session layer with SCTP protocol are the following:

- To have a robust handover mechanism which provides seamless connectivity across changes in the network by preserving communication
- To have a comprehensive mobility solution that addresses both change in the host's IP address and the problem of long network disconnections

From the other part, the objectives are to install, to perform performance evaluation of NMIP, SCTP and MPTCP protocols and then to validate their behaviours using a real EPC (test bed). Those terminal-based protocols are relevant with the challenge "Keep signaling under certain levels" as they avoid a centralized anchor point, such function introduced by MIPv4, PMIPv6, GTP, etc., and the challenge "Mobility adaptation to the UE/application requirements" as there will have no additional cost for non moving UEs.

Using those three protocols we will measure KPIs about HO success, E-E packet loss rate, signalling load, RTT, packet delay and jitter.

2.3.1.2 Validation scenarios

Figure 7 illustrates the evaluation setup for the first part.

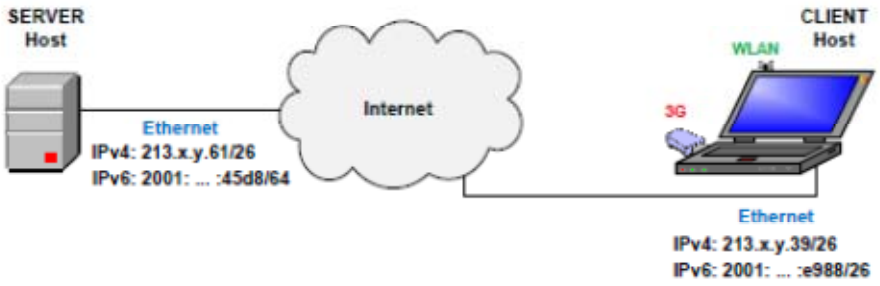


Figure 7 - Evaluation setup.

The components of the evaluation setup are as follows:

1. File transfer application
   - implemented for testing the session layer
   - uses the Session Layer API
2. Server
   - locates the file's content

- delivers file's data to the session layer
- transmitted to the other host through the session

3. Client
   - requests the file's content
   - receives file's data from the session layer

For the second part, as NMIP, MPTCP, SCTP are end-to-end, each of them is compatible whatever the selected architecture scenario.

We will measure, through network tools (iperf and wireshark), the RTT, the throughput and the E-E packet loss for each protocol.

2.3.1.3 Validation tools

In the second part of the validation an internal test bed will be used for validation. We will use what we call a LTEbox where we have installed all the EPC functionalities (ePGW, MME, PGW, SGW, PCRF, and eNodeB) We will connect UEs NMIP, SCTP, and MPTCP compliant to make all our tests.

We will use network tools for our tests (iperf, tcptrace, netperf, wireshark).

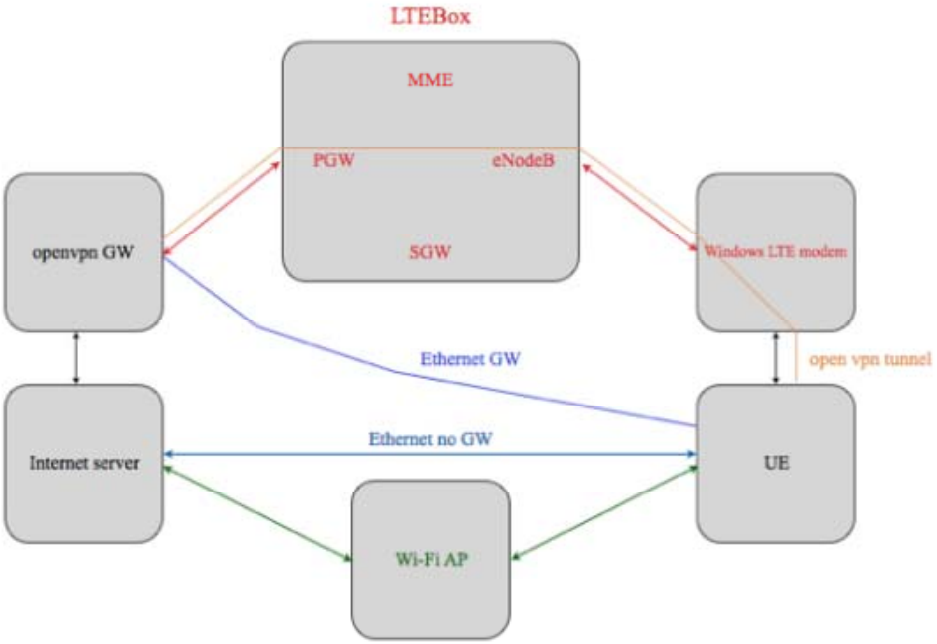


Figure 8 – Validation environment for end-to-end mobility protocols.

2.3.1.4 Expected results

For the first part, the expected experimental results are as follows:

- Session layer accepts application's request to suspend session
- Session layer gracefully detects and suspends disconnected sessions
- Session layer rapidly resumes from a suspension

Session layer integrates well with mobility-aware and mobility-unaware applications providing seamless mobility

For the second part of validations the expected results are the assessment of those protocols and their performances using EPC.

2.3.1.5 Results

SCTP:

The current implementation supports session suspension and resumption in both scenarios:

- When an application explicitly requests for a suspend
- When there is network disconnection (both long and short network disconnections)

We have tested the implementation by shifting between different available WLAN networks and also by doing a handover from Wi-Fi to 3G and vice versa. A streaming video is used as the application to test session continuity.

NMIP, SCTP, MPTCP analysis:

The testbed uses the LTEBox that implements the core network in a rack. As shown in Figure 8, the rack is composed by one eNodeB, one MME, one S-GW and one P-GW. The UE is a laptop running a Linux operating system on which is connected a LTE USB dongle for association with the eNodeB in the LTEbox. As we lack of a reliable LTE's Linux driver for the UE, we used a virtual machine running Windows OS as a modem for the validations. However, this has implied the use of a VPN tunnel between the Linux system and the OpenVPN GW (see Figure 8) crossing the Windows host which acts as a "modem". If there was a working Linux driver for the LTE dongle, the Windows host and the OpenVPN GW would be not required.

IEEE 802.11 Wi-Fi and IEEE 802.3 Ethernet 1Gb/s interfaces on the UE have been used to test the multi streaming and/or the HO capabilities. For Ethernet we have done two different kinds of connections, the first one goes through the openVPN GW, the second one goes directly to the Internet server so that one can evaluate the cost of the OpenVPN GW in this specific setup.

First each transport protocol have been tested on all types of connections (Ethernet no GW, Ethernet GW, LTE, Wi-Fi). Furthermore as these three protocols are based on data stream, the use or not of the Nagle's algorithm is tested. The aim of the Nagle's algorithm is to avoid sending small packets but waiting for more data to optimize the ratio user data size and total message size (with the headers).

For NMIP and MPTCP that are based on TCP, it is managed by the setting of the TCP_NODELAY option, referred as *nodelay* in the following figures and when the option is not set, it is referred as *delay*.

The tools iperf and its equivalent for SCTP nagle_snd/rcv have been used for this test. A typical test is a ten seconds measurements on one direction (either downlink or uplink)) for a given packet size. Because of variability of measurements, especially for the radio links, each test is repeated ten times, and the mean and the standard deviation are computed. The packet sizes in bytes that have been considered for the tests are: 10, 20, 50, 100, 200, 500, 1000, 2000, 5000 and 10000.

NMIP:

In Figure 9, the difference between the *no delay/delay* cases is clear for the Ethernet no GW curves. The limitation is due to the bandwidth of the Ethernet link (1Gbit/s). The influence of the packet size is clear; the throughput is becoming constant after the packet size reaching the MTU value (1350 bytes).

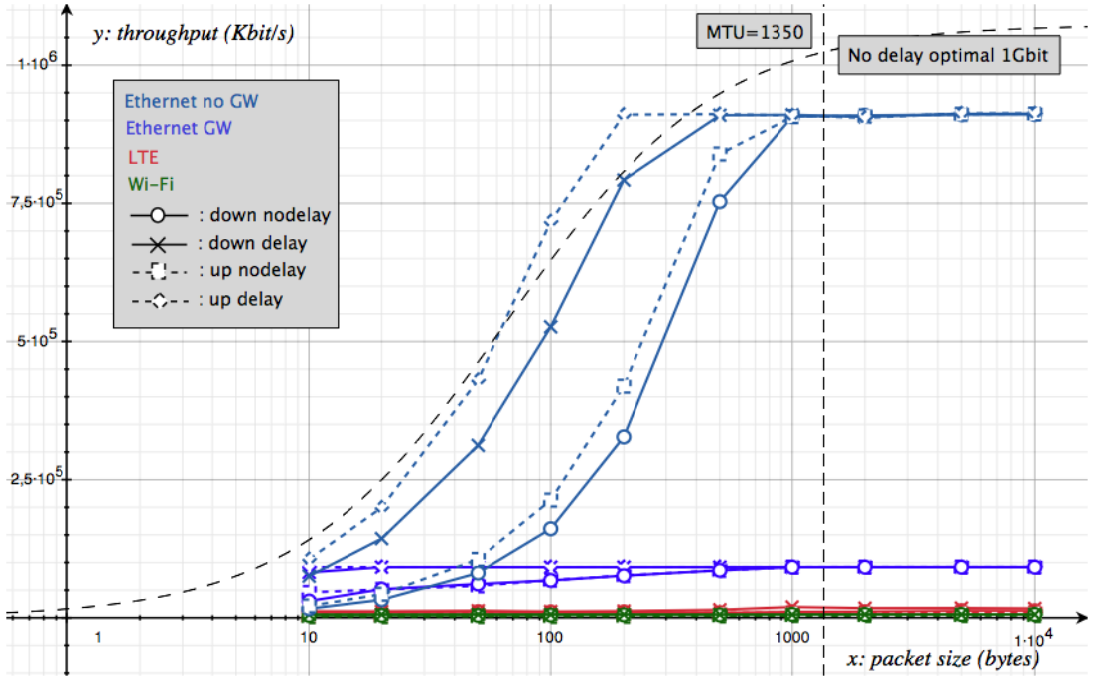


Figure 9 - NMIP Ethernet LTE Wi-Fi

However, for the *Ethernet GW*, the limitation is essentially due to the processing power of the openVPN gateway to manage the openVPN tunnel which depends on the number of packets received. The gap in performance between the delay and no delay option is equivalent in Ethernet GW and Ethernet no GW cases.

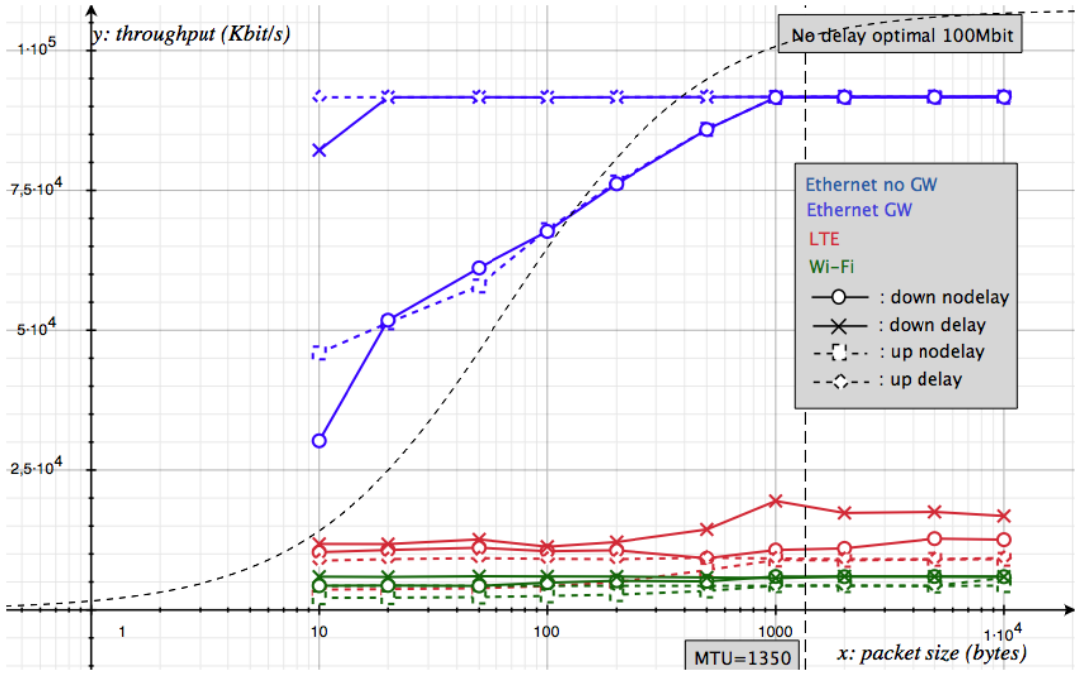


Figure 10 - NMIP Ethernet gateway

For the two radio links LTE and Wi-Fi the results are different, depending on radio characteristics like the throughput limitation for LTE of 20Mb/s on downlink and 10Mb/s on uplink. Once again the *no delay* option use decreases the performance. For Wi-Fi, our platform lab is not well adapted to realize Wi-Fi performance testing; neighboring Wi-Fi APs (more than forty at the time of measurement) generates bad radio conditions.

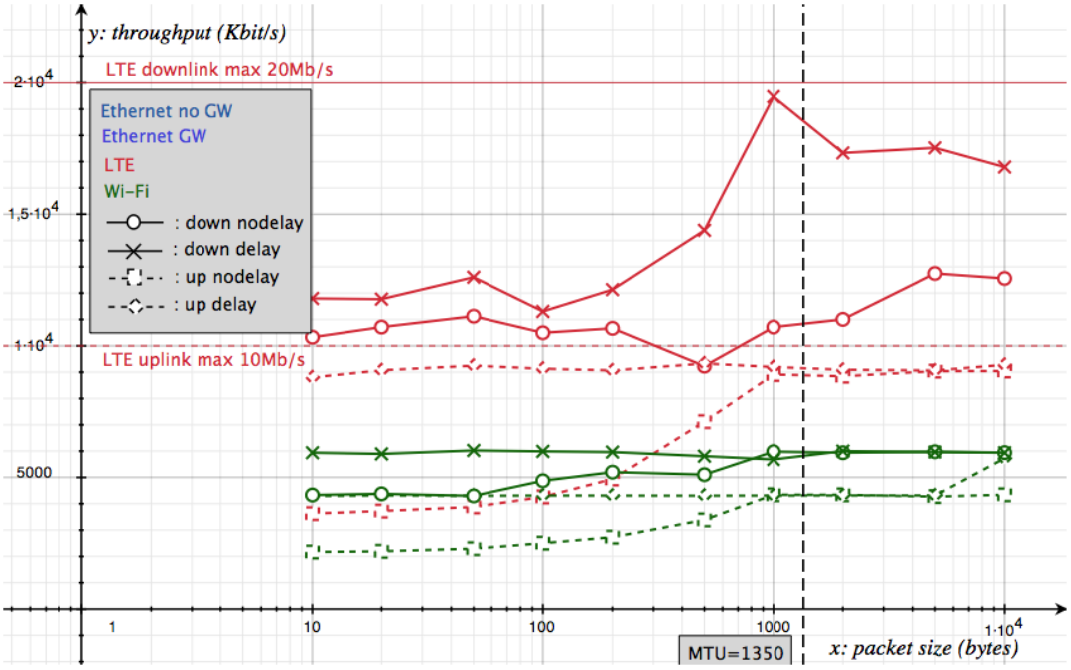


Figure 11 - NMIP LTE Wi-Fi

MPTCP:

The TCP_NODELAY option is always taken into account by MPTCP and therefore reducing the available throughput as shown in the figure below.

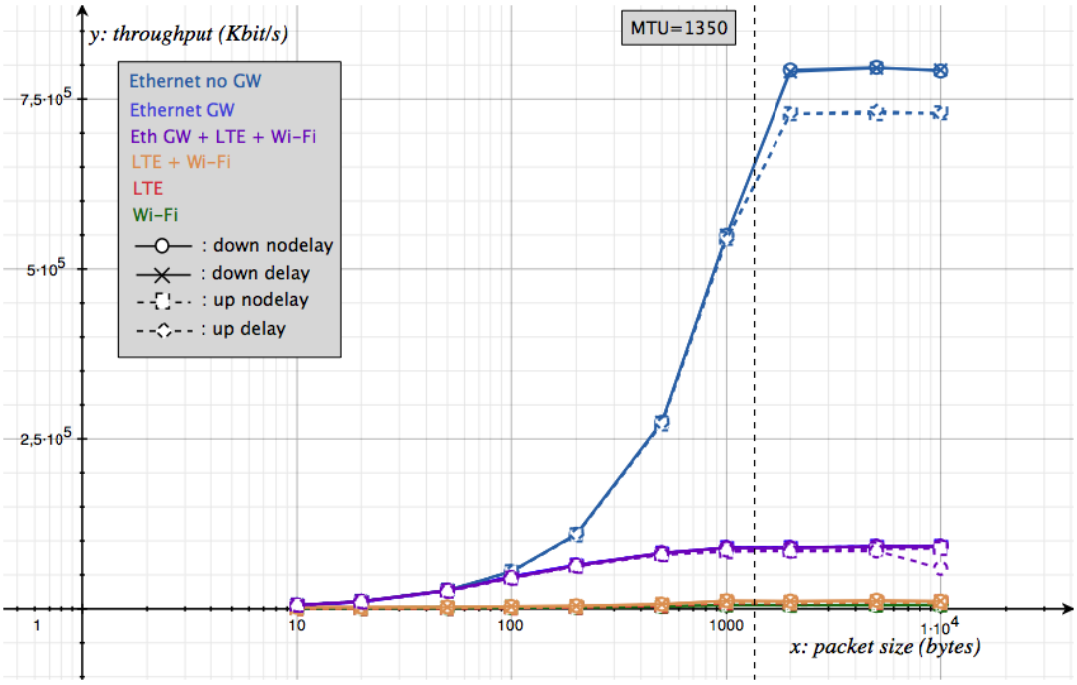


Figure 12 - MPTCP Ethernet No Gateway

Then several multipaths solutions have been tested, the first one is with three networks: Ethernet GW, LTE and Wi-Fi and the second one is realized with LTE and Wi-Fi networks. On the first test, the Ethernet link is a lot faster than the two others, the multipath (Ethernet & LTE & Wi-Fi) is not better than the best solution (Ethernet).

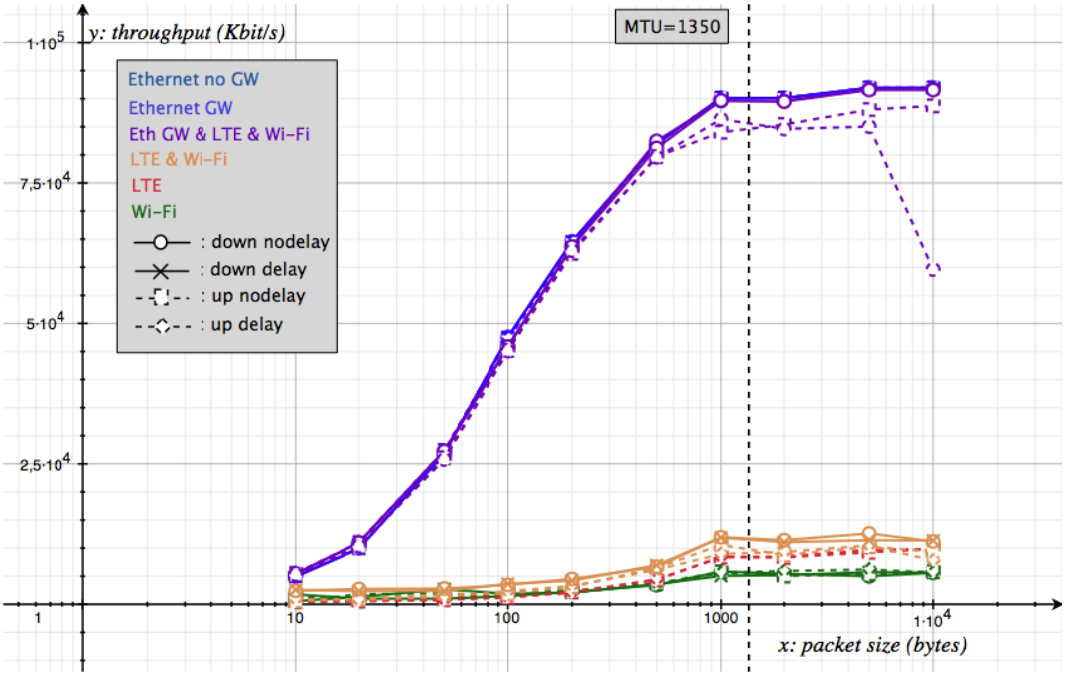


Figure 13 - MPTCP Ethernet Gateway

When the links have similar throughput like in the case of LTE and Wi-Fi simultaneous use, the multipath solution gives better result, above the values of the best interface (LTE in this case).

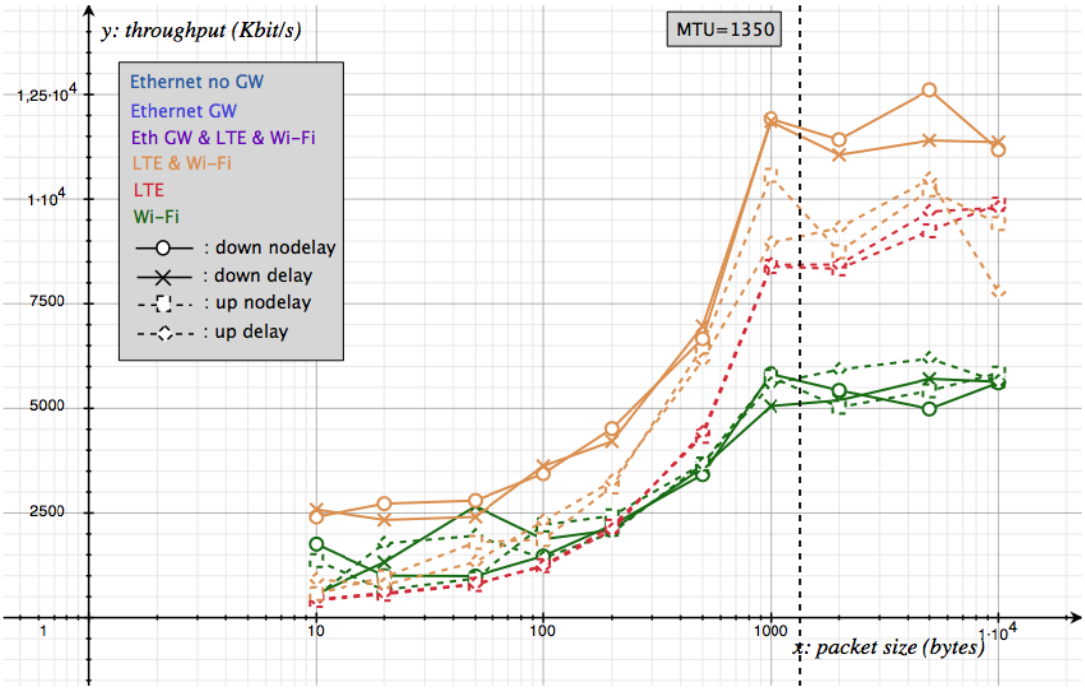


Figure 14 - MPTCP LTE Wi-Fi

As shown on the following figure the gain is not optimal. The optimal solution being represented by the sum of the LTE and Wi-Fi throughput (the black curve) in Figure 15.

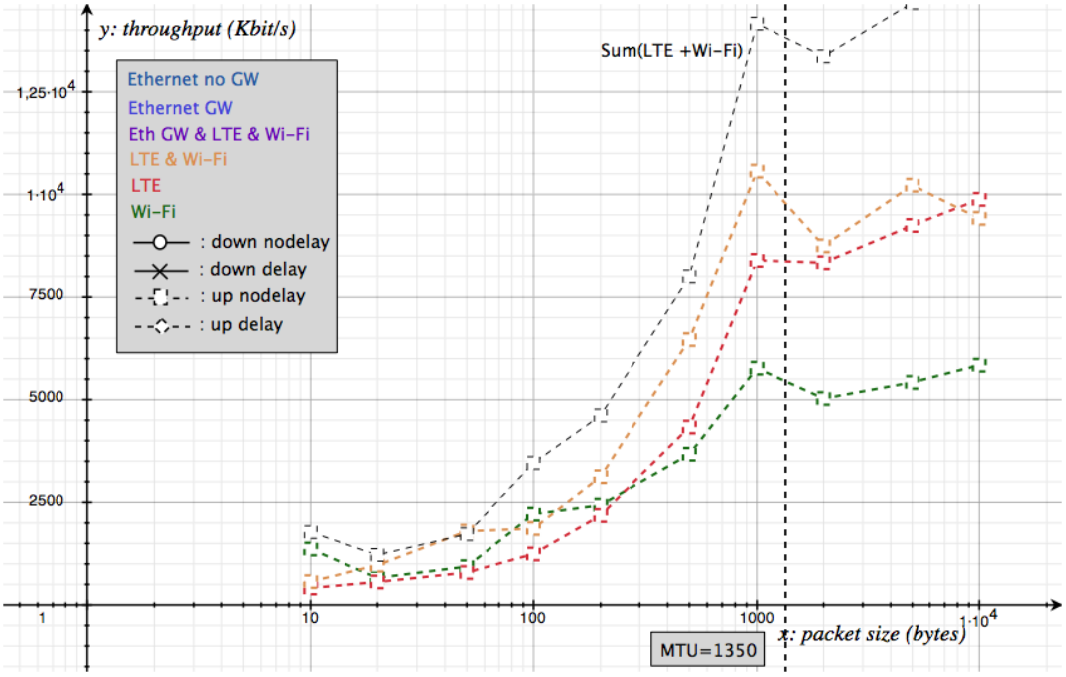


Figure 15 - MPTCP LTE + Wi-Fi

SCTP :

The effect of the TCP_NODELAY is clearly visible here for the *Ethernet no GW* curves for small packets size until reaching approximately the MTU value where the best results are for the delay option.

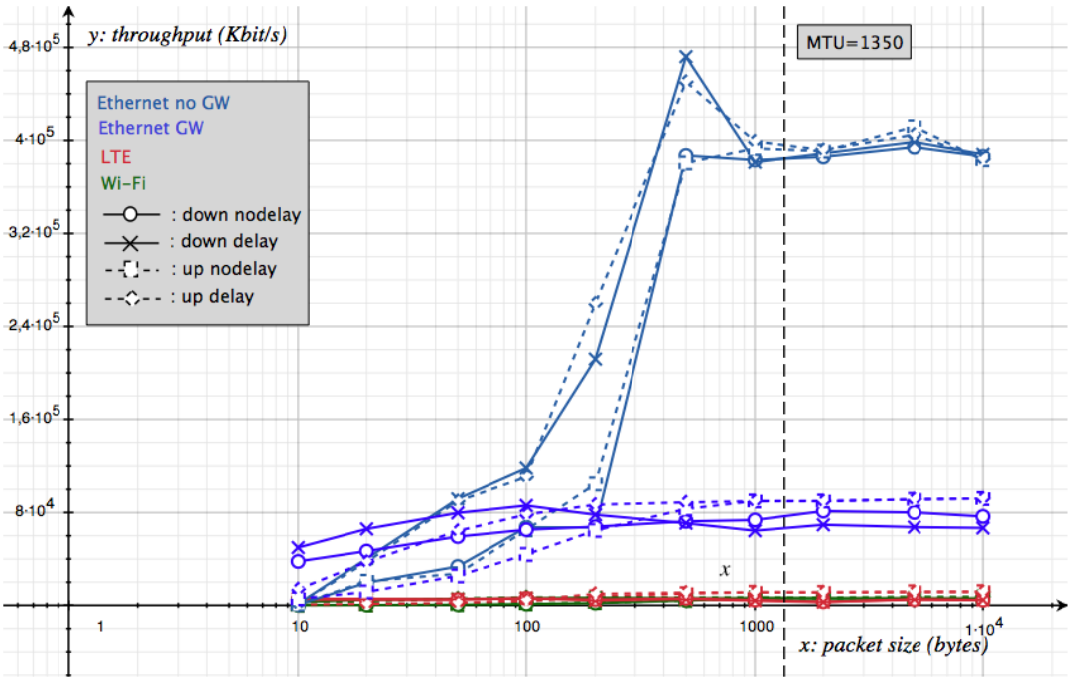


Figure 16 - SCTP Ethernet No Gateway

It is also the case for the *Ethernet GW*.

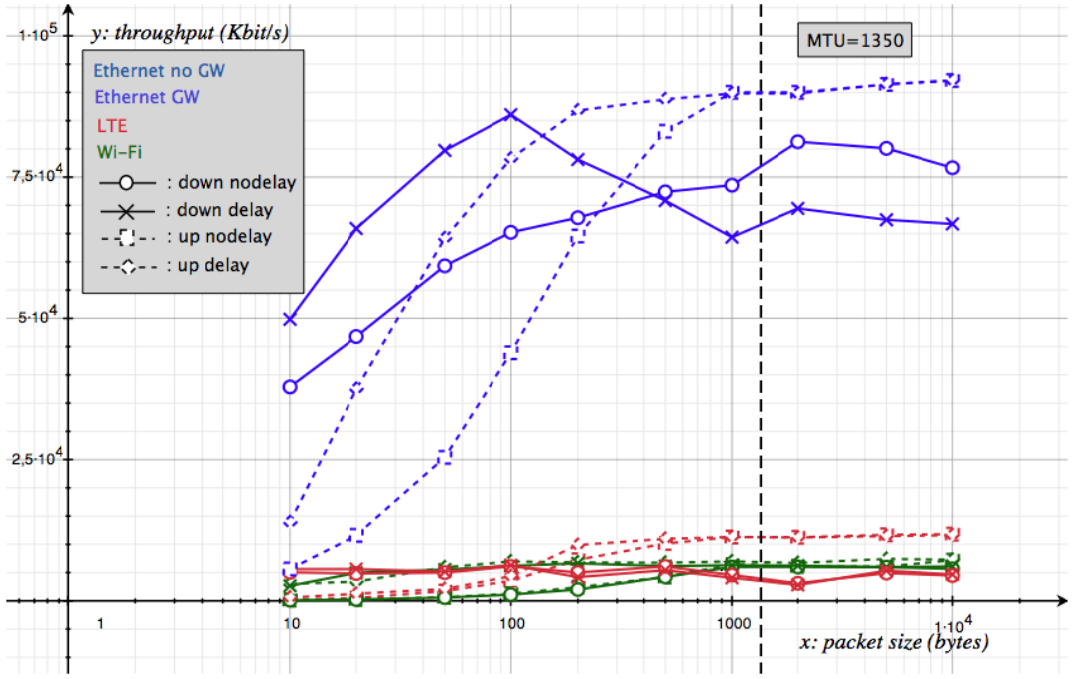


Figure 17 - SCTP Ethernet Gateway

For the LTE there is almost no difference between the *delay* curves and the *no delay* ones.

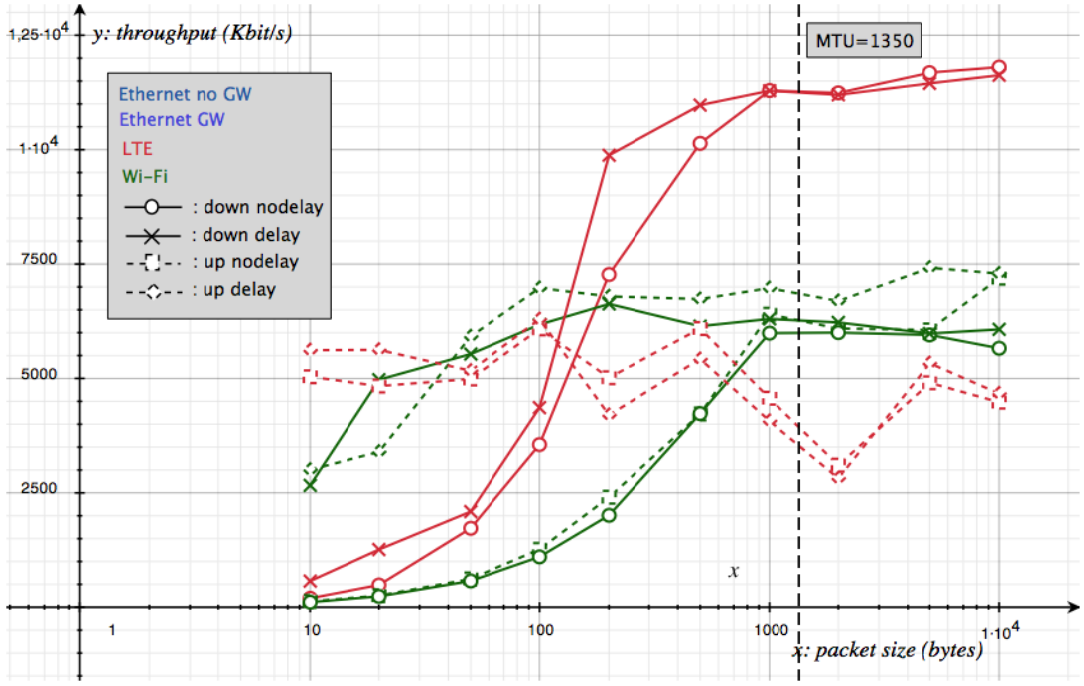


Figure 18 - SCTP LTE Wi-Fi

The results for the three transport protocols for downlink traffic without the *no delay* option are presented in the following figures.

NMIP is the most efficient protocol for the throughput criteria.

SCTP has better performance with small packets size when MPTCP, due to the fact that packets are sent when data are ready like with the TCP_NODELAY option, is more efficient on packets with bigger size.

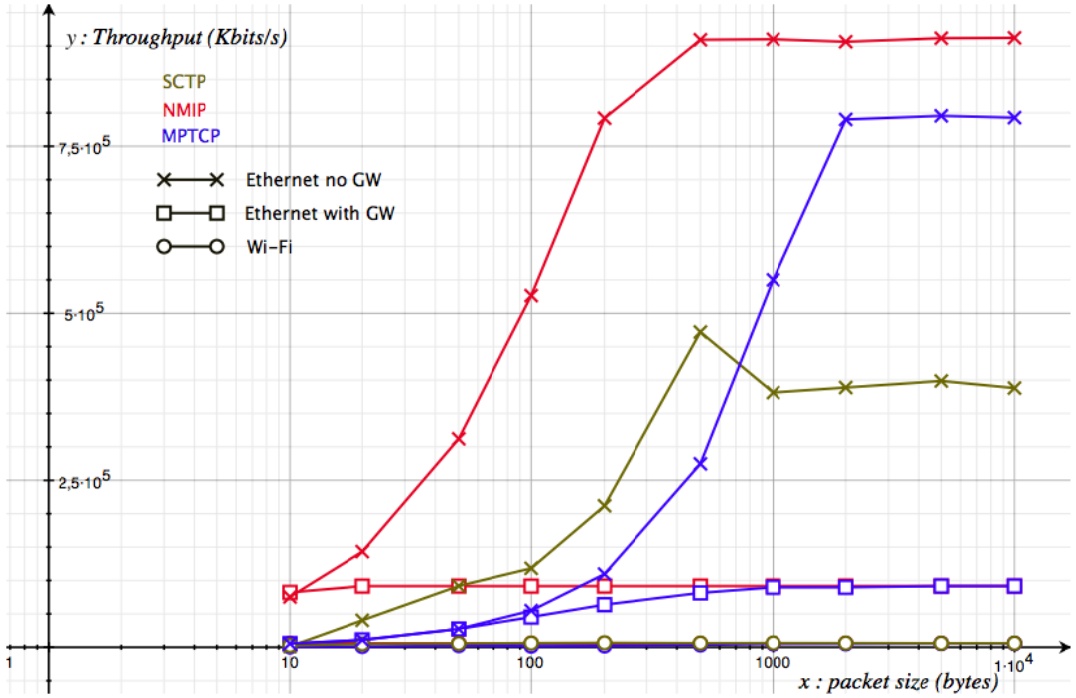


Figure 19 - NMIP SCTP and MPTCM on Ethernet No Gateway

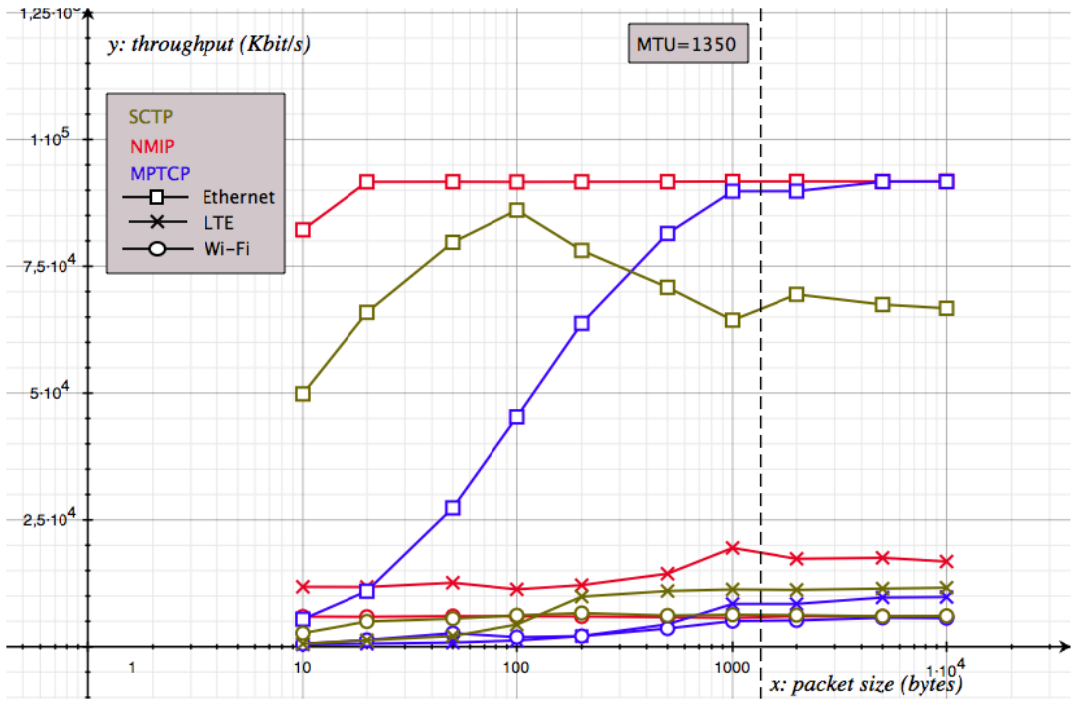


Figure 20 - NMIP SCTP and MPTCP for Ethernet Gateway LTE Wi-Fi

2.3.1.6 Applicability of the results

SCTP:

- This solution is believed to provide seamless mobility solution that addresses both changes in IP address and long network disconnections
- The session layer preserves communication upon changes in host's location
- It suspends communication upon long and short network disconnections and resumes communication upon network connectivity

Comparison of SCTP, NMIP, MPTCP:

For NMIP and MPTCP, the implementation of these protocols is relatively recent and probably lacks of maturity compared to TCP.

NMIP may be a good choice if the throughput efficiency is the main criteria.

The interest of SCTP is the ability to manage the failover issue by allowing the secondary path to be use when the first one fails.

2.3.1.7 Partners involved

ALBLF (NMIP, SCTP, MPTCP), Ericsson (SCTP)

2.4 Flat and distributed mobility management

This section describes the validation of flat or distributed mobility management protocols. These protocols need network infrastructure. Several candidate technologies are investigated.

2.4.1 Functional and performance validation of SIP based Ultra Flat Architecture (SIP-based applications)

2.4.1.1 Objectives

The objective here is to validate UFA concept and evaluate the performance of the mobility procedure for SIP-based applications. The overall mobility procedure is described in D2.2.

UFA handover delay is contains the items shown by Figure 21:

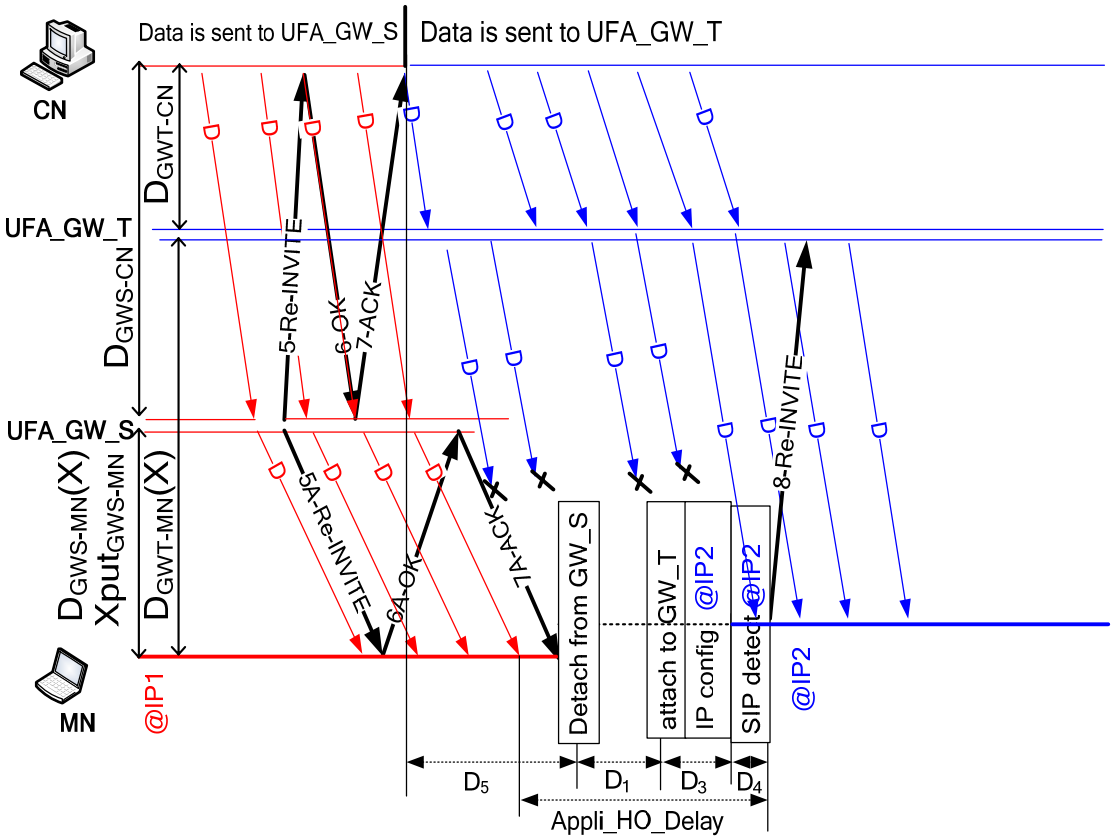


Figure 21 – UFA handover delay components.

2.4.1.2 Validation scenarios

To measure the handover delays, the hard handover case is considered as it is the most restricting one. The handover delay is measured at the application level (Appli_HO_Delay). It

is the delay in the MN between the last application data (D) packet received before handover, and the first application data packet received after handover.

2.4.1.3 Validation tools

UFA concept has been implemented on a testbed that consists of:

- Correspondent Node (CN): A desktop running Fedora Core 7 with kernel version 2.6.23.
- 2 Gateways (GW): Two desktops running Ubuntu 8.10 with kernel version 2.6.28.2. They act as Wi-Fi Access points and have 3com Wireless a/b/g PCi adapters based on Atheros chipsets.
- Mobile Node (MN): A laptop running Ubuntu 8.10 with kernel version 2.6.28.2. It has a PCMICIA wireless Netgear a/b/g card based on Atheros chipset. It uses a Wi-Fi link to connect to the GWs.

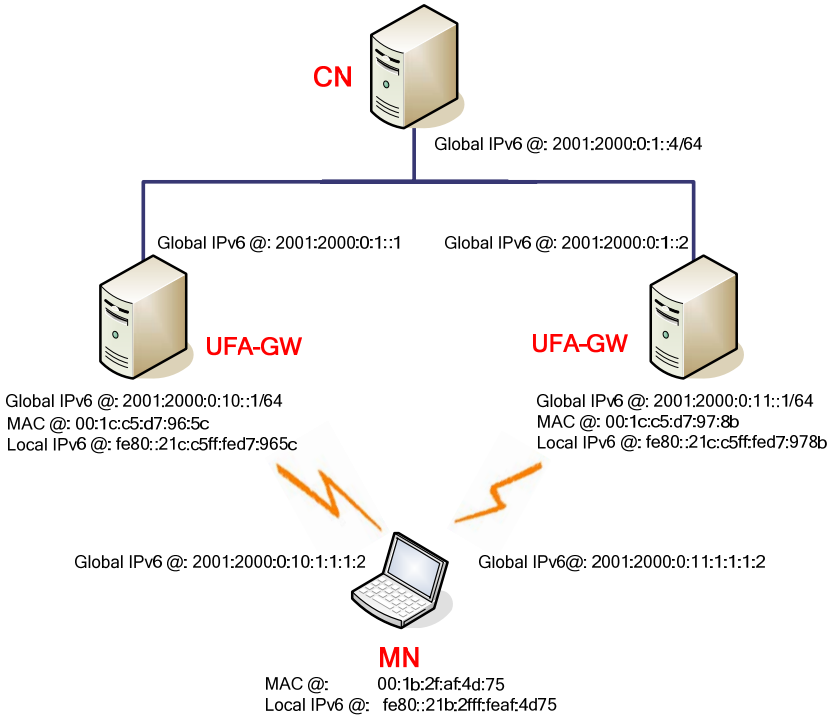


Figure 22 – Testbed configuration.

2.4.1.4 Expected results

Measurements of handover delay components.

Low application handover delay.

2.4.1.5 Results

For UFA, the different handover delays measured on the testbed are as follows:

- The number of lost packets is 4 based on RTCP packets as shown in Figure 23.
- The application handover delay Appli_HO_Delay is 80ms based on the number of lost packets (4*20ms). Figure 24 shows that Appli_HO_Delay is 60ms (packets 952 -> 957).

The difference between the two results (80ms and 60ms) may be due to the fact that wireshark captures packets while the MN interface is not completely attached.

- D1 is 60ms.
- D2 does not exist for UFA.
- D3 is 20ms, it corresponds to the delay necessary for the MN to enforce the IP configuration received from the Source UFA Gateway.
- D4 is 70ms. This delay does not impact the application handover delay (Appli_HO_Delay value (80ms)) measured on the testbed, as buffering in the Target UFA Gateway in not implemented. Indeed, D4 represents the delay in the MN at the

SIP layer to detect the new IP address and send message 8 (SIP Re-INVITE) to the Target UFA Gateway (UFA_GW_T). If this latter has buffered any data received from the CN or the Source UFA Gateway (UFA_GW_S), the reception of message 8 would trigger data sending to the MN.

- D5 is 0ms as messages 7A and 7 (SIP Re-INVITE) are received at the same time by the MN and the CN (delay on the links UFA Gateway-MN and UFA Gateway-CN links is low).

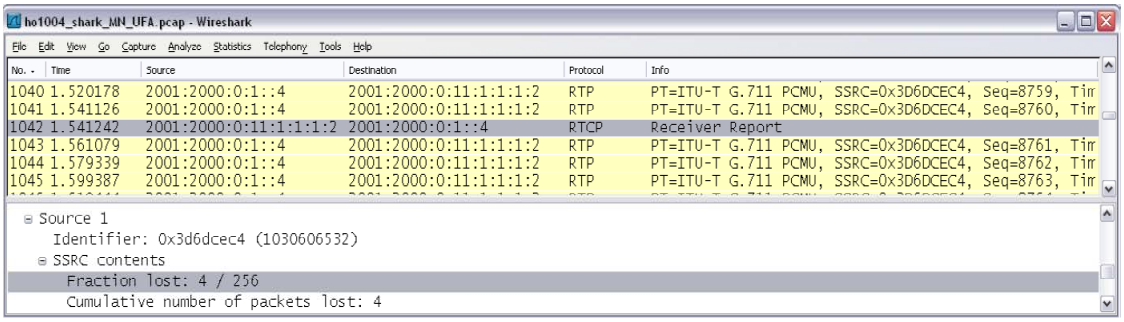


Figure 23 – Number of lost packets in UFA on the MN wireshark.

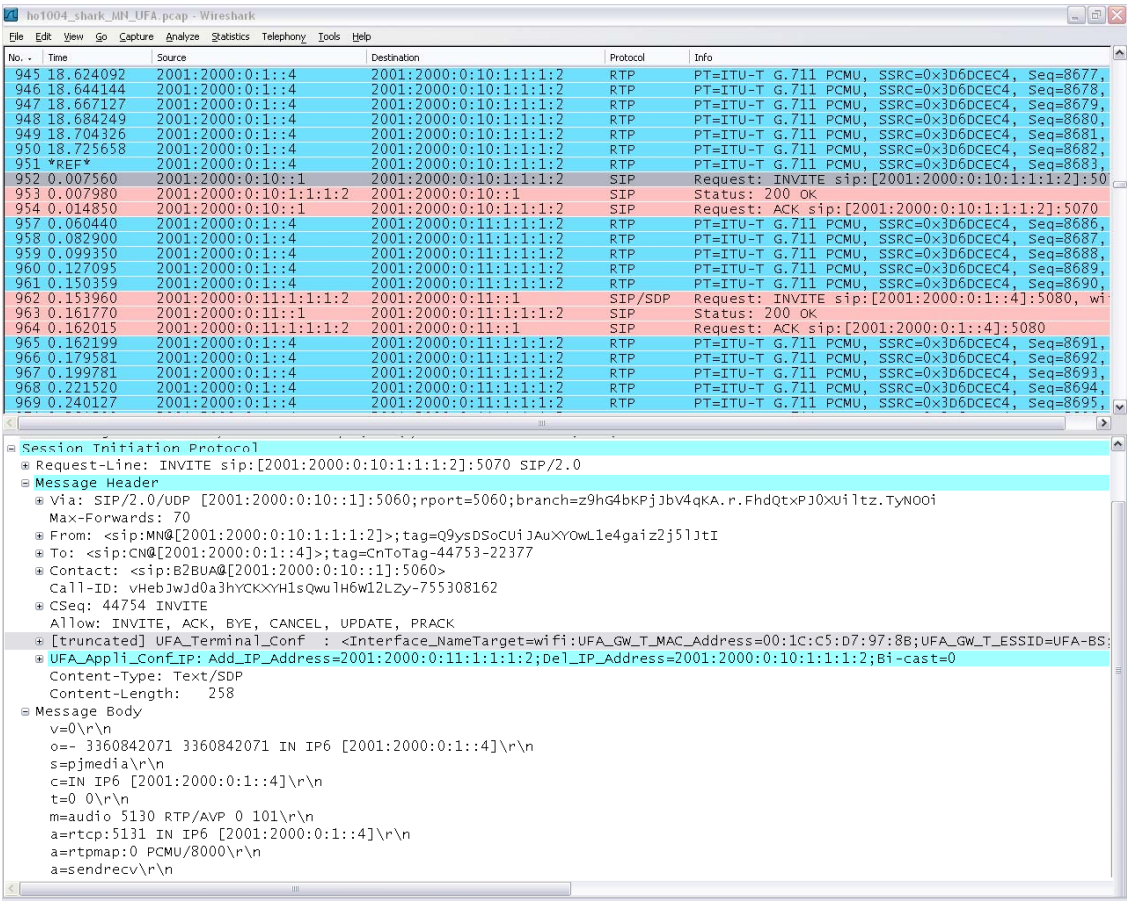


Figure 24 – Application handover delay in UFA on the MN wireshark.

2.4.1.6 Applicability of the Results

Results applicable for UFA-SIP, any application.

2.4.1.7 Partners involved

Regarding the evaluation of SIP-based UFA mobility management, Orange is the only partner involved in the validation

2.4.2 Functional and performance validation of SIP based Ultra Flat Architecture (non-SIP based applications)

2.4.2.1 Objectives

The objective here is to validate UFA concept and evaluate the performance of the mobility procedure for non-SIP based applications. The overall procedure for mobility in UFA is described in D2.2 [2].

Regarding non-SIP based applications, the specific case of applications based on SCTP in the user plane has been considered. In terms of mobility, these applications can be managed either by mobile SCTP protocol or by UFA. Therefore, to show UFA benefits, a comparison with m-SCTP is performed hereafter.

SCTP applications combined with m-SCTP face performance problems in hard handover situations: high network handover delay [6], high transport handover delay [6] [7] and throughput under-utilization [8]. In the following, based on a thorough analysis of these problems, we underline the limitations of current solutions based on terminal mechanisms and identify the need of anoverall mobility management architecture with network-controlled mechanisms. We then show the benefit of UFA and compare its performances to m-SCTP.

A-SCTP problems during hard handover

A-1 SCTP and m-SCTP overwiew

Multihoming feature brought by SCTP enables to establish an SCTP association between two endpoints taking into account a set of IP addresses and an SCTP port for each of these endpoints. Then, a path from one endpoint to a destination endpoint is characterized by one of the destination endpoint addresses. An endpoint chooses one path (called primary path) for data sending, the other paths (called secondary paths) being only used to retransmit data lost over the primary path. Like TCP, SCTP rely on specific data transmission and congestion control mechanisms to ensure reliable data delivery and efficiently use the available network resources. As paths may have different congestion states, SCTP sender separately maintains for each of them a set of congestion control parameters. These are the congestion window (cwnd), the slow-start threshold (ssthresh) and the Retransmission TimeOut (RTO). Cwnd limits the size of data a sender can send over a particular path without requiring any acknowledgement (SACK). To transmit data over a given path, SCTP first sets the congestion parameters to their default values (cwnd=2MTU [1], ssthresh=65536bytes, RTO=3s) and enters in slow start mode during which cwnd size increases exponentially. When cwnd reaches ssthresh, SCTP switches to congestion avoidance mode during which cwnd increases linearly. To control data delivery over a given path, the sender triggers a retransmission timer (T3-rtx) each time a packet is sent on that path. When T3-rtx reaches RTO (T3-rtx expiration) and data is still not acknowledged (non reception of SACK), the packet is considered as lost and SCTP falls back to slow start mode on that path with cwnd equal to 1 MTU, ssthresh divided by two and RTO doubled. The lost packet is retransmitted on a secondary path considering the congestion parameters of that path. On the other hand, if the sender receives 3 SACKs indicating that a given packet is missing (through GAP ACK Blocks field [1]) and the current T3-rtx has not expired yet, SCTP switches to fast retransmit mode by immediately retransmitting the missing packet on a secondary path without waiting for the expiration of the current T3-rtx .

To handle mobility, a new extension of SCTP called Mobile SCTP (**m-SCTP**) has been introduced in [4]. It makes possible the dynamic addition and deletion of IP paths to an established association through the usage of m-SCTP signaling messages (ASCONF). When a Mobile Node (MN) acquires a new IP address, it sends an ASCONF (ADD IP) to its Corresponding Node (CN) so that CN can consider the new address as a secondary path. Then, if MN wants that its new address is considered by CN as a primary path, it sends to CN an ASCONF (SET PRIMARY). After receiving an ASCONF (ADD IP), CN performs path verification towards the new address to become able to send data to that address. Path verification consists in sending a HEARTBEAT message to the new address and waiting for the reception of HEARTBEAT ACK. On the MN side, MN is allowed to use the new address for data sending only after receiving the ACK response to ASCONF (ADD IP), confirming that CN has received the ASCONF (ADD IP).

[1] MTU: Maximum Transmission Unit

A-2 SCTP problems

SCTP encounters a set of problems during hard handover when used with m-SCTP. We consider scenarios where MN is receiving data from CN and performing hard handover from one gateway (GW) to another one, both GWs belonging to different IP subnets. As illustrated in Figure 25, hard handover includes different time periods: (D_1) the time necessary for MN to detach from the Source GW (GW_S) and attach at Layer 2 to a Target GW (GW_T), (D_2) the time necessary to receive IP subnet information from GW_T through Router Advertisement (RA) [99], (D_3) the time necessary for MN to configure its interface with the new IP address and be able to use it (including DAD), and (D_5) the time necessary to exchange m-SCTP signaling messages (ASCONF (ADD-IP), ASCONF ACK, HEARTBEAT, HEARTBEAT ACK, ASCONF (SET-PRIMARY), ASCONF ACK). Network handover delay is the Layer 2 and Layer 3 disconnection delay and is equal to (D1+D2+D3). Transport handover delay (SCTP_HO_Delay) is the delay perceived by SCTP layer on MN side and represents the time between the last packet received before HO and the first packet received after HO.

Typical values for network handover delay (D1+D2+D3) vary between 1.5s and 3s [7]. These values directly impact the transport handover delay (see Figure 21). Indeed, during (D_1+D_2+D_3) period, CN continues to transmit data on MN old address which is no more reachable. Each time T3-rtx expires, CN retransmits the non acknowledged packet towards the same address (the only one known by SCTP) after setting (cwnd=1MTU, ssthresh=ssthresh/2, RTO=RTO*2) and arming T3-rtx with the new RTO value. When m-SCTP signaling is received by CN, CN cannot transmit data to the new IP address of the MN before T3-rtx expires on the old MN address [1], [6] and [7]. Thus, the larger the network handover delay (D_1+D_2+D_3) is, the higher the number of T3-rtx expiration is, the higher RTO is, and the higher the transport handover delay is.

On the other hand, if T3-rtx has never expired when m-SCTP signaling is received by CN (i.e D1+D2+D3 <1s[2]), CN transmits immediately new packets towards the new MN address without waiting for T3-rtx expiration [1], [6]. The SACKs generated by MN after receiving these new packets indicate the missing packets lost during handover. This trigger on CN fast retransmits to recover all of the missing packets. During the recovery period, cwnd associated with the new path remains constant [1] [9]. Consequently, even for low network handover delay SCTP performances are decreased due to packet losses.

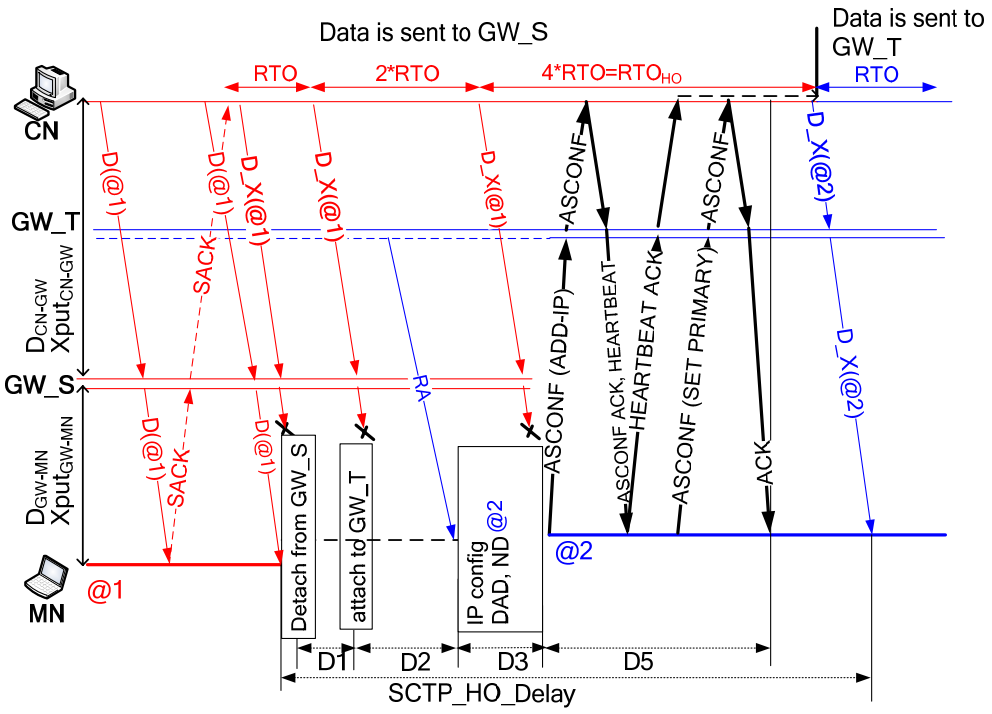


Figure 25 – Handover delay components with m-SCTP.

[2] minimal RTO value is 1s [1].

Another problem encountered by SCTP concerns throughput under-utilization on the new path after handover. Indeed as this path is initiated with default values for SCTP congestion control parameters (cwnd=2MTU, ssthresh=65536, RTO=3s), a period of time is necessary for the cwnd to reach the optimal value enabling the maximum usage of the network resources available on the link. This value is equal to the product of the bottleneck link throughput (Xput) and the sender-receiver link round trip time (RTT), named BDP (Bandwidth Delay Product) [10].

B- Related work

Regarding the long transport handover delay problem [6] and [7], supposing a high network handover delay, propose a solution based on m-SCTP that triggers a particular SCTP configuration we call **m-SCTP+**. Upon the reception of ASCONF (SET-PRIMARY), CN immediately retransmits data without waiting for the current T3-rtx expiration. The disadvantage of such solution is that it requires cross-layering mechanisms within MN to detect hard handovers and send ASCONF (SET-PRIMARY) message in addition to ASCONF (ADD-IP). Moreover, as hard handover is not the only case where ASCONF (SET PRIMARY) may be sent, triggering m-SCTP+ configuration may be inappropriate. [6] additionally shows enhanced performances for SCTP in case of low network handover delays; however it does not indicate the mobility architecture enabling such low delays.

Link throughput under-utilization problem is also encountered by TCP. Work in [11] addresses this problem as well as the long transport handover delay issue. It proposes a TCP-HO solution where MN reports to CN handover events and the BDP of the new link. CN stops transmission during one RTT and then begins transmission with cwnd equal to BDP. The disadvantage of this solution is that it does not specify how MN gets the BDP of the new link and does not take into account the delay necessary to get this information from the network. Reference [8] has the same disadvantage.

M-SCTP is only an end-to-end mobility signaling protocol: it does not provide tools to optimize mobility execution and relies on terminal mechanisms. The above analysis has proven that terminal mechanisms are not sufficient to deal with SCTP problems as they lack the necessary information to perform on-time and fine SCTP configuration tuning. Therefore an optimized mobility management architecture with network controlled-mechanisms driving SCTP configuration shall be defined. The first requirement of the target architecture is to reduce the network handover delay being the cause of the long transport handover delay. The second requirement is to support a proactive mechanism able to determine the throughput available on the target link without impacting the network handover delay. The third requirement is to provide SCTP with explicit triggers regarding handover events and adequate SCTP configuration containing new endpoint IP addresses and adapted congestion control parameters. This latter requirement could not be performed without using a dedicated signaling protocol that interacts with SCTP protocol.

C- UFA: a network-controlled architecture solving SCTP problems

UFA meets a part of the requirements discussed in the previous section: 1) it implements network-controlled cross-layer techniques driving terminals' configuration at all layers; 2) it provides a mobility procedure with proactive mechanisms and a reduced network handover delay; 3) it relies on SIP which is suitable to transport explicit triggers and information. The network control is enabled through the implementation in the UFA gateway (UFA_GW) of a SIP Back-To-Back User Agent (B2BUA) that modifies and generates SIP messages.

C-1: Support of SCTP applications by UFA

Each SCTP application is managed through a SIP session. The SIP session transports the SCTP application characteristics to the UFA_GW so that it can control the handover of this application. A SIPcrossSCTP (SxS) module is implemented within MN, CN and UFA_GW to maintain the binding between SIP sessions and SCTP applications and ensure the interaction between them.

UFA mobility procedure is based on two phases as shown in Figure 26: 1) a preparation phase (messages 1, 2, 3, 4) aiming at pre-determining the MN OSI layers configuration after its HO and the new CN SCTP layer configuration due to MN HO; 2) an execution phase (messages 5, 6, 7, 5A, 6A, 7A, 8) aiming at providing MN and CN with the predetermined

configuration. Both phases are controlled by the source UFA_GW (UFA_GW_S) as detailed hereafter. When UFA_GW_S anticipates the need of HO for MN because of coverage loss, it sends to a set of candidate UFA_GWs a RESOURCE QUERY REQUEST (1) that includes application characteristics, user profile, etc. Each of these candidates answers in RESOURCE QUERY RESPONSE (2) with *Allocated Throughput (*$Xput_{GW\_T\text{-}MN}$*)* according to the received information and available resources. UFA_GW_S then selects a target UFA_GW (UFA_GW_T) and pursues HO preparation by sending CONTEXT TRANSFER (3) to UFA_GW_T. UFA_GW_T pre-determines an IP address for MN (*Add_IP_Addr*), checks its uniqueness, confirms the *Allocated throughput,* calculates the associated $BDP_{GW\_T\text{-}MN}$ and includes them with other UFA_GW_T related information in ACK message (4) towards UFA_GW_S. Based on the received message, UFA_GW_S builds two SIP re-INVITE messages (5, 5A) sent to CN and MN respectively. SIP Re-INVITE (5) message towards CN includes *UFA_Appli_Config* header and $BDP_{GW\_T\_MN}$. SIP Re-INVITE (5A) message towards MN includes the same *UFA_Appli_Config* header and *UFA_Terminal_Config* header:

- *UFA_Appli_Config* header is depicted in Table 1. It indicates the new SCTP association addresses. With this header m-SCTP signaling is no more needed: the reception of message 7A by MN directly validates Add_IP_Addr as a new source address; and the reception of message 7 by CN directly validates Add_IP_Addr as a new destination address to MN.

- *UFA_Terminal_Config* is depicted in Table 2. It contains the reconfiguration necessary for MN Layer 2 and Layer 3 to handover to UFA_GW_T.

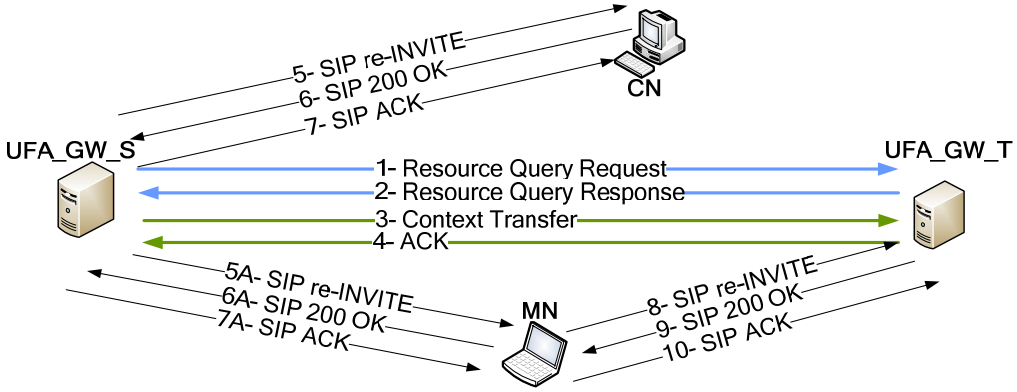


Figure 26 – UFA handover procedure.

Table 1 - SIP header for application configuration (*UFA_Appli_Config*).

Add_IP_Addr	2. The new MN address.
Del_IP_Addr	4. The old MN address.

Table 2 - SIP header for Layer 2/Layer 3 configuration (*UFA_Terminal_Config*).

UFA_GW_T_MAC_Addr	Used for Layer 2 HO
UFA_GW_T_ESSID	
UFA_GW_T_Channel	
UFA_GW_T_IP_Addr	Used for Layer 3 HO
UFA_GW_T_Netmask	
Add_IP_Addr	

UFA handover timing diagram is illustrated in Figure 27. When MN attaches to UFA_GW_T it sends a SIP Re-INVITE (8) message to UFA_GW_T. UFA_GW_T buffers data received from CN until reception of SIP Re-INVITE (8). We define D4 as the time necessary for SIP layer to detect IP address change and send SIP Re-INVITE (8). Compared to m-SCTP, UFA mobility mechanism enhances the network handover delay. Indeed the equivalent D2 delay does not exist and the equivalent D3 is very low as MN address determination and Duplicate Address Detection are performed proactively to HO execution. In section 2.4.1, (D1+D3+D4) was measured on a testbed to 150ms. This value is valid for SCTP as it is independent of the transport layer.

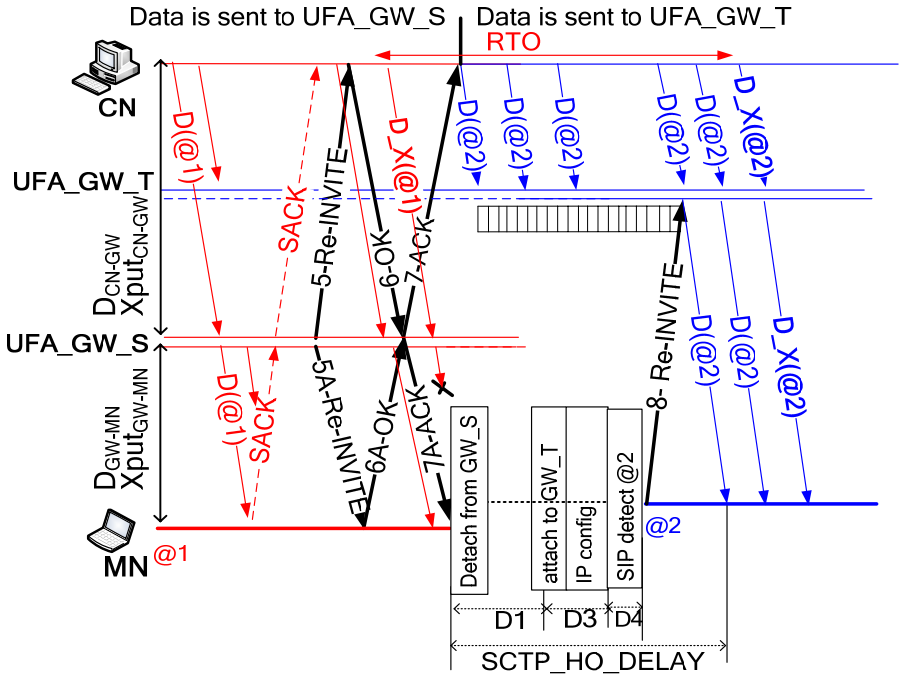


Figure 27 – Handover delay components with UFA.

2.4.2.2 Validation scenarios

We define three incremental UFA options depending on the configured SCTP parameters upon the reception UFA SIP messages for handover:

- **UFA**: is the minimal and basic SCTP configuration to support SCTP applications handover. It supposes the consideration of SIP headers presented in the previous section.
- **UFA+**: as handover delay in UFA is low, performance problems raised in section II.B appear. UFA+ resolves these problems by triggering on CN side after receiving message 7 immediate sending of lost packets before any new packet, preventing thus cwnd from remaining constant.
- **UFA++**: is based on UFA+ and solves in addition the link throughput under-utilization problem raised in section II.B. With UFA++, when $BDP_{GW\_T\text{-}MN}$ is received by CN within message 5, the SxS module calculates the SCTP congestion control parameter values related to Add_IP_Addr and informs the SCTP layer to immediately apply the calculated values. Cwnd is set to $BDP_{GW\_T\text{-}MN}$ which cancels the time necessary to attain this value, ssthresh is set to $BDP_{GW\_T\text{-}MN}$ and RTO is kept to 3s.

$BDP_{GW\_T\text{-}MN}$ is calculated by UFA_GW_T during handover preparation procedure using formulas (1) and (2) and considering the transmission delay of a 1500 bytes-length packet for the assessment of $RTT_{GW\_T\text{–}MN}$. $RTT_{CN\text{-}GW\_T}$ can be determined based on measurements performed by UFA_GW_T as described in [16].

$$BDP_{GW\_T\text{-}MN} = RTT_{CN\text{-}MN} * Xput_{GW\_T\text{-}MN} \quad (1)$$

$$RTT_{CN\text{-}MN} = RTT_{CN\text{-}GW\_T} + RTT_{GW\_T\text{–}MN} \quad (2)$$

During throughput allocation, UFA_GW_T checks whether its free buffer size is compliant with the rule of thumb (buffer_size = $BDP_{GW\_T\_MN}$) [17] and allocates it accordingly.

2.4.2.3 Validation Tool

In this section, UFA, m-SCTP as well as their enhanced configurations (UFA+, UFA++, m-SCTP+) are compared.

We construct a simulation model using Network Simulator 2.33 [18]. During a given simulation time (500 seconds), CN sends a file to MN. Hard handover of MN is simulated by periodically switching between two GWs. The switching periodicity determines the number of handovers (HO_nbr) occurring during data downloading. Links between CN and GW (respect. GW and MN) are characterized by a propagation delay D_{CN-GW} and a throughput $Xput_{CN-GW}$ (respect D_{GW-MN}, $Xput_{GW-MN}$). The receiver buffer size is 65536bytes. For m-SCTP and m-SCTP+, we consider 2s for the network handover delay (D1+D2+D3). Simulations are conducted using different network scenarios given in Table 3.

Table 3 - Considered network scenarios (Sc) for simulation

	Sc1	Sc2
D_{CN-GW} (ms)	10	100
$Xput_{CN-GW}$ (Mbps)	10	10
D_{GW-MN} (ms)	2	2
$Xput_{GW-MN}$ (Mbps)	1	Variable (0.1 …3)
HO_nbr	Variable (1..13)	6

2.4.2.4 Expected results

Better performances for UFA compared to m-SCTP.

2.4.2.5 Results

To compare the different performances, we measure different pertinent key performance indicators such as the transport handover delay, the mean throughput and the size of downloaded data during the simulation time (500s). Due to limited space, we only show here results for the most global indicator, which is the size of downloaded data.

We first compare m-SCTP, m-SCTP+, UFA, UFA+ and UFA++ for the network scenario Sc1 (see Table 3) considering different number of handovers (HO_nbr). A HO_nbr value corresponds to a given MN velocity. For exemple, HO_nbr=6 corresponds to a pedestrian walking at 5 km/h in an area covered by 100m-diameter cells or a user in a car travelling at 51 km/h in 1km-diameter cells. Figure 28 gives the additional data volume m-SCTP+, UFA, UFA+, UFA++ enable to download compared to m-SCTP. For m-SCTP, the downloaded data volume is 62, 61, 60, 59, 57 Mbytes for HO_nbr equal to 1, 3, 6, 9, 13 respectively. We observe that all UFA options enable to download more data volume than m-SCTP. Moreover they are more efficient than m-SCTP+ considered in the state of the art as the best enhancement to m-SCTP performance with regards to the long transport handover delay: m-SCTP+ enables a gain ranging from 0.2% to 2% compared to m-SCTP; and UFA enables a gain ranging from 0.4% to 7.8% compared to m-SCTP. For this network scenario Sc1, UFA+ and UFA++ do not provide remarkable gains compared to UFA as both D_{CN-GW} and BDP_{GW_T-MN} are low. These enable in UFA lost packets to be recovered rapidly (SACKs are returned rapidly) and cwnd to attain rapidly BDP_{GW_T-MN}.

We therefore compare the performances of UFA, UFA+ and UFA++ using the network scenario Sc2 (see Table 3) considering higher values for D_{CN-GW} and BDP_{GW_T-MN} (higher value for $Xput_{GW-MN}$). We set the receiver buffer size to 200000bytes in order to take into account the high throughputs ($Xput_{GW-MN}$ is 2Mbps or 3Mbps). Figure 29 shows the additional data volume UFA+ and UFA++ enable to download compared to UFA. For UFA the downloaded data volume is 6, 28, 57, 116, 168 Mbytes for $Xput_{GW-MN}$ equal to 0.1, 0.5, 1, 2, 3 Mbps respectively. We observe that compared to UFA:

- UFA+ enables a gain varying from 2% to 7% for an $Xput_{GW-MN}$ varying from 1 Mbps to 3 Mbps, and
- UFA++ enables a gain varying from 4% to 9% for an $Xput_{GW-MN}$ varying from 1 Mbps to 3 Mbps.

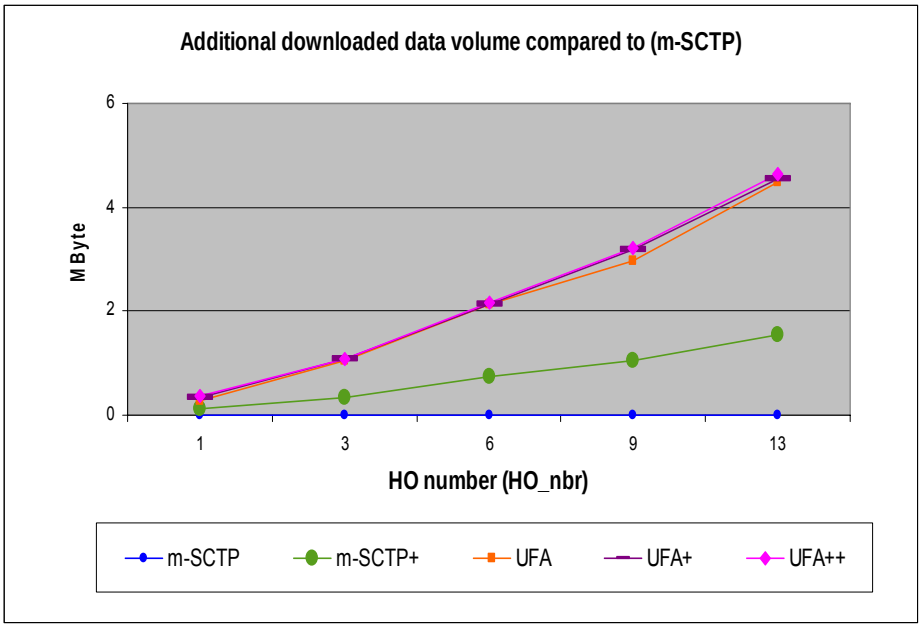


Figure 28 - Performance comparison for network scenario Sc1.

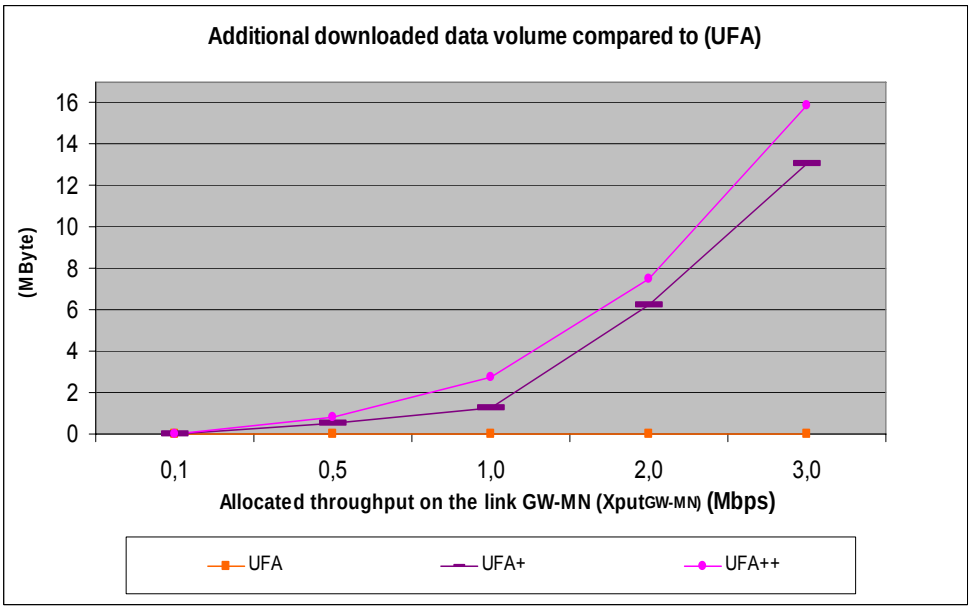


Figure 29 - Performance comparison for network scenario Sc2.

We conclude that UFA+ and UFA++ provide better performances than UFA. In general, although the additional downloaded data volume may appear relatively low, this one shall not be neglected as it has been calculated during a short time period (simulation time=500s) and for a single terminal. The gain for an operator is important as it is proportional to the number of terminals and the duration of data downloading (higher than 500s).

2.4.2.6 Applicability of the Results

Results applicable for SCTP based applications application managed by UFA-SIP.

2.4.2.7 Partners involved

Regarding the evaluation of SIP-based UFA mobility management, Orange is the only partner involved in the validation

2.4.3 Functional and performance validation of PMIPv6 Route Optimization

2.4.3.1 Objectives

Proxy mobile IPv6 (PMIPv6) is specified to operate on S5, S8, S2a, and S2b interfaces. To that end PMIPv6's functional elements are located on the PGW for the local mobility anchor (LMA) and on RAN's gateways (e.g., SGWs, ePDGs, A-GWs, etc.) for the mobility anchor gateways (MAGs). PMIPv6 has been specified in [RFC5213]. This mobility management protocol relies on a centralized approach where the LMA (PGW) is:

- Responsible of tracking locations of UEs
- The data anchor of all traffic
- The signalling anchor
- The IPv6 anchoring point of all UEs for traffic coming from the Internet.

Furthermore, the location of the LMA on the PGW helps to complete on-flow services that are part of the functions of the PGW such as per-user packet filtering, lawful interception, charging, etc.

The route optimization solution addresses the problem of centralized mobility anchoring (challenge C.Mo.3) in PMIPv6. The idea is to distribute the data anchor function of the LMA and some functions of the PGW on intermediate anchors (IAs) between MAGs. The objectives of the solution are twofold: propose an approach that reduce unnecessary load at the LMA and provide a set of methods that allows transferring the data anchoring role from the LMA to distributed servers (SGWs, ePDGs, PGWs, Intermediate Anchors, or specific servers) close to MAGs and /or UEs. The transfer of role would allow having a mobile functionality that would optimize routing within a PMIPv6 domain. Hence, the specific KPIs are:

- Throughput of data traffics: reduced load at the LMA should in some cases (according to the underlying transport network) reduce the number of concurrent flows at the LMA and then increases their performances.
- Packet loss: related to the throughput of data traffics' KPI, by reducing the number of concurrent flows, one would also reduce the probability of packet loss.

Number of routing hops: in a distributed deployment, the number of IP hops between two MAGs (passing through the LMA) should be important because of triangular routing. The routing optimization solution will help to reduce the number of hops by relying on close-by data anchoring server(s).

2.4.3.2 Validation scenarios

PMIP-RO is intended to be used in different setup.

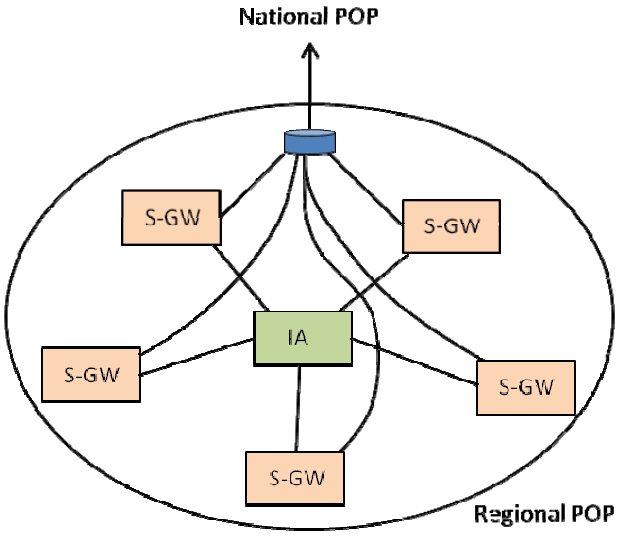


Figure 30 - Intra-POP scenario of routing optimization.

One may consider that one IA will be located in sub-areas of the domain (e.g., one POP) to handle optimization that could occur between close by MAGs (see Figure 30). However, the operator may choose different interconnection approaches such as a mesh-like scheme or a tree-like one. Different performance evaluation will be proposed considering this topology choice.

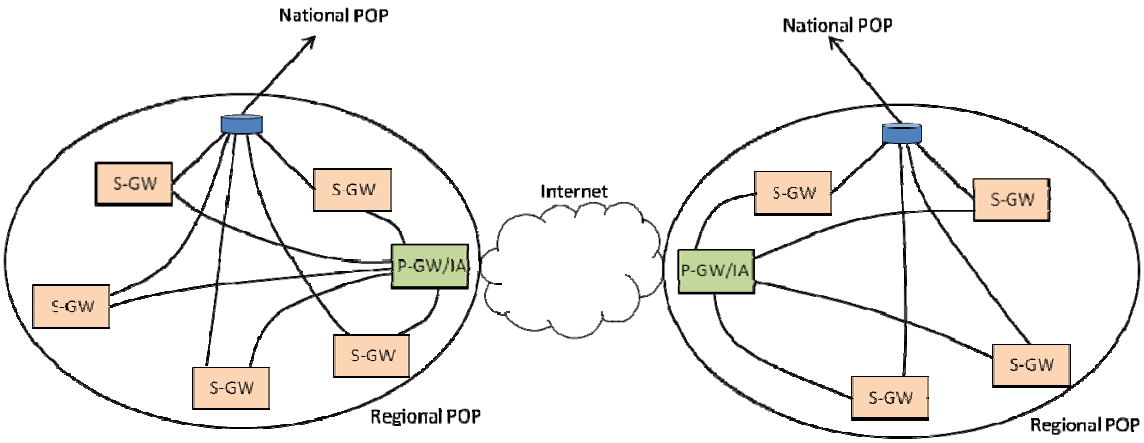


Figure 31 - Inter-POP scenario of routing optimization.

In an inter-POP routing optimization scenario, the intermediate anchors could be co-localized with regional PGWs to ensure rapid offloading in the Internet (see Figure 31). Here, the optimized route would pass through both IAs at the regional level bypassing the national POP where all un-optimized data traffic converges.

2.4.3.3 Validation tools

A functional validation is proposed for PMIP-RO. The validation relies on a laboratory testbed that implements the current specification of PMIPv6 from the RFC5213 and the PMIP-RO specification developed in D2.2.
PMIP-RO is implemented in C (the programming language) directly in our implementation of PMIPv6 and runs in userspace on Linux machines (Ubuntu and Debian). Note that running in userspace introduces a slight delay between the time a signaling message is passed from the kernel space and the userspace and the time when the related network layer modification is committed. Through a configuration file, it is possible to select which role will be played by the program, i.e., MAG, LMA, or IA.

As presented on Figure 32, the testbed is composed by at least 4 machines interconnected together to form the core network (EPC). A Fast Ethernet 100 Mbit/s switch interconnects the LMA and MAGs, while the IA uses 3 independent Fast Ethernet wired links with the MAGs and the LMA. Therefore, traffics from clients are aggregated between MAGs and the LMA, while they can be isolated from the others by passing through the IA. Clients are connected to MAGs with IEEE 802.11 Fast Ethernet wired links.
Note that this architecture may be considered as an offloading one, as the IA may be located out of the core network.

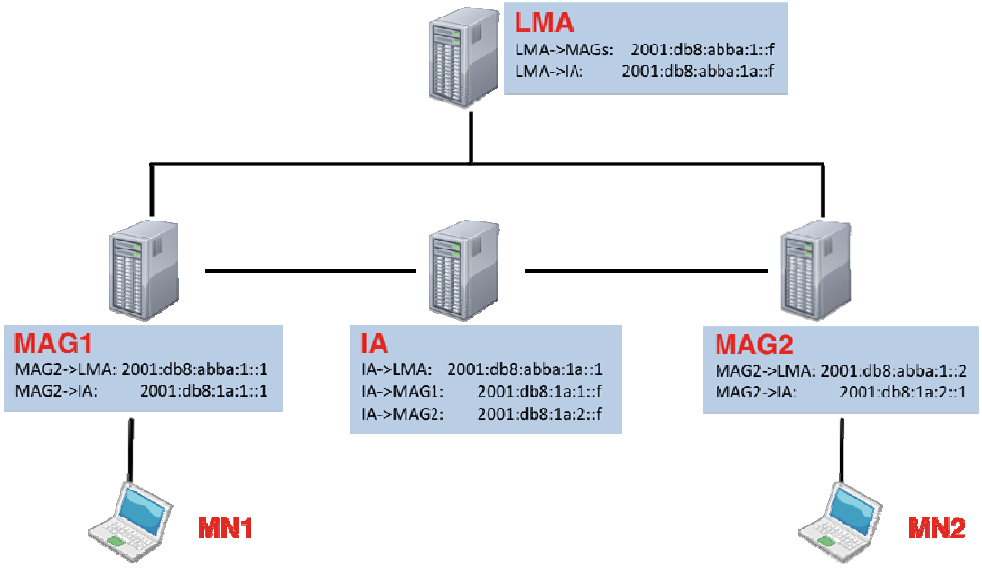


Figure 32 - PMIPv6 Routing optimization testbed.

The number of clients varies according to the validation scenario. However, at the attachment with a MAG, a client (MN1 or MN2 in Figure 32) obtains an IPv6 prefix from the PMIPv6 system through a Router Advertisement message. The attributed IPv6 prefix is anchored at the LMA and advertised by the MAG on the "home network link" (the link between the client and the MAG). Using the stateless address auto-configuration procedure, the client constructs its own IPv6 address by merging the received prefix with its MAC address. In one scenario, a corresponding node (CN) is used and connected beyond the LMA (emulating a server in the Internet) uses a fixed IPv6 address with a default route towards the LMA. The CN runs as a virtual machine on the LMA.

The performance of PMIP-RO is measured using the bandwidth measurement tool iPerf. According to the scenario, a MN or the CN will act as a TCP or UDP server and traffics will be generated in the network between MNs.

In all cases, the routing optimization is triggered when the LMA detects at least two MNs attached to different MAGs and after tens of second to have time to generate the traffic flows manually.

2.4.3.4 Results

We first focus the results on the analysis of a data stream behavior during the routing optimization initiation procedure (LRI), i.e., during the modification of the generic data path (MAG→LMA→MAG) to the optimized one (MAG→IA→MAG). Specifically, we are interested to evaluate if the specified procedure is efficient enough to have a reduced impact on an on-going communication, i.e., packet loss, delay, etc. To that end, TCP traffic is considered, as the included congestion control mechanism is sensitive enough to highlight packet reordering and retransmissions which may occur. Through monitoring of end-to-end bandwidth, TCP sender congestion window and, evolution of TCP sequence number, we will be able to evaluate the impact of PMIP-RO on the flow.

The scenario considered for this analysis is the following: MN2 is attached to MAG2 and runs an iPerf IPV6 TCP server. The MN1 is attached to MAG1 and establishes the iPerf TCP stream with MN2. During a certain period of time, the traffic follows the generic routing path passing through the LMA. At time 85 seconds (a semi-random delay), the routing optimization is triggered by the LMA, making the IA the new data anchor of the TCP flow eventually. The impact of this optimization is presented below.

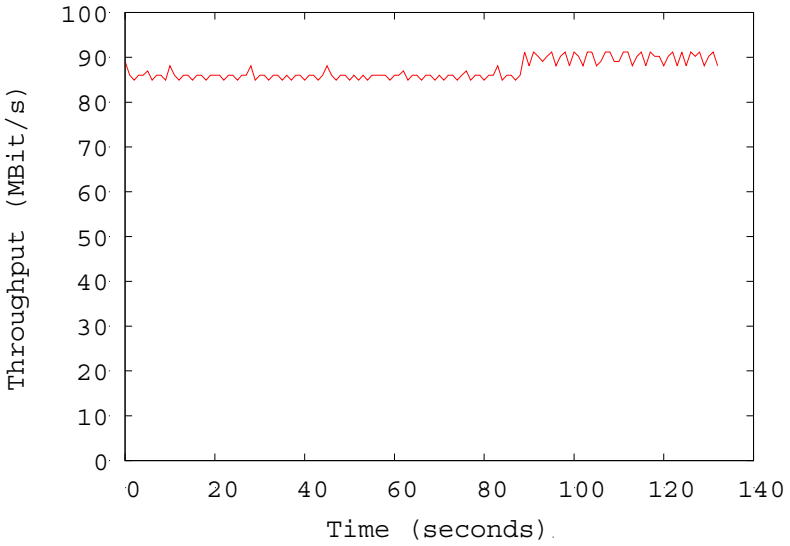


Figure 33 - Throughput evolution before and after routing optimization. The optimization is triggered at time 85 seconds.

On Figure 33, the end-to-end throughput of the TCP flow is presented. The data path capacity is 100 Mbit/s. One observes first that as there is only one flow running in the network, the achieved throughput is close to the data path(s) capacity. At time 85 seconds, there is a small improvement of the throughput marking the routing optimized state. The small improvement was not expected as there were no concurrent flows but do highlight that the throughput has not been negatively affected during the LRI procedure, i.e., there is no (visible) drop in throughput at time 85 seconds.

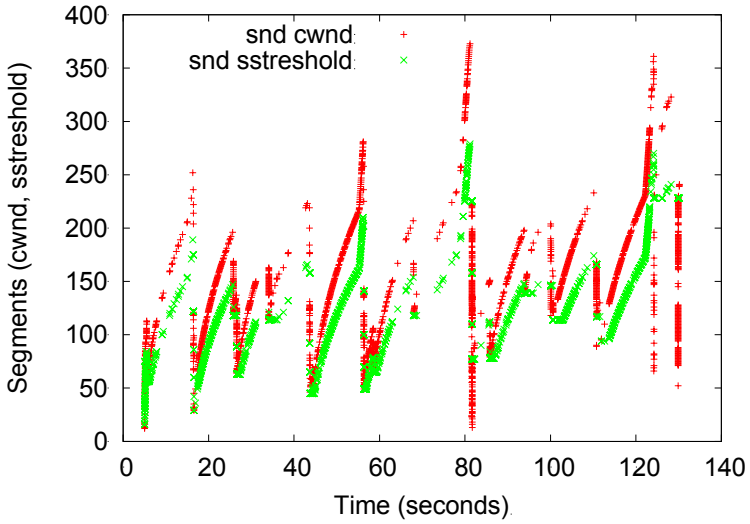


Figure 34 - Evolution of the TCP sender congestion window (snd_cwnd) and slow start threshold (snd_sstreshold). The optimization is triggered at time 85 seconds.

To get the impact of the LRI procedure more in details, we analyze the sender congestion window during the transmission. We can observe on Figure 34, that the congestion window during the whole experiment follows the expected TCP pattern. At time 85 seconds, one observes that the congestion window was small and in the additive increase step of the TCP congestion control procedure but no clear impact of the LRI procedure.

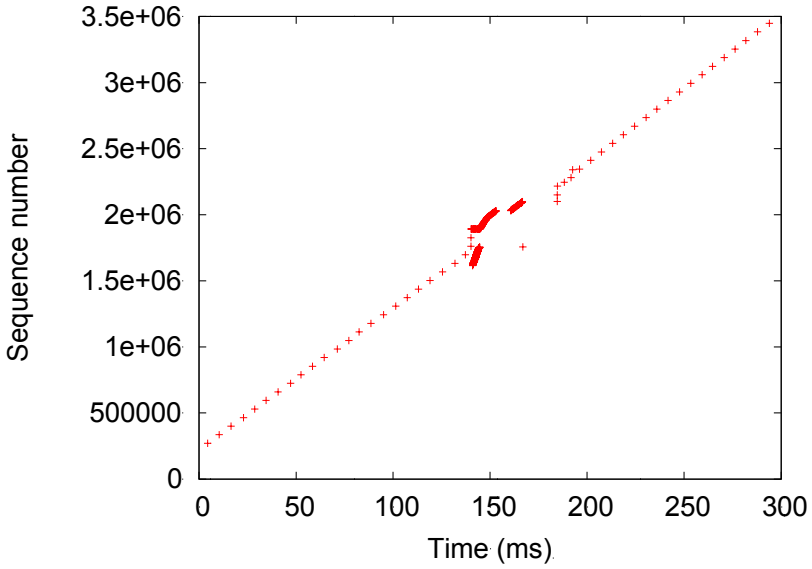


Figure 35 - Evolution of the TCP sequence number during the optimization procedure.

To conclude this analysis, we monitor the evolution of the TCP sequence number during the LRI procedure (see Figure 35). One observes clearly an impact of the modification of the data path on this parameter. The effect last for a little less than 50 ms and do show some retransmission of packets due to missing acknowledgment packets. However, it is important to note that the evolution of the TCP sequence number goes back to normal rapidly and that there is no drop of connection.
We then consider that the optimization procedure has a negligible impact. This observation is confirmed by the analysis of the evolution of the congestion window and the end-to-end throughput on a TCP flow in our setup.

We now evaluate the performance improvement of a TCP traffic in presence of concurrent flows. To this end, our setup is slightly different. We consider the presence of a CN beyond the LMA which will host two iPerf UDP servers. We limit the bandwidth between MAG1 and LMA as well as between MAG2 and LMA to 40 MBit/s. Mobile node 2 hosts a iPerf TCP server and initiates an UDP bidirectional flow with CN at 16 MBit/s, i.e, one UDP flow is initiated from MN2 to CN at 16 MBit/s, CN in turn initiates an UDP flow towards MN2 with the same targeted bandwidth. We then have 32 Mbit/s of UDP traffics between MN2 and CN. Mobile node 1 initiates one UDP bidirectional flow with CN at 14 MBit/s and a TCP traffic with MN2. We then have 28 MBit/s of UDP traffics between MN1 and CN. Because the links are limited to 40 MBit/s, the available bandwidth for the TCP flow is around 12 MBit/s on the link between MAG1 and LMA, and 8 MBit/s on the link between LMA and MAG2. At some points the TCP flow is optimized and we analyze its performance on Figure 36.

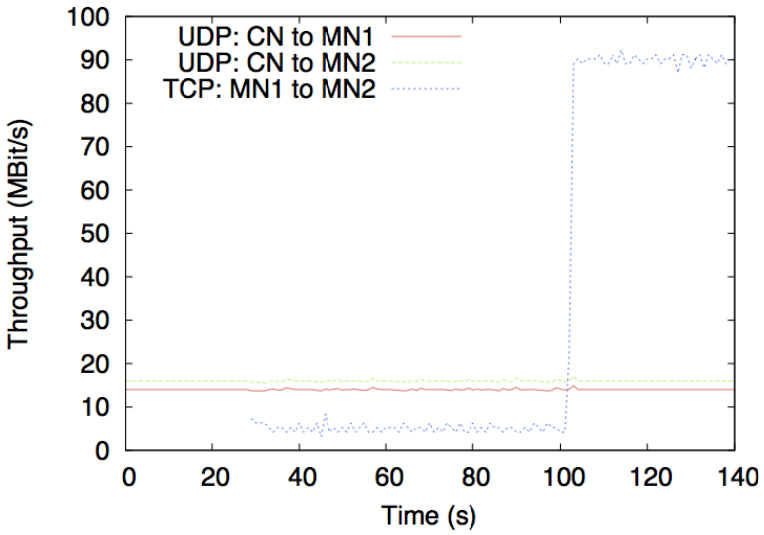


Figure 36 - TCP flow performance before and after optimization.

One observes that before the TCP flow initiation, the UDP traffics are stable, i.e., they are not affected by the limited available bandwidth. When the TCP flow is initiated at time 29 second, the reached TCP throughput is around 4 and 7 MBit/s, while the UDP traffics undergo some fluctuations. At time 102 second, the routing optimization is enabled for the TCP flow and one observes a rapid increase of the throughput which reaches 90 MBit/s one second later. Meanwhile, the UDP traffics stop fluctuating to remain stable at their targeted bandwidths.

2.4.3.5 Applicability of the results

PMIP-RO is an extension to PMIPv6. Recall that the LMA in PMIPv6 is considered as located on the PGW and MAGs on SGWs, ePDGs, or other RANs' specific gateways (S5, S8, S2a, S2b interfaces). In such a context, the routing optimization solution takes advantage of the underlying network infrastructure to interconnect MAGs directly (reducing the indirection caused by the LMA) or by relying on intermediate servers (IA) located close to them.
Nowadays, the centralized functional architecture does assume direct communications between SGWs and PGWs as they could be in some cases collocated at national POPs. However, in distributed deployments, where SGWs could be located at different POPs than PGWs (e.g., SGWs at local POPs and PGWs at regional POPs), one may assume interconnection between SGWs to optimize data traffic routing of users below the same POP for instance.

2.4.3.6 Partners involved

On this route optimization topic within the 3GPP EPC relying on PMIPv6, CEA LIST is the only partner involved.

2.4.4 Functional and performance validation of PMIPv6 with NEMO support

2.4.4.1 Objectives

As stated in MEVICO, the proportion of mobile networks will dramatically increase in the next few years. The examples of such networks are airplanes, trains, vehicular, or boats. The support of such kind of networks is then crucial for the 3GPP EPC core network. So far, PMIPv6 does support mobility management of mobile UEs through "home links" mobility management, i.e., the link between the MAGs and the UEs. However, such an approach do not consider, in terms of routing, the possibility that the UE (here the mobile router) provides connectivity to interconnected UEs (here local fixed nodes "LFNs"), i.e., a moving network. The objective of our solution is then to provide an extension to PMIPv6 in terms of data structures and routing operations to support moving networks (C.Mo.8).

The key performance indicators are:

- The PMIPv6 database overhead: the modifications of PMIPv6 data structures will introduce more information in the signalling messages and will add complexity in the routing procedures.
- Handover performance: we will evaluate the impact of the evolution of data structures on the handover delay.

2.4.4.2 Validation scenarios

The considered deployment scenarios of the 3GPP EPC in MEVICO should not have any impact on the extension's performance. Therefore, the validation scenario considers a typical PMIPv6 architecture with at least two MAGs to perform handovers. The expected validation scenario is depicted in Figure 37.

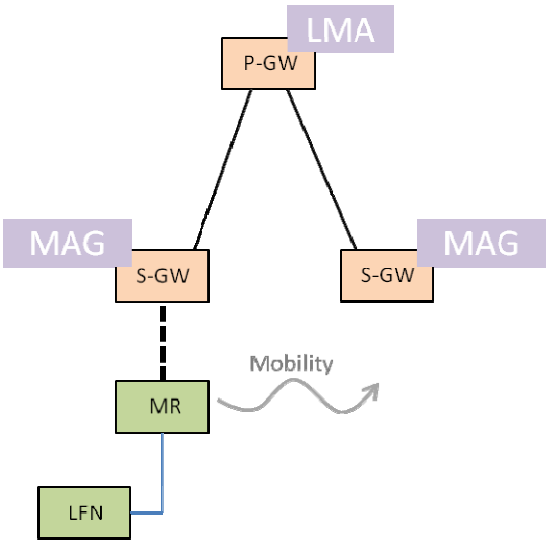


Figure 37 - Validation scenario of PMIPv6-NEMO.

From the two specified solutions to bring support of moving networks in PMIPv6 (see Section 2.9.1 in D2.2 [2]), only the second requires modification of the PMIPv6 functional elements. For the sake of clarity, the latter solution is selected for functional validation.

In a non-moving network, the support of a sub-network can be ensured by a functionality of DHCPv6 called Prefix Delegation (DHCPv6-PD). IPv6 prefixes are provided to the requesting router that will be, eventually advertised to leaf UEs (LFNs). However, in a moving network context, adapted mechanisms to ensure routability of delegated prefixes must be integrated in the mobility management protocol (here, PMIPv6).

Hence, this solution considers the interaction of PMIPv6 with one DHCPv6 server located in the core network, collocated with the LMA or not. In the DHCPv6 architecture, a delegating router provides delegated prefixes, which could be the DHCPv6 server. We consider that the delegating router is the DHCPv6 server and the terms will be used alternatively. The support of moving networks is enabled by the support of DHCPv6's prefix delegation functionality between the mobile router and the DHCPv6 server and by the management of the delegated prefixes in the PMIPv6 architecture.

In a nutshell, the functional validation then requires implementation of new functionalities in different network elements.

On the MAG:

- A DHCPv6 relay
- Functions to interpret received prefix delegation request messages and to extract delegated prefixes from the DHCPv6 server.
- Function to replace the DUID value by the UE's PMIPv6 MNID in DHCPv6 messages relayed to the DHCPv6 server.
- Function to update routing tables and routing rules according to approved delegated prefixes.
- Modification of clients' data structures to integrate delegate prefixes.
- Management of delegated prefixes by the MAGs. It includes: 1) the generation of a PBU message (and new mobility option) to be sent to the LMA including the delegated IPv6 prefixes, 2) modification of routing rules and routing tables at the MAG.

On the LMA:

- Modification of the client data structure to incorporate the delegated prefixes in the LMA.
- Handling of received PBU including delegated prefixes and generation of PBA message to be sent to the MAG to acknowledge the delegated prefixes.
- Modification and management of new routing table entries.

2.4.4.3 Validation tools

The validation will be performed on the CEA LIST PMIPv6 testbed presented in Section 2.5.4.3. The extension will be implemented on laptops as PMIPv6 platform, mobile router and LFN. Common performance evaluation tools will be used to consolidate the results, e.g., iPerf, tcptrace, wireshark, etc.

2.4.4.4 Expected results

As we propose a functional validation for a new service, we consider the service is validated if we are able to provide an IPv6 address to one LFN and that the implemented mechanisms within PMIP functional elements are able to route and deliver LFN's data traffic to a corresponding node (CN) beyond the LMA.

The expected results are: 1) a clear interaction between the DHCPv6 signaling plane and the PMIPv6 architecture. 2) Implementation of all functions to intercept and handle those signaling messages. 3) Correct management of delegated prefixes in the PMIPv6 architecture. 3) The support of handovers.

2.4.4.5 Results

According to the presented validation scenario, the current status of the implementation allows the basic support of moving networks in PMIPv6. Below, is the table which summarizes the current status of functionalities that were planned to be implemented.

Functionality	Status
Interception of DHCPv6 messages	Supported
Parsing of DHCPv6 messages	Supported
Modification of provided DHCPv6's DUID by registered user's PMIPv6 MNID	Supported
Interception of allocated IPv6 prefixes (Prefix Delegation)	Supported
Interception of released IPv6 prefixes (Prefix Delegation)	Not supported
Interception and management of status messages	Not supported
Relay forward and relay reply messages	Supported
Extension of binding list structure to register allocated Prefixes	Supported
Transmission of allocated prefixes in PBU to LMA	Supported
Transmission of allocated prefixes in PBA to MAG	Supported
Parsing and registration of received prefixes in LMA	Supported
Modification of routing table in LMA	Supported
Modification of routing table in MAG	Supported
Modification of routing rules in MAG	Supported
Prefixes lifetime management	Not supported

2.4.4.6 Applicability of the results

Considering the deployment of the extension, the solution will be directly applicable to any 3GPP EPC network supporting PMIPv6. The level of distribution of gateways should have no impacts on the performance of the extension.

This solution does not require modification of LFNs, which are considered as standard IPv6 capable UEs. However, it is sometimes assumed that mobile router runs the NEMO protocol, which is not compatible, so far, with PMIPv6. Therefore, the mobile router should be modified to send Prefix delegation request to a DHCPv6 server in the EPC and to handle the received prefixes accordingly. The NEMO functionality must be deactivated while the router is handled by PMIPv6.

2.4.4.7 Partners involved

On the support of moving network in PMIPv6 topic, CEA LIST is the only partner involved.

2.5 User access authorization

This topic validation covers terminal attachment related functional and performance validations. Within that, new L3 access authorization schemes are investigated. The main challenges to be covered here are reduction of security setup overhead and seamless interworking with different access technologies.

2.5.1 Performance evaluation of new HIP access authorization methods compared with IKEv2-based methods in the operator controlled Wi-Fi accesses

2.5.1.1 Objectives

This topic is related with the performance problems described in D2.1 [1], Section 5.1.2, i.e., the first phase of service continuity during inter-GW handover. The results are expected to support our decisions on which technologies and authentication methods should be selected in distributed EPC where the first IP gateway may be located in the national, regional, or local Point of Presence (POP).

This work evaluates the performance benefits of Host Identity Protocol Diet Exchange (HIP DEX), and HIP DEX with Authentication and Key Agreement (AKA) protocols. HIP DEX AKA provides similar functionality to the Internet Key Exchange protocol v2 (IKEv2) with EAP-AKA that controls user access authentication and authorization of USIM based UEs in non-managed non-3GPP access networks. With HIP mobility extension, it provides the same functionality as MOBIKE, i.e., supports IP mobility while the UE remains attached to the same ePDG. Both services provide mutual authentication and establish an IPsec security association pair to protect the path between the UE and the ePDG in the network layer.

Several challenges of the EPC have been identified in the mobility work package. This technology may tackle the following challenge:

- Reduce security setup overhead. This challenge becomes more important if GWs are distributed and pushed down to regional and local POPs, because at each inter-GW handover user access authorization must be controlled.
- We plan also to examine whether seamless handover is achievable for real-time services during re-attachment procedures without making further optimizations of the technologies.

HIP DEX AKA scheme reduces both message and computational overhead of IKEv2. HIP DEX and HIP DEX AKA are both lightweight due to the use Elliptic Curve (EC)-based public/private key cryptography instead RSA-based keys. The HIP DEX protocol forfeits signatures and hash function from the security parameter negotiation and uses CMAC for message authentication. These changes are expected to lead to lower CPU utilization on the UE and GW and make it more suitable for constrained devices and in architecture scenarios where frequent inter-GW handovers take place.

In case of IKEv2 with EAP-AKA authentication the IP gateway (ePDG) relies on the authentication service of a centralized AAA server. The AAA server is responsible to query authentication vectors for the user from the Home Subscriber Service (HSS). This communication path is simplified because in case of HIP DEX with AKA the gateway provides the network-side authentication service, i.e. the gateway directly asks authentication vectors from the HSS.

Scalability of this solution will be investigated later as it is anticipated that the protocol requires signalling optimizations towards HSS in this regard. The solution will be optimized in the later part of the MEVICO project and further evaluated. It should also be noted that frequent inter-GW handovers of a UE will lead to complete re-authentications. Neither IKEv2 with EAP-AKA, nor HIP DEX AKA solves this challenge because they cannot handle the change of the identity of the responder side. (i.e., lack of re-binding key material to new public/private key pair, lack of context transfer from previous gateway). This challenge is targeted by HIP-based Ultra Flat Architecture, described in Section 1.1.1, which could coexist with HIP DEX AKA.

Key performance indicators of the authentication process are:

- Process load on different network entities given in average number of CPU cycles per one authentication round.
- L3 authentication delay. This KPI is important in case of single interface devices, when a reactive HIP DEX AKA authentication is triggered during handover. HIP authentication delay together with L2 attachment and authentication delay are the main influencing factors of service interruption delay. Service interruption delay directly affects the QoE of streaming, real-time interactive services, but has also effect on the achievable TCP throughput, UDP packet loss in case of single interface devices. Multi-interface devices may exploit multihoming features to prevent service interruption during attachment of an interface.

2.5.1.2 Validation scenarios

This validation topic will compare the performance of the following L3 authentication methods. The compared technologies provide initial authentication service and negotiate IPsec security associations between the UE and the IP gateway of the operator.

- IKEv2 EAP-AKA: this is the standard method for user access authorization in non-3GPP access networks.
- IKEv2 EAP-TLS: can be used for certificate-based authentication of the user. This method has high network and computational costs, to be used for reference in the comparison. It is not used in 3GPP.
- IKEv2 PSK: provides pre-shared key based authentication. This method requires manual configuration on every peer, hence it cannot be used in large deployments. It will be evaluated in order to obtain reference values for the comparison. It represents the most lightweight traditional IKEv2 based method.
- HIP DEX AKA: it is a lightweight HIP method providing AKA-based user authentication in non-3GPP access. Currently it could be used in non-managed non-3GPP access where protection of data is needed on the path to the ePDG. In long-term timescale, if EPC is flattened, HIP DEX AKA could be used for uniform L3 user access authorization, in case of resource constrained devices.
- HIP DEX: lightweight, elliptic curve cryptography-based authentication method. It could be used to protect path between SEGs or only in very specific cases between the UEs and first IP GW of the operator. Its usage is restricted because some mechanism must guarantee for the participants the knowledge on which are the acceptable remote peer identities (public keys).
- HIP Base Exchange (BEX): it is taken into consideration for comparison, to show the gains of HIP DEX compared to traditional HIP. It could be used in cases where the peers preliminary know by other mechanisms the acceptable public keys/HIT tags of remote peers.
- HIP BEX with certificates (HIP BEX CERT): original HIP method using certificates, i.e., public keys are certified by trusted third-party Certificate Authority. Its counterpart configuration in IKEv2 is the certificate-based authentication. Both methods are appropriate for inter-network element authentication and authorization, or certificate-based user access authorization. However the Public Key Infrastructure has scalability problems, there is not only one root Certificate Authority but multiple in every country. The evaluation of this method has low priority, but could be relevant to see the processing and message overhead due to the addition of certificates.

The MEVICO project investigates the question of appropriate distribution of EPC functions. The topological distribution of the network elements has influence on the distance of the network elements. Furthermore, the distance of network elements influences the KPI “L3 authentication delay”. As a consequence, in order to help the decision on appropriate distribution level and to know the influence of the distribution level on the L3 authentication delay, we will compare the authentication methods in three reference scenarios. Each scenario reflects a fixed distribution case of the network in a homogenous, idealistic network in a big country. We are only interesting in maximum delay values, i.e., the worst case distances between network elements.

The following figures show the reference scenarios for centralized, distributed and flat network, respectively. The figures reflect functional and topological aspects at the same time.

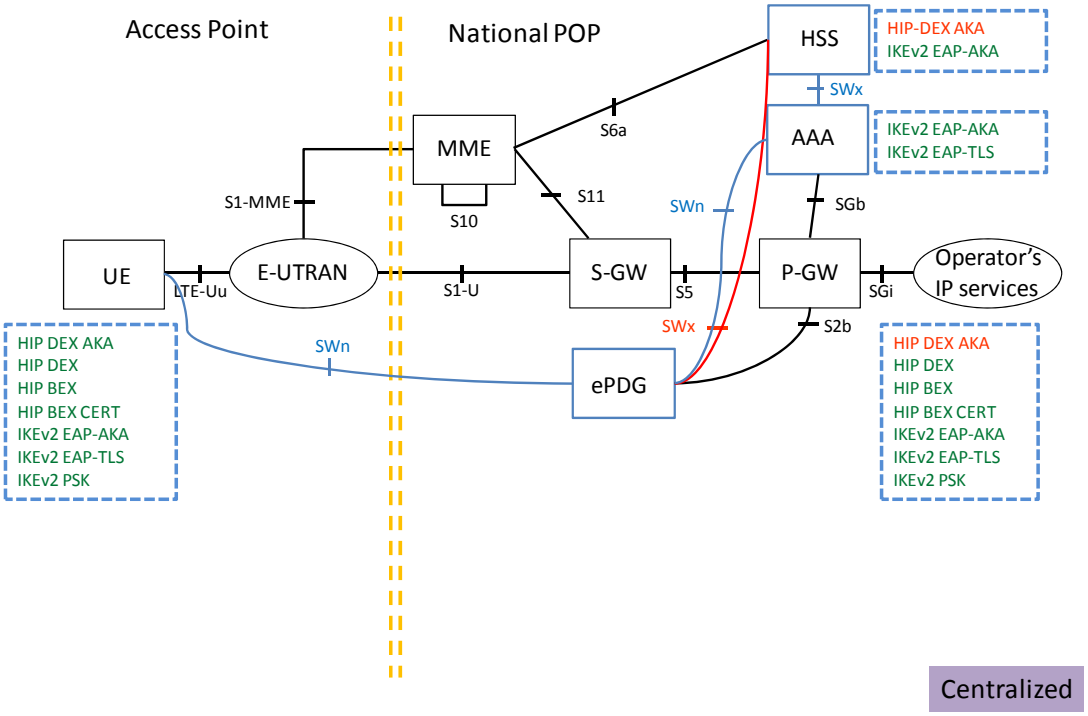


Figure 38 – Centralized scenario for the compared L3 authentication methods.

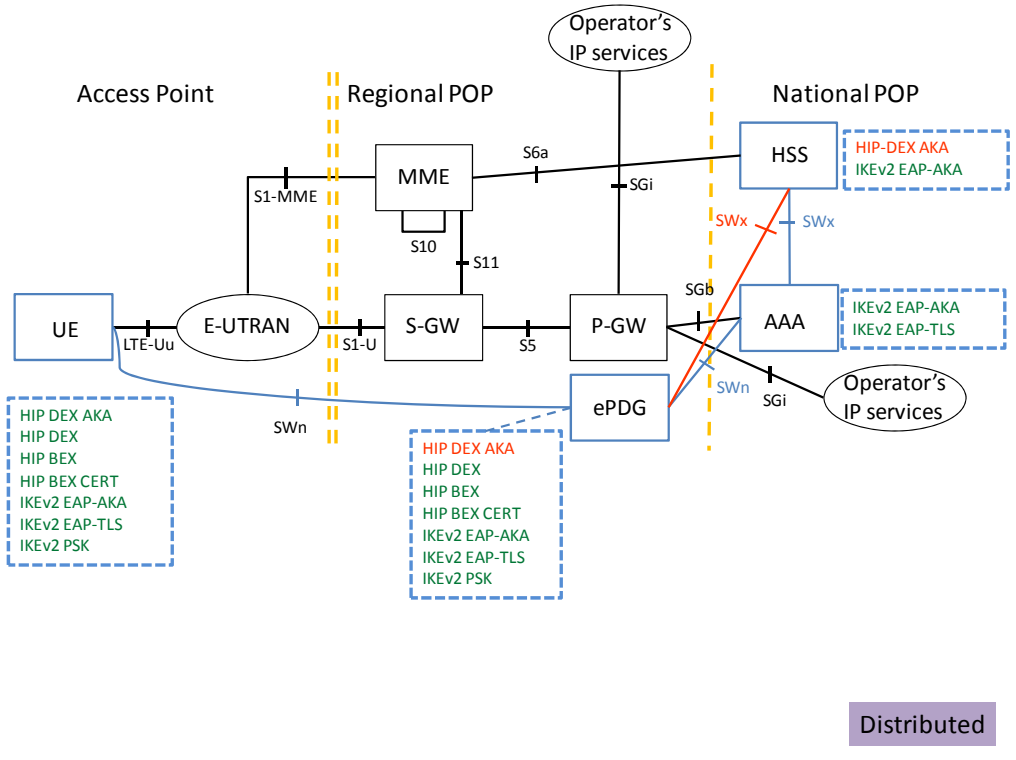


Figure 39 – Distributed scenario for the compared L3 autentication methods.

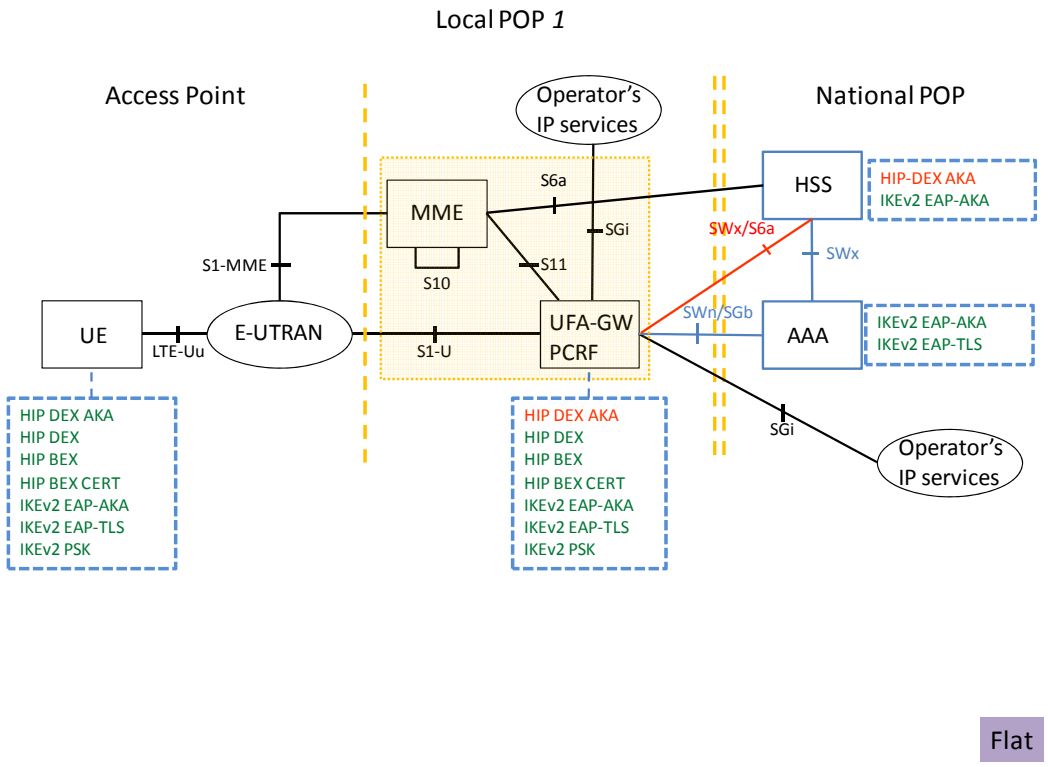


Figure 40 – Flat scenario for the compared L3 authentication methods.

The distribution level depends on whether the core functions are placed in the national, regional, or local Point of Presence (POP). HSS and AAA server will remain centralized in every case. The gateway (GW) that provides the first IP hop for the UE (ePDG, UFA GW) is pushed down to the proximity of the eNodeBs as the scenario becomes more and more distributed. The highest distribution level is represented by the flat scenario, where the GW is near the eNodeBs. In this case the backhaul network is reduced to 1-2 hops from the eNodeBs to the GW.

The network elements and connections in blue reflect the network functions and interfaces where the authentication methods have modules. The red lines and interfaces show that in case of HIP DEX AKA the GW (ePGD, UFA-GW) asks authentication quintuplets directly from the HSS using the SWx (or S6a) interface of the HSS. This requires a new interface in the GW. Note that in this case the AAA server functionality is integrated at the gateway side hence the proposed method may reduce the overall security setup overhead by the shortening of the message paths and reduction of computational overhead on the AAA server. The network interfaces in these scenarios are realized between remote network elements.

The L3 authentication delay will be evaluated in all the three scenarios using two different access networks

- Wi-Fi (IEEE 802.21g)
- HSDPA downlink and UMTS uplink. LTE would be more appropriate but is not available in the demonstrator where we measure the authentication delays.

The Wi-Fi case is relevant in all scenarios, while the UMTS case is only relevant in the flat scenario. In the non-flat scenarios UMTS (or LTE access) uses standard user access authorization to the 3/4G network, i.e., the MME is the authentication server and gets quintuplets from the HSS through the S6a interface for the authentication of the UE. The eNodeB provides the network-side endpoint of the security association. L3 access authentication with the analyzed methods would only cause duplicated authentication of the UE (on different layers) and over-protection on the radio link. In these cases, L3 protection could only be useful, if the path between the eNodeB and the serving PGW was untrusted and not protected by other means. However this scenario has low relevance, because security between different sites of the network provider is often provided using network security services (by IPsec tunnels between SEGs).

The CPU utilization of the methods will be evaluated only in case of the flat scenario with Wi-Fi access, because the network delays are not expected to influence the average processing cost in the UE and the highlighted network elements. The HSS is a legacy Huawei equipment, hence we cannot measure the CPU utilization of the methods there. Both IKEv2 EAP-AKA and HIP DEX AKA call the same commands on the HSS when they ask authentication quintuplets through the Cx interface. However, in our configuration, in case of IKEv2 EAP-AKA the AAA server asks three quintuplets at once, and the HIP DEX AKA prototype asks one quintuplet per authentication. Asking multiple authentication vectors at once is currently not part of the GW functionality in case of HIP DEX AKA. There is a trade-off, i.e., how many authentication vectors to ask from the HSS: in the flat network scenario, a GW may require only one authentication vector because the MN may move to another GW before re-authentication, however in distributed and centralized scenario requesting multiple authentication vectors at the same time might reduce security overhead of HIP DEX AKA. As a consequence, in this validation, a sub-optimal usage of HIP DEX AKA will be measured in case of distributed and centralized GW.

This means that the HIP DEX AKA utilizes three times more the HSS currently.

CPU utilization is more relevant on the UE, the GW, and the AAA server. Only the IKEv2 EAP methods use the AAA server. The CPU utilization on the UE is important due to its high correlation with the energy consumption. The CPU utilization of the GW and the AAA by the different methods is relevant for network dimensioning purposes.

2.5.1.3 Validation tools

Two real-life test-beds are used for this validation and have been described in D2.1 Section 6.1.1 and 6.1.2. The first test-bed is mainly used to obtain the KPIs of this validation. The second test-bed is used for feasibility study of HIP DEX and HIP DEX AKA-based authentication and bootstrapping.

Reference scenarios are realized partially by emulation because evaluation in large network is not available. Distances can be emulated by adding one-way network delays with a traffic emulator. In this validation, the netem [1] traffic emulator is applied in the first test-bed.

After having discussions with operators, average reference values have been estimated for the one-way delays between the network elements. These values could be used by any validation requiring reference values for these average network delays.

Table 4 presents the reference parameters for the three scenarios.

Table 4 – Additional network emulation delays for the three distribution scenarios

	centralized	distributed	flat
UE - P-GW/ePDG/SGW	30ms	15 ms	10 ms
GW - AAA	5 ms	15 ms	30 ms
AAA - HSS	5 ms	5 ms	5 ms
GW - HSS	30 ms	15 ms	30 ms

2.5.1.4 Expected results

We expect significant reduction of the CPU utilization cost by light-weightened methods, moreover reduction of the authentication delay due to less message exchange. The exact proportions of the KPIs in case of the compared methods will be seen from the results.

2.5.1.5 Results

The results of the performance measurements are presented in the following sections. The evaluated performance indicators are computational cost, memory cost, authentication delay and message complexity of one authentication flow.

2.5.1.5.1 Computational cost

Figure 41 presents the mean and variance of the CPU cost of one re-authentication flow in case of each authentication method on the UE, the GW and the AAA server.

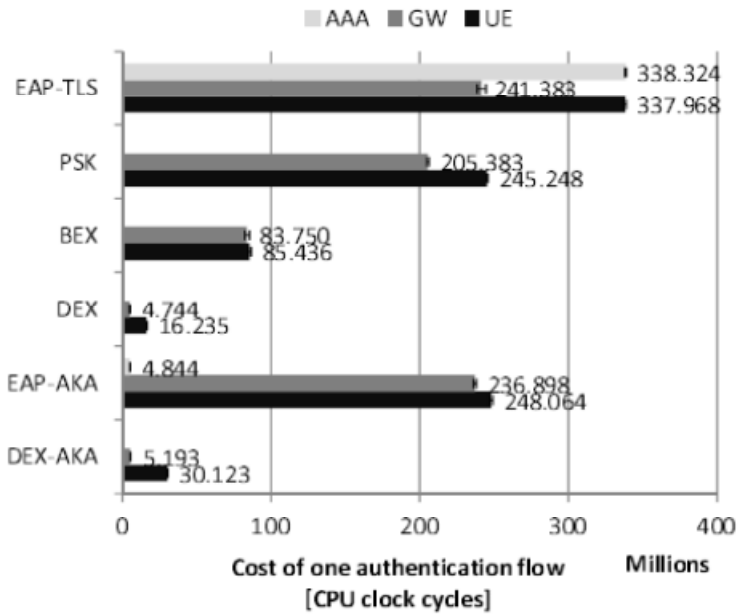


Figure 41 – Cost of one authentication flow in term of number of CPU clock cycles.

The result show that the reduction in the amount of non idle CPU intervals occupied by the HIP DEX-AKA authentication process is significant compared to IKEv2 EAP-AKA in relative terms. On the UE 12%, on the GW 2% is the proportion of the computational cost of HIP DEX-AKA versus IKEv2 EAP-AKA. The results for the other authentication methods are presented as well by the figure. DEX proves to be the less demanding method, however it is less appealing for mobile operator based environment where the USIM based subscriber authentication is already in place.

In absolute terms, the frequency and cost of re-authentications is so low, that no significant influence is on the battery consumption of the UE. CPU typically consumes less than 10% of the total energy consumption. The frequency of re-authentications can be estimated as the sum of the intensity of GW change due to mobility and the frequency of lifetime expiration. In order to reduce the frequency of complete re-authentications, support of GW change should be added to these authentication methods.

The results in CPU clock cycles enable the derivation of CPU utilization time for CPUs having similar architecture but running on different frequencies than the ones used in our measurements. This might be relevant for example in case of the UE, which supports 600MHz and 1.6GHz CPU frequencies. The average CPU utilization time by one authentication flow is illustrated in Figure 42 for the UE, GW and AAA server.

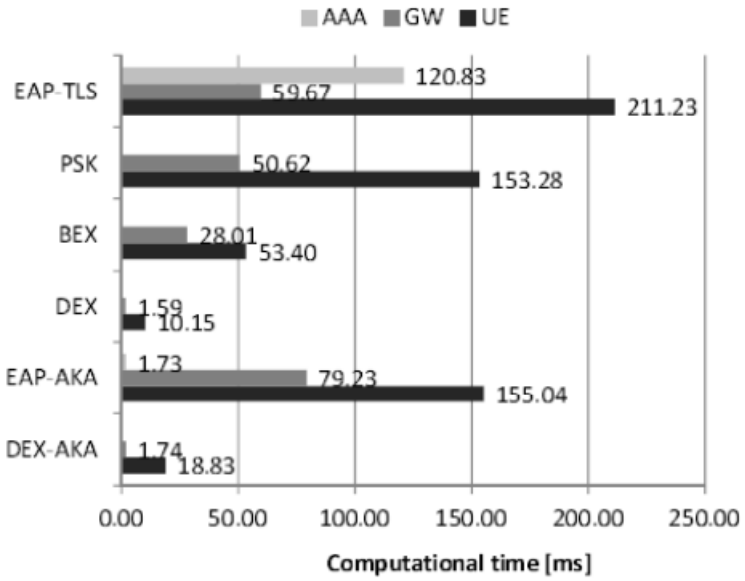


Figure 42 – Computational times of authentication methods.

Table 5 shows the proportion of computational time compared to the overall authentication times described in Section 2.5.1.5.3 in case of the different reference scenarios.

Table 5 – Proportion of computational times in the authentication delay.

	centralized	distributed	flat
WiFi access			
DEX-AKA	7.1%	9.3%	9.4%
EAP-AKA	40.6%	50.4%	49.3%
DEX	8.6%	15.2%	20.3%
BEX	39.8%	56.0%	65.7%
PSK	70.5%	89.6%	97.1%
EAP-TLS	39.4%	45.2%	39.8%
HSDPA/UMTS access			
DEX-AKA	3.3%	3.7%	3.7%
EAP-AKA	18.5%	20.3%	20.0%
DEX	3.2%	3.9%	4.1%
BEX	14.3%	16.1%	17.0%
PSK	33.3%	37.5%	38.8%
EAP-TLS	16.3%	17.3%	16.4%

It can be seen from the table that HIP DEX-AKA and HIP DEX reduce the computational time part significantly compared to the other methods.

2.5.1.5.2 Memory cost

The heap and stack memory size occupied by the different methods have been evaluated. The profiles include the initialization of participating daemons and one successful authentication.

Figure 43 shows the mean of the peak heap memory sizes allocated during the initialization of software and one authentication flow. The variance of the peak values was zero or near zero.

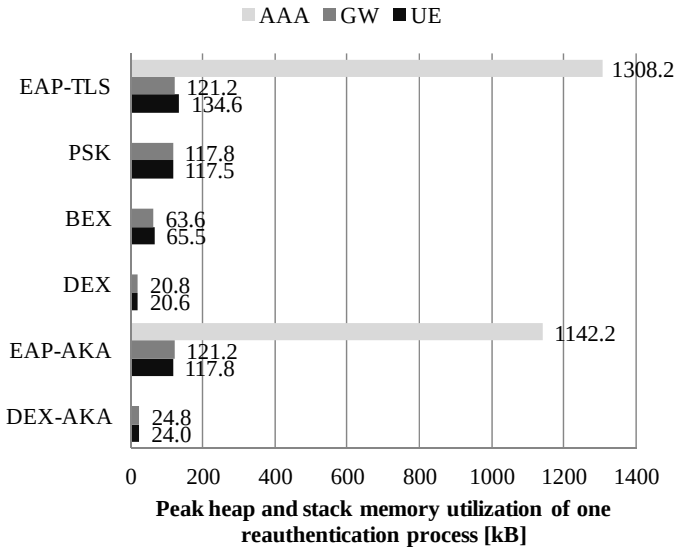


Figure 43 – Memory utilization of the authentication methods.

The results show, that reduction of memory utilization of HIP DEX-AKA, and HIP DEX is significant compared to the other authentication methods. Comparing HIP DEX-AKA with IKEv2 EAP-AKA, HIP DEX-AKA provides 80% gain on the UE and the GW. The AAA server is not utilized by HIP DEX-AKA.

2.5.1.5.3 Authentication delay

Figure 44 and Figure 45 show the authentication delays of the methods measured by the tshark tool, in case of Wi-Fi (802.11g) and HSPA access, respectively, for the centralized, distributed and flat network scenario.

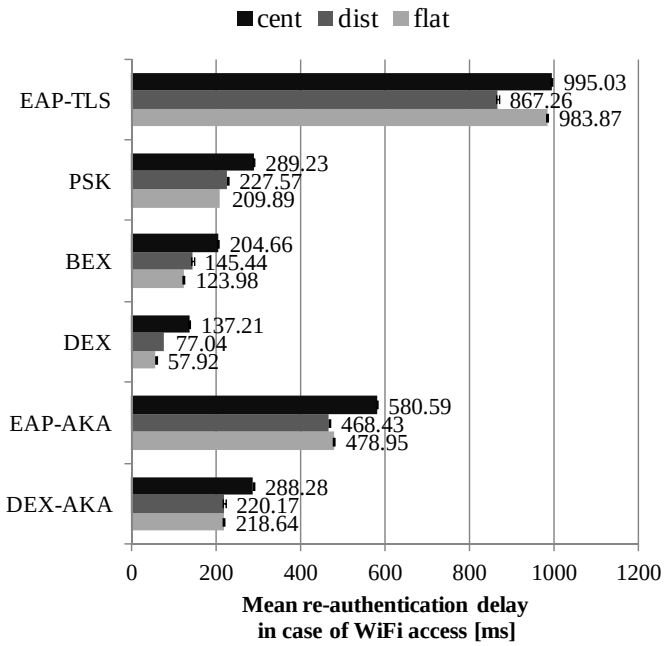


Figure 44 – Mean authentication delay in different distribution cases of the network, in case of Wi-Fi access.

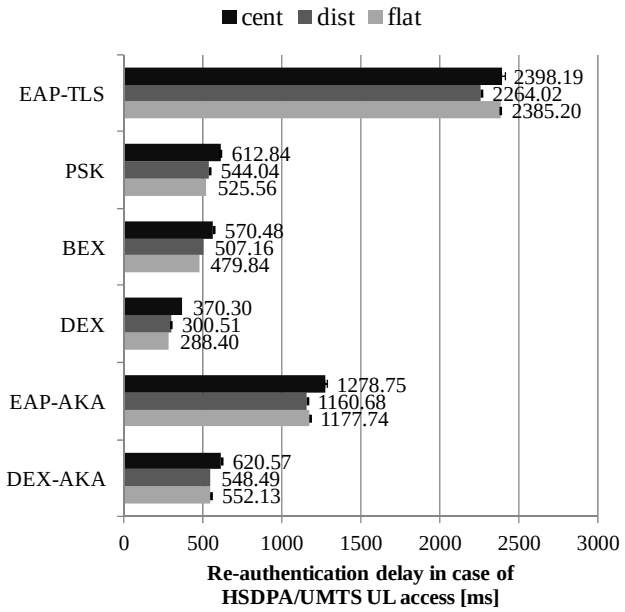


Figure 45 - Mean authentication delay in different distribution cases of the network, in case of HSDPA/UMTS UL access.

The additional emulated latencies between the network elements reflect very long paths in the network, hence the results represent re-authentication delays in a considerably large-sized mobile operator network.

Real-time service constraints are different for different service types. 3GPP TS 23.203 specifies the following packet delay budgets between the UE and the Policy Control Enforcement Function (PCEF) for Guaranteed Bitrate Services: 50ms for real-time gaming, 100ms for conversational voice, 150ms for live (interactive) video streaming and 300ms for buffered video streaming. PCEF is located in the Packet Data Network Gateway which contains the Gi interface via the Internet.

The results show that HIP DEX-AKA highly improves the authentication delay compared to IKEv2 EAP-AKA. Regarding the packet delay budget, only in case of the Wi-Fi access it results authentication delays between 150 and 300 ms. I.e. HIP DEX-AKA could be used in case of buffered video streaming, if the overall delay is less than 300 ms together with other delay factors. Note that no one of the evaluated authentication methods was designed with fast re-authentication in focus in case of break-before-make handovers.

It also can be stated that under the conditions of the reference scenarios, for the authentication methods applicable in large environment, i.e., for IKEv2 EAP-AKA, IKEv2 EAP-TLS and HIP-DEX AKA, the scenario causing the lowest delay was the distributed scenario. However, in case of IKEv2 EAP-AKA and HIP DEX-AKA, the flat scenario also results very close results.

2.5.1.5.4 Message complexity

In the case of the investigated authentication methods, the control messages appear in two different parts of the network. A message is either relayed through the RAN and backhaul between the UE and GW, or through the aggregation and core network among the GW, the AAA server and the HSS.

Figure 46 shows the average number of signaling messages (i.e. IP datagrams) in the RAN and backhaul network, and in the aggregation and core network. The values were got from the Wireshark traces of the authentication methods. In case of IKEv2 EAP-AKA, the assumption was that the AAA server returns to the HSS for authentication vectors after every third authentication on average

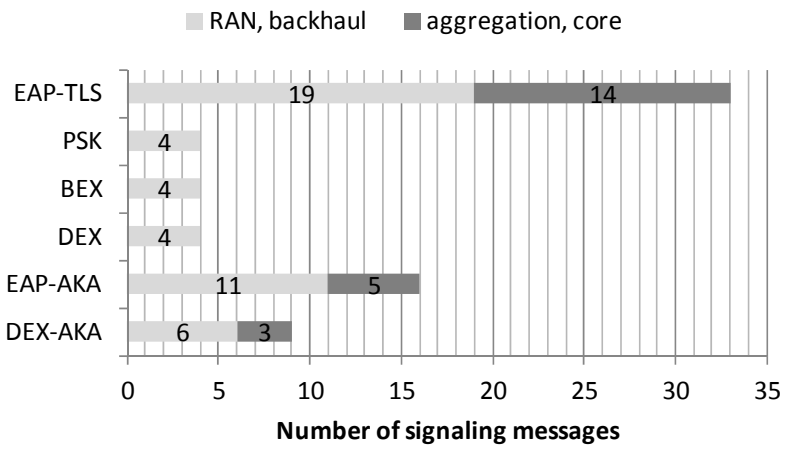


Figure 46 – Number of signaling messages for one authentication flow.

Figure 47 presents the volume of messages in Bytes transfered through the two network parts.

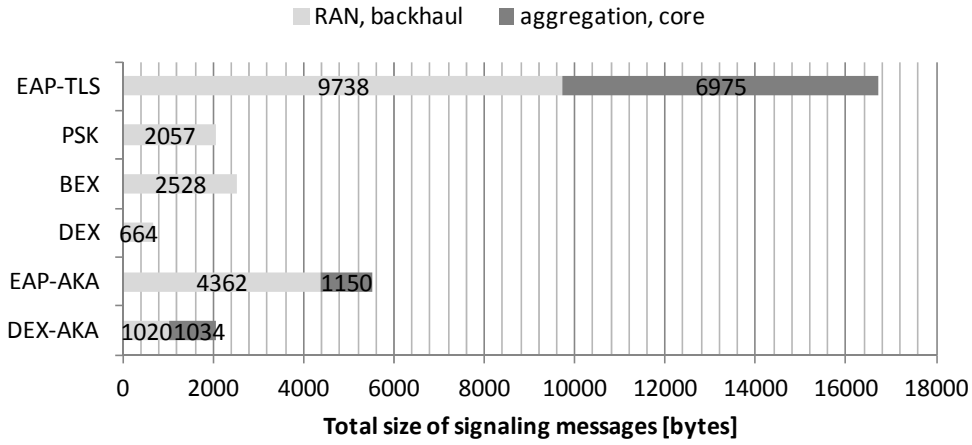


Figure 47 – Total size of signaling messages for one authentication flow.

Both in terms of number of messages and in terms of the total size of the messages at different parts of the network, HIP DEX-AKA significantly outperforms IKEv2 EAP-AKA. The ratio of the number and total size of messages is 56% and 37%, respectively, between HIP DEX-AKA and IKEv2 EAP-AKA.

Another important aspect is the number of control messages charging the aggregation and core network. HIP DEX-AKA requires on average two messages less per re-authentication than IKEv2 EAP-AKA.

2.5.1.6 Applicability of the results

As previously mentioned, currently HIP DEX AKA could replace IKEv2 with EAP-AKA used in non-managed non-3GPP access networks, where the network part between the radio access point and the ePDG is untrusted, and the operator may not have control on the L2 security within the access network.

In long-term, if one distributed network entity provided IP connectivity and access to the operator's value-added IP services (e.g., IMS services), then a uniform L3 access authentication and authorization could exchange the current network access authorization (AKA) and IMS access authorization (e.g., the IMS AKA).

In a totally flat network protecting the path between the eBodeB and the UE on L2 with standard AKA and encryption could be enough. In that case running a separate L3 access authorization must be justified. It depends on the trust model, whether L3 authorization is required, e.g., by the IMS domain. Currently, IMS requires L3 authorization of the user and protection of the signalling path until the P-CSCF.

A possible approach to reduce security overhead due to L2 and L3 access authorization is to reuse the L2 security context and generate cryptographically independent key material for HIP and IPsec. This cross-layer authorization approach was proposed in [20]. However this approach would lead further from this validation topic, and requires a variant of HIP which e.g., does not use DH key exchange but generates keying material from L2 security contexts.

2.5.1.7 Partners involved

CWC has developed HIP DEX AKA and HIP DEX prototype. BME MIK performs measurements on its testbed, in orther to get KPIs in different scenarios. Common publication is expected from this work.

2.5.2 Suitability analysis of different L3 authentication methods for the MEVICO architecture and requirements

2.5.2.1 Objectives

The objective of this validation is to compare the authentication methods described in the previous section in a broader aspect. The methods will be ranked not only based on their performance related KPIs but other criteria as well.

Figure 48 depicts the criteria of the evaluation. Criteria are organized in a hierarchic way: there are main criteria, such as security, performance, deployment and functionalities, and sub-criteria, such as support mutual authentication under the security criterion.

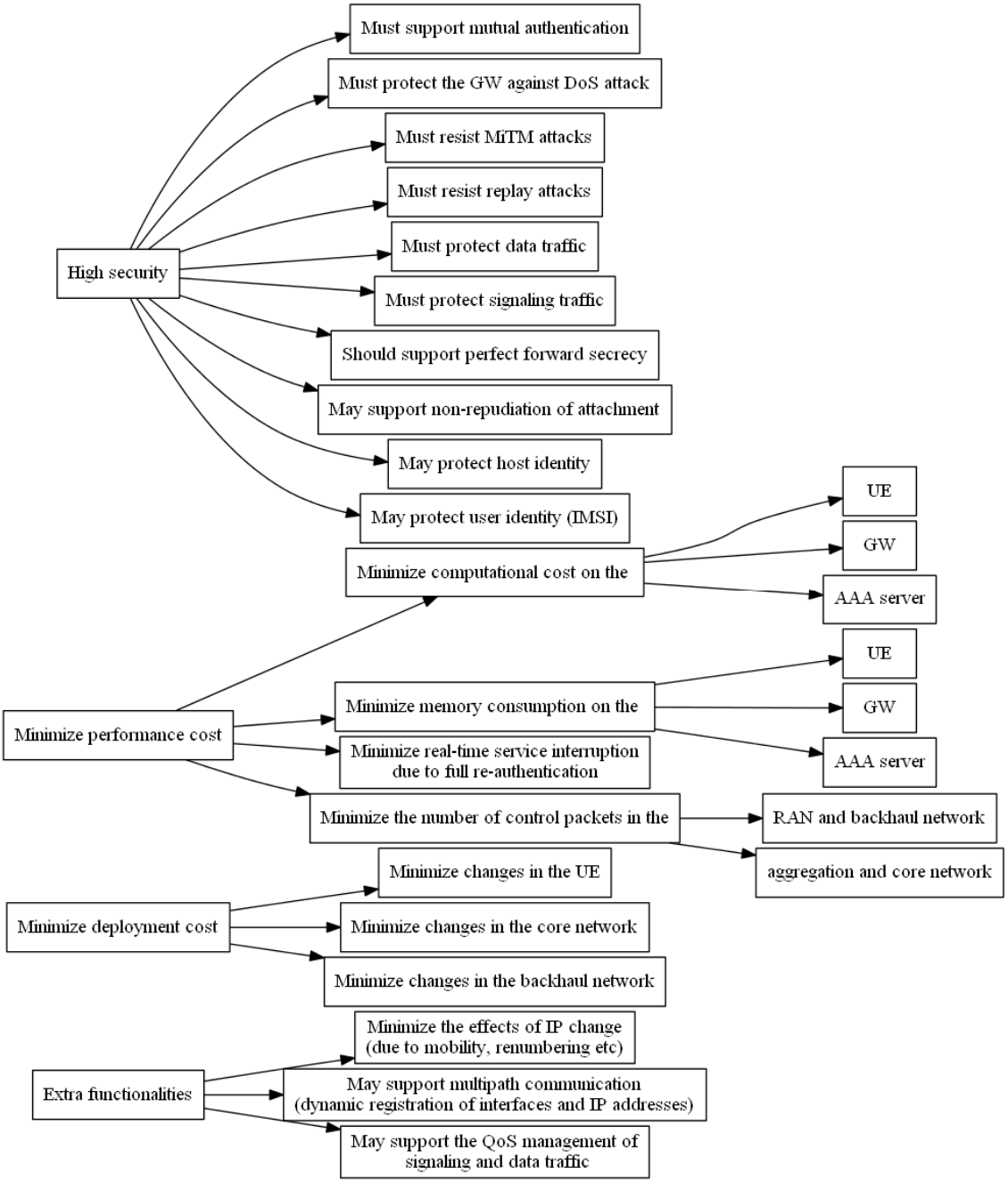


Figure 48 – Criteria tree.

2.5.2.2 Validation scenarios

The reference scenarios are the same as described in Section 2.5.1.2, i.e., centralized, distributed and flat. Note that this only influences the real-time service interruption time KPI.

2.5.2.3 Validation tools

The Multiplicative Analytic Hierarchy Process (MAHP), a multi-criteria or muti-attribute decision making method is applied as validation method. The decision making process contains the following steps:

- Define criteria weights
- Define criteria metrics for the sub-criteria
- Define mapping from metrics to grades, including normalization of the metrics and also reflecting the real influence of a metric value on the decision making.
- Obtain KPIs for every sub-criterion in every reference scenario.
- Apply a decision making tool, i.e., a score aggregation method, to get the terminal score for every authentication method in every scenario.

Several score aggregation methods exist, such as the Multiplicative Analytic Hierarchy Process [21] or the most widespread Simple Weighted Sum method. Both of them require normalized KPI metrics and the criteria weights as input and are applied to calculate the scores of the compared alternatives under different criteria. All score aggregation methods apply different approaches to calculate the terminal score.

Multiplicative Analytic Hierarchy Process, similarly to its ancestor – the Analytic Hierarchy Process [22] – decomposes the complex matching problem of evaluating multiple alternatives under multiple criteria to many pair-wise comparisons of the alternatives per each criterion. The relative goodness of the alternatives is reflected by an exponential function of the difference of the performance grades of the alternatives.

Grades are obtained by mapping the performance metrics on a geometric scale (i.e., a stepwise logarithmic function), as depicted in Figure 49.

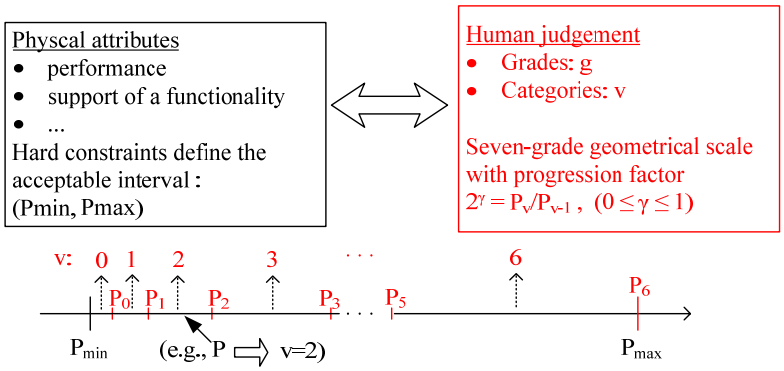


Figure 49 – Human judgment of KPIs approximated by a geometrical scale.

This scale approximates the behaviour of human judgment, and enables to neglect small differences of KPI values that basically represent the same level of satisfaction for the decision maker. On the other hand similarly behaving alternatives at the proximity of steps may be judged very differently. This problem can be prevented by manual adjustment of grades.

In general, it is not mandatory to use the geometric scale to get the performance grades from performance metrics. Any mapping function could be used, the important issue is that grades should reflect the satisfaction levels of the decision maker.

2.5.2.4 Expected results

This evaluation is expected to show the real benefits and disadvantages of selecting a given authentication method in a given scenario.

Even with simplified criteria metrics, the evaluation requires to investigate in a systematic way the criteria, the alternatives, hence we can discover the deficiencies of the methods during the analysis.

Criteria weights and network parameters highly influence that which signalling scheme should be considered as best alternative. That's why criteria weights must be fixed in the beginning of the decision making process based on a common agreement of decision makers. Note that a later analysis may focus on the robustness of the decision, i.e., it will be possible to analyze the sensitivity of the ranking of alternatives under perturbed criteria weights. Regarding the KPIs under performance criterion, particularly the real-time service-interruption delay, the network parameters of the reference scenarios influence the obtained grades.

2.5.2.5 Results

2.5.2.5.1 Definition of key performance indicators and performance grades of the methods

Security features of the methods

The security features of the methods are evaluated under each security criterion. Typically the metric used is the support (1) or non-support (0) of a feature, except in some cases where finer-grade scales have been used to express different levels of support of a feature. The higher is the grade the better is the method under a given criterion.

Table 6 summarizes the grades of the methods obtained under each security criterion, and the explanation of the grades.

Table 6 – Grades of the auhentication methods under security criteria.

	DEX-AKA	EAP-AKA	DEX	BEX	PSK	EAP-TLS	Meaning of grades
Must support mutual authentication	3	3	2	2	1	3	3: strong mutual authentication based on certificates or AKA 2: strong authentication of self-certifying identities, but lack of HIT verification 1: weak preshared key based authentication
Must protect the GW against DoS attack	1	2	1	1	2	2	2: optional cookie-based DoS protection in IKEv2, GW controls the cookie distribution 1: puzzle-challenge based DoS protection. The attacker could have high computational capacities, hence it is hard to set the good level of puzzle. GW has less control on the access authorization than in case of cookie-distribution
Must resist MiTM attacks	1	1	1	1	1	1	1: Authentication of entities, crypthoraphic binding of key material, symmetric-key based signature protection on control messages prevents MiTM attacks.
Must resist replay attacks	1	2	1	2	2	2	1: weaker resistance to replay attacks in case of DEX than in case of BEX or IKEv2, the initiator (UE) does not contribute to the freshness of the messages and keys, hence replaying R1/R2 messages can lead to HIP/Ipsec state establishment in the initiator. If the initiator added a nonce to the communication and key derivation, e.g., in I2, then this attack type could be mitigated. 2: both peers contribute to the freshness of the communication by using random nonces from an enough large interval.
Must protect data traffic	1	1	1	1	1	1	1: all methods negotiate IPsec transport, containing encryption, integrity protection, message origin authentication, anti-replay protection
Must protect control traffic	1	1	1	1	1	1	1: all methods provide confidentiality, integrity protection, message origin authenticity
Should support perfect forward secrecy	0	1	0	1	1	1	0: perfect forward secrecy is not provided in case of DEX, because it uses static DH key generation. If a long-term secret, such as the private key or the static DH secret established with a given peer, is compromised, previously captured confidential information can be revealed by the attacker. 1: perfect forward secrecy is guaranteed, due to ephemeral Diffie-Hellmann key exchange, i.e., always different DH key is negotiated.
May support non-repudiation of attachment	0	0	0	1	0	1	0: digital signature of the entities is not contained in the control message sequence. 1: digital signature of the entities is provided (based on TLS certificates in IKEv2 with EAP-TLS, or private key of HIP host identities in HIP BEX) Note: non-repudiation would also need secure logging of control messages with secure time-stamps. This is not part of the standards.
May protect host identity	0	1	0	1	1	1	1: host identity is sent encrypted, or in case of BEX, using the BLIND privacy extension 0: host identity is sent in plaintext
May protect user identity (IMSI)	1	1	0	1	0	1	1: user identity (IMSI) is sent in an encrypted block (in EAP or HIP control messages). 0: user identity is sent in plaintext

Performance features of the methods

Performance features of the compared authentication methods are described in Section 2.5.1.5. The decision making process requires the assignment of grades to the performance metrics under each performance criterion. Two different approaches have been applied in the performance grade assignment for the criteria.

1. If there are no objective requirements in existing standards for a given performance metric, satisfaction of decision maker is relative to the amelioration compared to the IKEv2 EAP-AKA method on the given network entity. IKEv2 EAP-AKA method is part of current 3GPP standard, hence its performance is chosen as reference. The grade assignment follows a geometric scale with progression factor 2, i.e., if a method provides 50% gain, then its grade increases by one. If a method performs 50%*50%=25% of IKEv2 EAP-AKA, then it obtains two grades better than IKEv2 EAP-AKA.
2. If there are objective requirements for performance values in the 3GPP standard then grade assignment function follows the standardized values. This is the case for the authentication delay metric, where the packet delay budgets for real-time applications, i.e., 50, 100, 150, 300 ms, have been taken from the 3GPP standard as the borders of the steps in the grade assignment.

Computational cost

Performance grade assignment functions are depicted in Figure 50, Figure 51 and Figure 52, for the computational costs on the UE, GW, and AAA server, respectively. Grade assignment follows the first approach, i.e., the IKEv2 EAP-AKA method is selected as the reference method. The performance of IKEv2 EAP-AKA is signaled using the circle symbol. The performance of other methods is measured relative to IKEv2 EAP-AKA, following a geometrical scale with progression factor 2.

Note: the absolute value of the grades is not relevant for the score aggregation method applied later. The differences between the obtained grades of the methods under a given criterion are relevant for the final decisions.

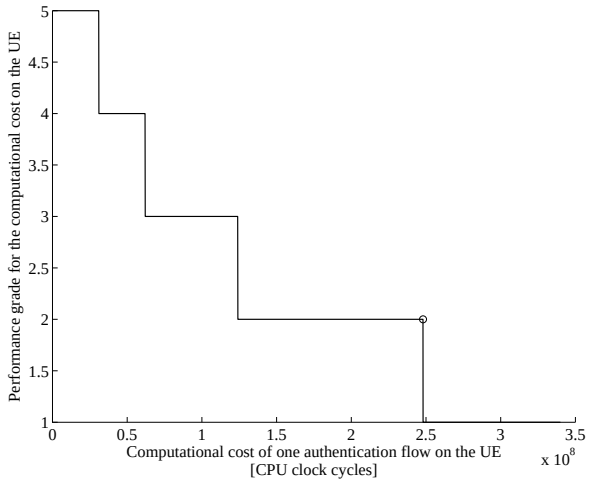


Figure 50 – Performance grade assignment function for the computational cost on the UE.

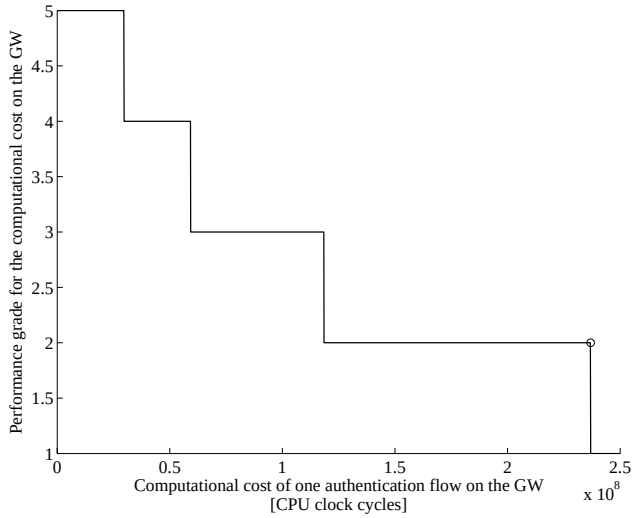


Figure 51 – Performance grade assignment function for the computational cost on the GW.

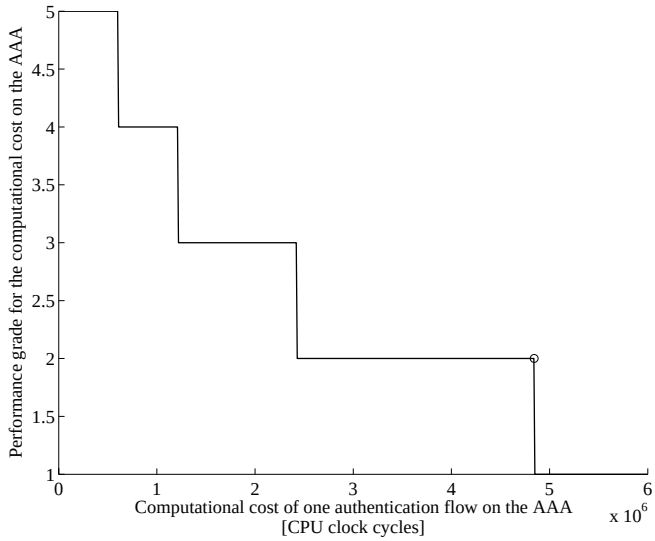


Figure 52 - – Performance grade assignment function for the computational cost on the AAA server.

Memory cost

Performance grade assignment functions for the memory requirements of the authentication methods are presented in Figure 53, Figure 54, and Figure 55, for the UE, GW and AAA server, respectively. The methods which do not use the AAA server, i.e., DEX-AKA, DEX, BEX, and PSK consume 0 kB heap and stack memory on the AAA server, hence they get the highest grade, i.e., 5. The memory requirement of IKEv2 EAP-AKA method is taken as reference value, and it is represented by a circle symbol in the plots.

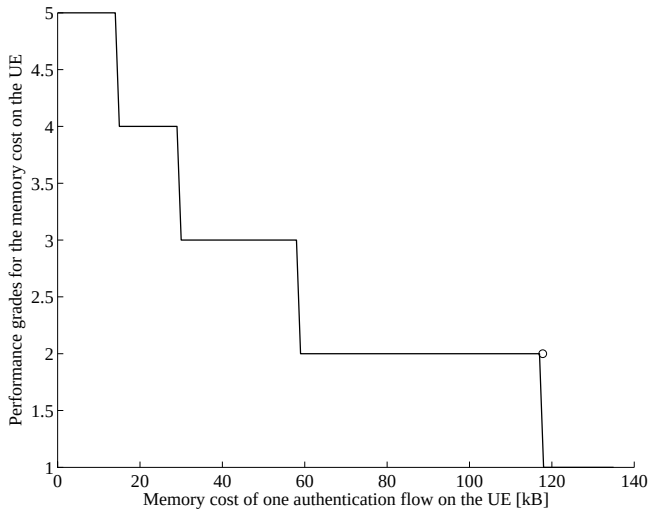


Figure 53 - – Performance grade assignment function for the memory cost on the UE.

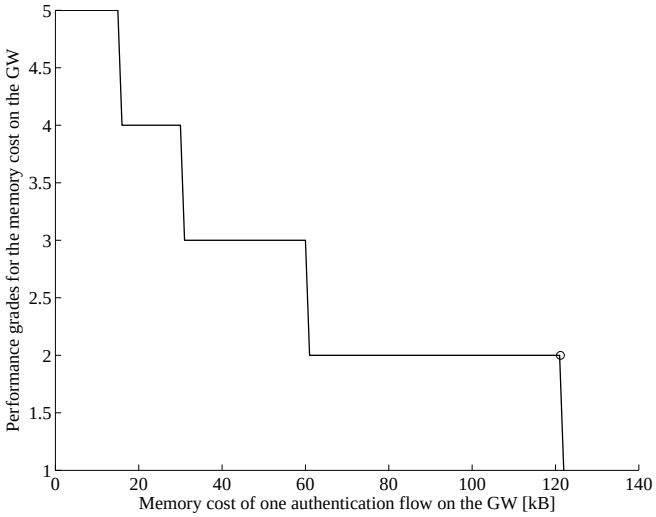


Figure 54 - – Performance grade assignment function for the memory cost on the GW.

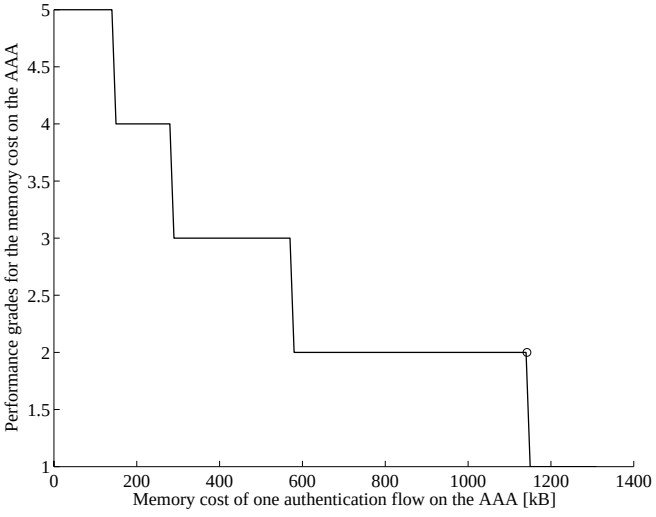


Figure 55 - – Performance grade assignment function for the memory cost on the AAA server.

Real-time service interruption delay

The performance grade assignment function for the real-time service interruption delay is depicted in Figure 56. This criterion is relevant only for real-time applications, and assigns grades based on the packet delay requirements of different application types. The grades represent the support or non-support of re-authentication time below a certain constraint value. TS 23.203 defines the following packet delay budgets for different GBR service types:

- 50ms: real-time gaming. If the authentication delay is lower than 50ms, then the method gets grade 4.
- 100ms: conversational voice. If the authentication delay of the method is between 50 ms and 100 ms, then the method gets grade 3.
- 150ms: Live (interactive) video streaming. If the authentication delay of the method is between 100 ms and 150 ms, the method gets grade 2.
- 300ms: Buffered video streaming. If the authentication delay is between 150 ms and 300 ms, the method obtains grade 1.
- For higher than 300 ms authentication delay, it is impossible to guarantee seamless service continuity for real-time GBR services. Hence for these application types, and in case of requiring seamless service continuity, authentication methods performing in this category should be rejected. Hence grade 0 means non support of real-time service continuity due to the high service interruption time.

Note: packet delay budget is the delay defined between the UE and the Policy and Charging Enforcement Function (PCEF) for Guaranteed Bit Rate (GBR) services.

Note: The grade assignment function could be further optimized since packet delay budget should cover the complete re-attachment and IP mobility management procedure for ongoing real-time sessions, and L3 authentication is only part of that. However, the current grade assignment function is considered good enough to differentiate between the authentication methods.

Note: the authentication delay of the methods depend on the topology of the network, hence the grades obtained in the different reference scenarios may be different.

Note: The grades obtained for real-time service interruption delay in different reference scenarios shall be used as indicator of the preference for each reference scenario.

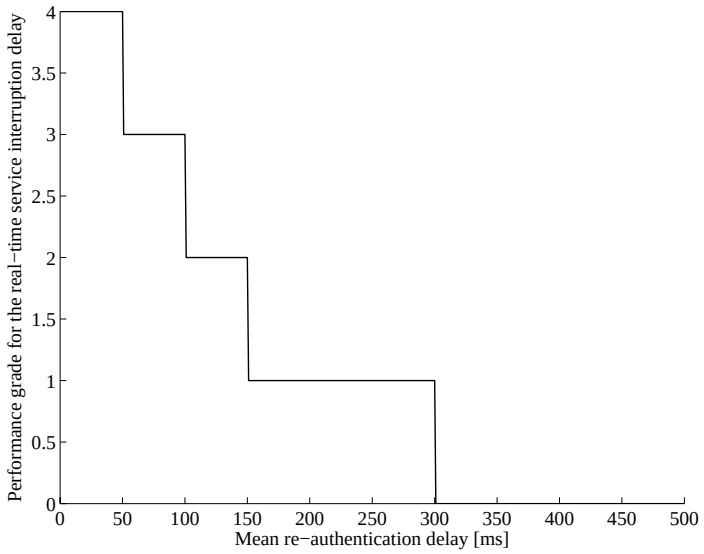


Figure 56 – Performance grade assignment function for the authentication delay of the methods

Message complexity

Performance grade assignment functions for the message complexity of the methods in different parts of the architecture are specified in Figure 57 and Figure 58. The performance of IKEv2 EAP-AKA is signaled by a circle symbol.

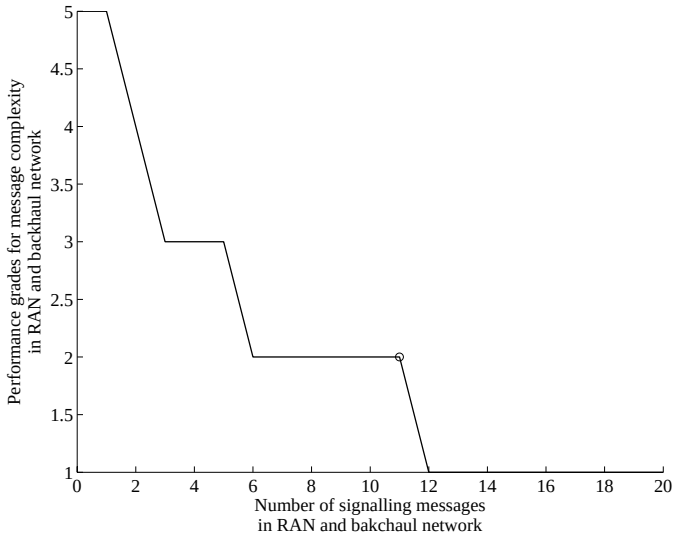


Figure 57 – Performance grade assignment function for the message complexity of the methods on the RAN and backhaul network.

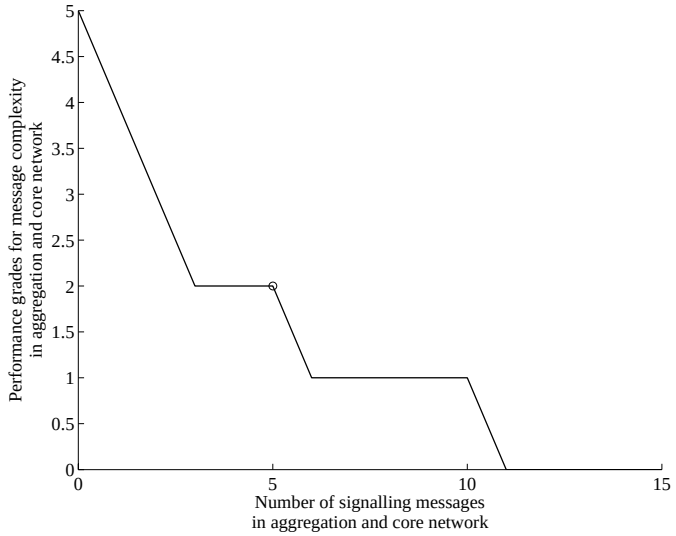


Figure 58 - Performance grade assignment function for the message complexity of the methods on the aggregation and core network.

Deployment features of the methods

The deployment costs of the authentication methods are measured by the required number of additional functional modules in different parts of the network, and the additional efforts required for configuring the methods in a large-scale network. The assumption is that the architecture supports IKEv2 EAP-AKA, hence the configuration and deployment costs of IKEv2 EAP-AKA are considered to be 0. Table 7 describes the deployment features of the authentication methods. The obtained values for required features could represent a possible lower-the-better grade assignment.

The grade assignment has been chosen very simple. All the positive values mean 0 grade, while 0 values mean 1 grade, so higher grade is better. The grade 0 means that the deployment of a method requires additional implementation of modules and/or configuration and management of the authentication methods. Hence we could avoid dealing with finer-grade deployment and management cost analysis in the decision process. At this stage, there are too many options and too much uncertainties for deployment cost analysis.

Table 7 - Deployment features of the methods

	DEX-AKA	EAP-AKA	DEX	BEX	PSK	EAP-TLS
Deployment requirements in the UE	1 - HIP	0	1- HIP	0	0	1 - TLS module
Deployment requirements in the GW	2 - HIP,AAA	0	1 - HIP	1 -HIP	0	0
Deployment requirements in the core network	2- HIPDNS,RVS	0	2- HIPDNS,RVS	2- HIPDNS,RVS	0	2 - TLS module, certificate management
Configuration requirements in the UE	0	0	1 - ACL with IDs of authorized GW's	1 - ACL with IDs of authorized GW's	1 - key management	0
Configuration requirements in the GW	0	0	1 - ACLs with IDs of authorized GW's	1 - ACL with IDs of authorized GW's	1 - key management	0
Configuration requirements in core network elements	0	0	0	0	0	0

Additional functionalities of the methods

The criterion on extra functionalities contains requirements for

- IP-mobility management,
- features that are necessary for multipath communication,
- QoS enforcement of the conveyed service data flows

IP mobility management means here the case when the IP address of one of the end-nodes of the IPsec security association changes for some reason (e.g., UE is moving to a new IP domain, renumbering of the network by the operator). The mobility management is expected to avoid complete re-association. This is solved by the security control protocol by the transfer of the new address to the peers and dynamic update of the security associations. All methods support this type of IP mobility management.

All methods support dynamic registration of interfaces and locators. Furthermore HIP has an extension called mHIP [30], for the provision of multipath communication.

QoS policy control is not part of the analyzed security control protocols, but should be solved by PRCF functions, and enforced by radio access network and transport network layer.

Table 8 summarizes the grades obtained for extra-functionalities of the methods.

Table 8 – Additional functionalities of the methods

	DEX-AKA	EAP-AKA	DEX	BEX	PSK	EAP-TLS
Minimize the effects of IP change (due to mobility or renumbering)	1 - supports using HIP mobility service	1 - supports using MOBIKE mobilty service	1 - supports using HIP mobility service	1 - supports using HIP mobility service	1 - supports using MOBIKE mobilty service	1 - supports using MOBIKE mobilty service
May support multipath communication (i.e., dynamic registration of locators, multipath feature)	2 - dynamic registration of interfaces and IP addresses supported, mHIP extension provides multipath communication	1 - dynamic registration of interfaces and IP addresses supported	2 - dynamic registration of interfaces and IP addresses supported, mHIP extension provides multipath communication	2 - dynamic registration of interfaces and IP addresses supported, mHIP extension provides multipath communication	1 - dynamic registration of interfaces and IP addresses supported	1 - dynamic registration of interfaces and IP addresses supported
May support E2E QoS management for data and signaling traffic	0 - not supported within the method, QoS policy control and enforcement by PCRF and transport network layer	0 - not supported within the method, QoS policy control and enforcement by PCRF and transport network layer	0 - not supported within the method, QoS policy control and enforcement by PCRF and transport network layer	0 - not supported within the method, QoS policy control and enforcement by PCRF and transport network layer	0 - not supported within the method, QoS policy control and enforcement by PCRF and transport network layer	0 - not supported within the method, QoS policy control and enforcement by PCRF and transport network layer

2.5.2.5.2 Definition of criteria weights by multiple decision makers

Multi-criteria decision making methods require the criteria weights as input. Weights reflect the importance of the criteria and can be derived by asking the opinion of multiple decision makers.

Local weights of branches are defined at each branching of the criteria tree (shown in Figure 48). The sum of local weights must be one at each branching. The global criteria weights (at the leaves) can be calculated by multiplying the weights of the branches starting from the root of the criteria tree to a specified leaf of the tree.

Two different approaches have been used for branch-level criteria weight definition: direct weight assignment, and logarithmic difference based weight definition.

Direct weight assignment should be used when few (e.g., 1 to 3) criteria fall under one branch of the criteria tree. In that case decision makers assign importance values to the criteria and the criteria weights are calculated by normalization. Accepted values are any

positive number assigned to each criterion. The proportion of the numbers determines the weights. Equation (1) shows the calculation of criteria weights.

$$c_i^{norm} = \sum_{d=1}^{D} \frac{P_d}{\sum_{j=1}^{D} P_j} \cdot \frac{C_{i,d}}{\sum_{i=1}^{M} C_{i,d}} \quad (1)$$

c_i^{norm} is the local criterion weight for the given branch of the criteria tree.

P_d is the importance of the decision maker, and can be any non-negative real number

$C_{i,d}$ is the importance of the criterion assigned by decision maker *d,* and can be any non-negative real number

D is the number of decision makers

M is the number of branching criteria

Logarithmic difference based weight assignment

In case of more than three criteria falling under a branch of the criteria tree, direct local weight assignment becomes difficult. In this case it is better to ask pair-wise comparison of the criteria from the decision makers, and calculate the final weights of the criteria from the pair-wise comparison values.

In the logarithmic difference assignment, a difference $\Delta_{i,j,d}$ between criteria i and j assigned by decision maker d means the subjective preference ratio $r_{i,j,d}$ between two critera described by Equation (2):

$$r_{i,j,d} = 2^{\gamma_d \cdot \Delta_{i,j,d}} \quad (2)$$

The progression factor of the geometrical scale is 2^{γ_d}, where γ_d value typically falls under the interval [0.5, 3]. If e.g., $\gamma_d = 1$, then $\Delta_{i,j,d} = 1$, $\Delta_{i,j,d} = 0$, $\Delta_{i,j,d} = -1$ mean that criterion i is twice more important, has the same importance or half important than criterion j, respectively. Changing the progression factor may change the granularity of steps in the pairwise preference ratio values.

Local criteria weights for a branch are calculated based on the logarithmic difference values assigned by the decision makers, and the importance of decision makers using Equation (3).

$$c_i = \prod_{d=1}^{D} \prod_{j=1}^{M} 2^{\frac{\gamma_d \cdot \Delta_{i,j,d} \cdot p_d}{M}} \left(\sum_{k=1}^{M} \prod_{d=1}^{D} \prod_{j=1}^{M} 2^{\frac{\gamma_d \cdot \Delta_{k,j,d} \cdot p_d}{M}} \right)^{-1} \quad (3)$$

c_i is the local criterion weight for the given branch of the criteria tree.

p_d is the weight of the decision maker d, and can be any non-negative real number

$\Delta_{i,j,d}$ is the relative difference between criteria i and j assigned by decision maker d

γ_d is the exponent of the progression factor of the geometric scale defining preference ratios

D is the number of decision makers

M is the number of criteria falling under the given branch of criteria tree

Seven decision makers from ALTO University, University of Oulu - Centre of Wireless Communication, Budapest University of Technology and Economics – Mobile Innovation Centre, Orange Labs have been asked to provide their opinion on criteria weights by filling criteria weight definition forms. The resulting local criteria weights are illustrated in Figure 59.

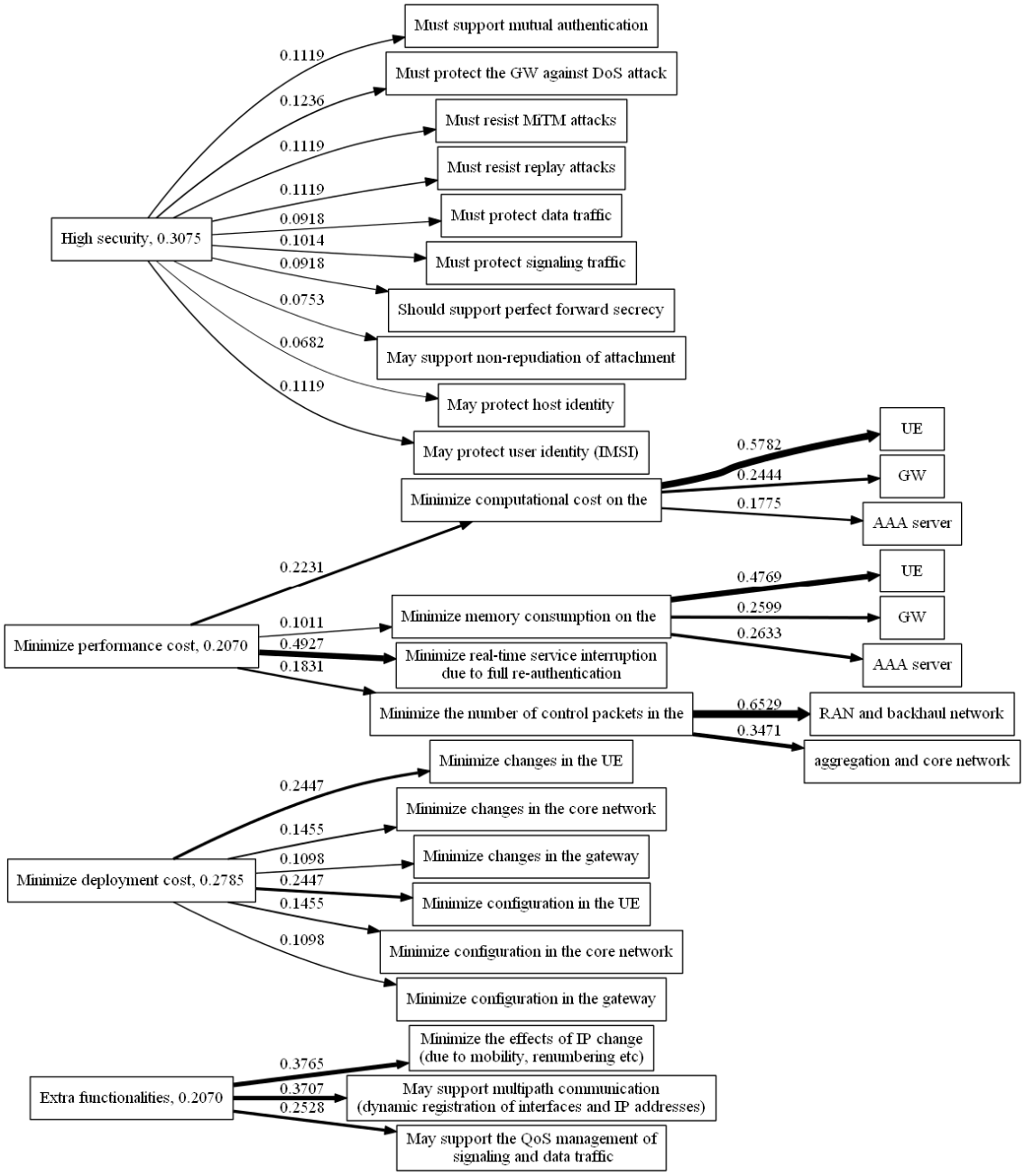


Figure 59 – Branch-level weights of the criteria

The reasoning behind weight allocations are summarized in the followings:

- From operator perspective, deployment cost minimalization is crucial. Then comes the performance cost and supported functionalities of the methods. However, in usage cases, where the subscriber needs high security, security might be even more important than deployment cost. This is reflected by the high weight of the security criterion. HIP/IPsec tunnelling shall be applied for UEs and MRs requiring high security, such as industrial remote monitoring and control. Note: the validation results contain perturbation analysis of the main criteria weights, which conduces to determine the robustness of HIP DEX-AKA method under different ranges of criteria weights.
- The weights of security sub-criteria reflect two approaches of the decision makers to set the weights. On the one hand, the provision of basic security services, such as mutual authentication, user identity protection, confidentiality, message origin and integrity protection of signalling data and user data is required, in slightly decreasing order of importance, because these basic security features mitigate the main threats. On the other hand, resistance to some attacks, mainly DoS attacks is considered very important by the decision makers. Support of non-repudiation of attachments,

perfect forward secrecy and host identity protection were considered less important security features.
Note: some of these security features have dependencies. E.g. mutual authentication, cryptographic binding of key material to authenticated identities and message authenticity protection are required to prevent man-in-the-middle attacks.

- Seamless user experience is the most important from the performance aspect. Hence real-time service interruption delay by re-attachment and re-authentication to new access networks is considered the most relevant performance criterion. The second most relevant criterion is the computational cost of the authentication process. The memory utilization is considered cheaper then CPU utilization and the arising energy consumption. Heap and stack memory consumption could be reduced thanks to processor evolution. CPU and memory cost is more crucial on the UE than on network elements. UE has limited resources so any extra impact on computational or memory consumption should be minimized. The processing and memory requirements in GW and or AAA can be handled with overprovisioning. Moreover, GW should be kept as generic as possible while AAA can be optimized for single task so memory and computational requirements are designed properly for the security tasks. The GW needs to serve a big number of customers with real-time requirements, while the AAA server is less frequently accessed. The network utilization caused by signalling messages is not so relevant but precedes memory utilization. Signalling is more crucial on the RAN and backhaul network than on the aggregation and core network. Control packets should be minimized in the RAN as it is the most energy hungry part of the whole system. Backhaul is often leased link and can be bottleneck.
- Regarding deployment cost, updating UE might be very complex since there are plenty of different models on the market. UE is more important for the success of a service or technology deployment. We could imagine any new service in the network: it can not work if the device is not compliant or cannot evolve. For example UMA, I-WLAN proves this comment, they are not successful since they need specific APIs on device.
- Regarding additional functionalities related to the security services, both IP-mobility management and multipath capabilities, i.e., dynamic registration of interfaces and IP addresses are important for the provision of seamless user experience. It is less relevant whether the security control protocol supports negotiation of QoS rules and enforcement of QoS policies is provided. QoS policy provisioning and enforcement shall be provided by PCRF and the transport network layer and L1/L2 in access network for the considered technologies.

2.5.2.5.3 Evaluation process

In the previous subsections, a criterion set has been defined for the evaluation of six different authentication methods. The criteria have been structured in the form of criteria tree.

This was followed by the evaluation of each method under each criterion that is in the set of leaves of the criteria tree depicted in Figure 48. For the evaluation of the methods under the leaf-criteria, either direct grade assignment has been used, or, in case of a quantitative KPI could be found for the criterion, the KPIs have been measured for the method. Following that, grades have been assigned to the methods using the previously described performance grade assignment functions.

The grades of the methods together with criteria weights form the set of input parameters for the final evaluation of the methods.

The evaluation is expected to provide results on the ordering of the methods under each element of the criteria tree, including the overall ranking.

The calculation of the terminal scores are based on the score aggregation method of the Multiplicative Analytic Hierarchy Process, slightly modified to handle the non-fulfilment of some criteria. The method allows for each leaf-criterion an optional rejection feature. If the option is set then zero grade means rejection of the authentication method. Otherwise all grades are acceptable.

Equations (4) describe the calculation of the terminal scores of the methods under each criterion.

$$t_{j\,(1)} = \prod_{i=1}^{M}\prod_{k=1}^{N}\begin{cases} 2^{\frac{\gamma_i(v_{i,j}-v_{i,k})c_i}{m_i}}, \ if\ \{v_{i,j}, v_{i,k}\} \in \{grades\_accept\} \\ 1, \ if\ v_{i,j} \in \{grades\_accept\}\ AND\ v_{i,k} = grade\_reject \\ 0, if\ v_{i,j} = grade\_reject \end{cases}, \ j = 1..N \quad (4)$$

$$m_i = \begin{cases} |\{j \mid v_{i,j} \in \{grades\_accept\}\}| \\ 1, if\ |\{j \mid v_{i,j} \in \{grades\_accept\}\}| = 0 \end{cases}, \quad i = 1..M$$

$$t_{j\,(2)} = \prod_{i=1}^{M} t_{i,j}^{c_i}, \ j = 1..N$$

Indexes:

i – index of criteria contained in a specific branch of the criteria tree.

j – index of different alternatives to rank, i.e., the authentication methods

Parameters:

M – number of criteria in the selected branch of the criteria tree.

N – number of alternatives

c_i– local weight of a criterion within a selected branch of the criteria tree

m_i– number of alternatives that obtained enough good grades under criterion i so that the these alternatives should not be rejected. If every grade is acceptable under criterion i, then for that i value it is always the first row that is used in the expression behind the bracket.

$v_{i,j}$ or $v_{i,k}$ – grade of alternative j or k under criterion i

Variables:

$t_{i,j}$ – terminal score of alternative j under criterion i

$t_{j,(1)}$– terminal score of alternative j under a criterion where the sub-criteria are the leaves of the criterion tree, i.e., under the sub-criteria grades had been assigned to the methods.

$t_{j,(2)}$ – generic expression to calculate the terminal score of alternative j under a criterion that is the parent of other criteria.

Methods getting higher terminal score are considered better under a given criterion.

In our analysis, 0 grade means rejection of the alternative under the following criteria:

- Must support mutual authentication, i.e., methods not fulfilling the “must” requirements should not be accepted
- Must protect the GW against DoS attack
- Must resist MiTM attacks
- Must resist replay attacks
- Must protect data traffic
- Must protect control traffic
- Minimize configuration in the UE, i.e., if additional key management or ACL management is required, and these functionalities are out of the scope of the authentication method, then the method should not be accepted.
- Minimize configuration in the core network.
- Minimize configuration in the GW.

2.5.2.5.4 Terminal scores of the alternatives with fixed criteria weights

Figure 60 shows the terminal scores of the methods under each criterion in distributed reference scenario. The left part is for HSDPA DL/UMTS UL access. The right part is for Wi-Fi access.

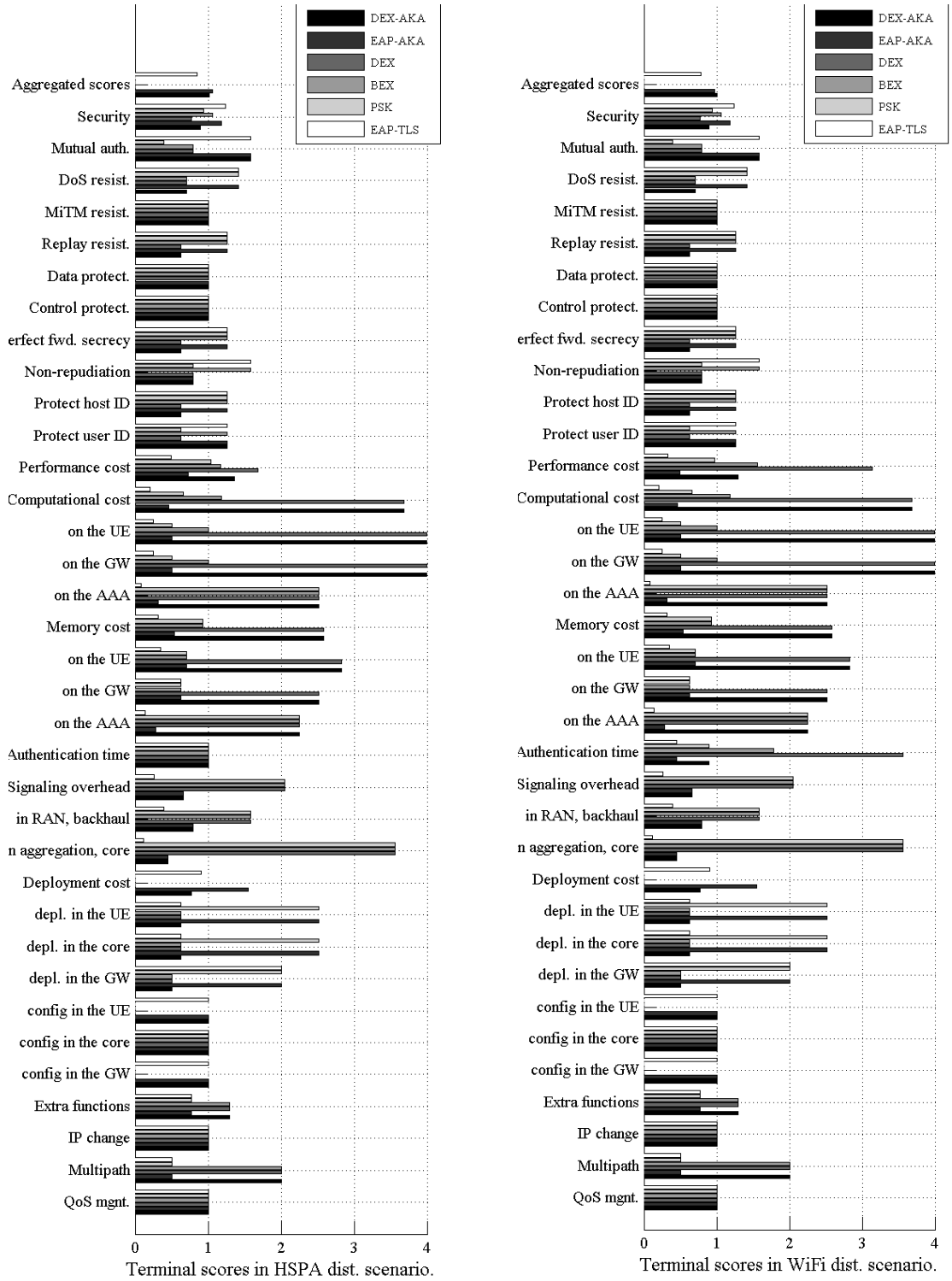


Figure 60 – Terminal scores of the methods under each criterion in distributed reference scenario.

The terminal scores obtained in the centralized and flat reference scenarios are very similar to these figures, because they differ only in the “Authentication time”, i.e., the metric applied for the real-time service interruption delay.

The “Aggregated scores” show the overall terminal scores under all criteria for the alternatives. The value of the terminal score is not important, just the ordering of the values is relevant to obtain the preference order of the methods. These values show that IKEv2 EAP-AKA is the most preferred method in case of HSDPA/UMTS access, while HIP DEX-AKA method is first in case of Wi-Fi access. The difference seems very small between the

two methods in both access types. The aggregated scores are the result of the actual trade-off between different criteria specific to each authentication method under the specified criteria weights.

IKEv2 PSK, HIP DEX and HIP BEX have been rejected, i.e. got zero terminal score, under the aggregated scores, because of their rejection under the configuration criteria in the UE and GW. This is due to the bad scalability of the management/configuration of these authentication methods in large-scale network, the key management cost of pre-shared keys in case of IKEv2 PSK, or management of access control lists based on HITs in case of HIP DEX and BEX.

The scores achieved under security and the related sub-criteria are the same in all reference scenarios, and are the direct consequences of the security features of the methods described previously.

Under performance criteria, the computational and memory cost of HIP DEX and DEX-AKA are the best among the methods. Regarding signalling cost, i.e., the number of control messages, HIP DEX-AKA performs the same as IKEv2 EAP-AKA, HIP DEX performs the same as HIP BEX and IKEv2 PSK due to the similarities in their grades.

Under authentication time criterion, in case of HSDPA/UMTS access, the methods got the same score, i.e., all got 0 grade due to non fulfilment of the 300 ms packet delay budget of real-time applications. In case of Wi-Fi access, HIP-based methods got positive grades, while IKEv2 EAP-TLS and EAP-AKA methods got 0 grades. This is reflected by the terminal scores under authentication time in case of Wi-Fi access. Note that if the support of real-time packet delay budgets during handover was a must criterion, then only HIP DEX-AKA could be acceptable among the authentication methods working in large-scale environment.

Among the deployment criteria, IKEv2-based methods naturally get higher terminal scores, since IKEv2 EAP-AKA is assumed to be supported already by the network. Under the extra functionalities criterion, HIP-based methods perform somewhat better due to their better multipath capability feature, then in case of IKEv2-based methods.

2.5.2.5.5 Terminal scores of the alternatives with running criteria weights

The robustness of the decisions can be analyzed by perturbing the parameters influencing the outcome. In this section we describe the results of an important robustness analysis, where the weights of the four main criteria are varied. The analysis is done by picking one of the main criteria, and linearly increment its value in the interval [0..1]. Meanwhile the weights of the other criteria are reduced proportionally, in such way that the sum of the weights of the main criteria remains 1, and the proportion of the other criteria related to each other remains as it was in the fixed case.

The rejected methods (IKEv2 PSK, HIP DEX and HIP BEX) get always 0 terminal score, except when the weight of the deployment criteria is set to 0, i.e., deployment cost does not matter.

The following figures show the variation of the aggregated terminal scores of each method in case of the distributed reference scenario with HSDPA/UMTS and Wi-Fi access, respectively, when perturbing the weight of

- the security criterion (Figure 61, Figure 62)
- the performance criterion (Figure 63, Figure 64)
- the deployment criterion (Figure 65, Figure 66)
- the extra functionalities criterion (Figure 67, Figure 68)

The results show that HIP DEX-AKA should be the preferred method in case of high performance and extra functionality requirements. Otherwise, higher security requirements bring IKEv2 EAP-AKA method the preferable alternative. If only security requirements count, IKEv2 EAP-TLS is the most preferable method, because it supports digital signature of the TLS-client and TLS-server, additionally to EAP-AKA method where no digital signatures are incorporated in the protocol. Higher deployment cost requirements favour the use of IKEv2 EAP-AKA.

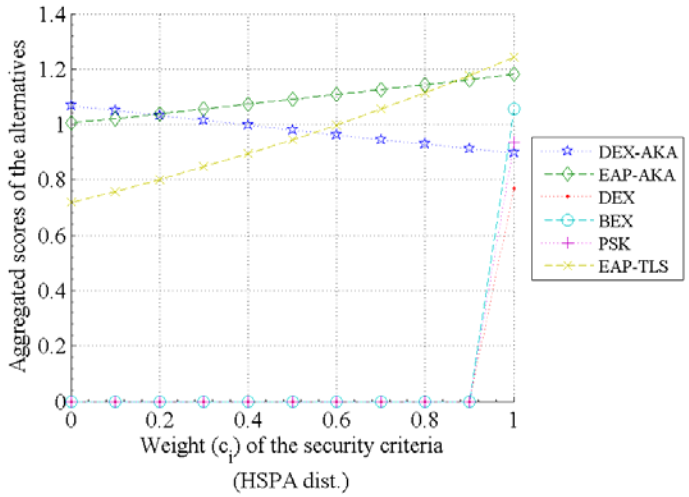


Figure 61 – Perturbation of the weight of security criterion in case of distributed scenario with HSDPA/UMTS access.

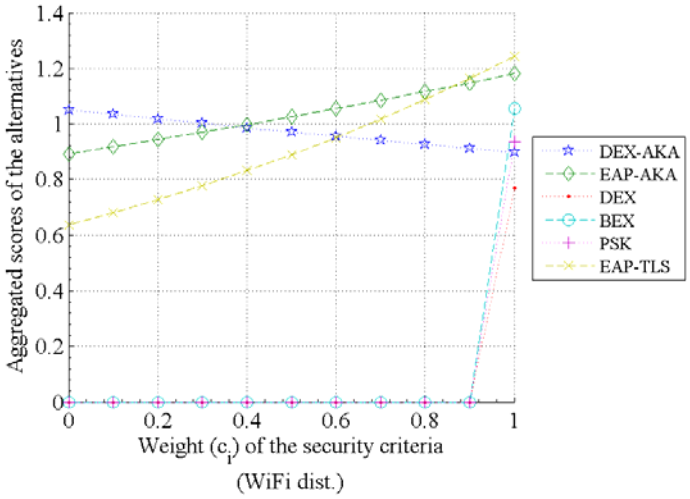


Figure 62 - Perturbation of the weight of security criterion in case of distributed scenario with Wi-Fi access.

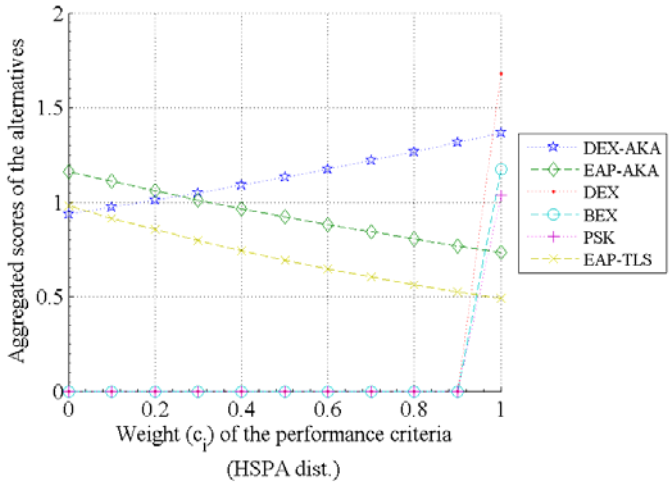


Figure 63 - Perturbation of the weight of performance criterion in case of distributed reference scenario with HSDPA/UMTS access.

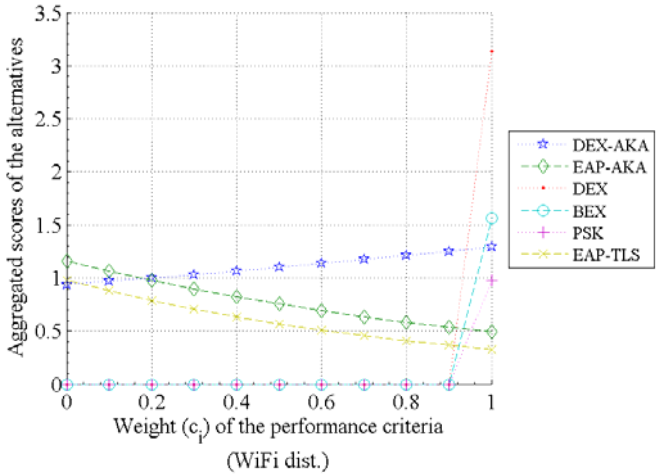


Figure 64 - Perturbation of the weight of performance criterion in case of distributed reference scenario with Wi-Fi access.

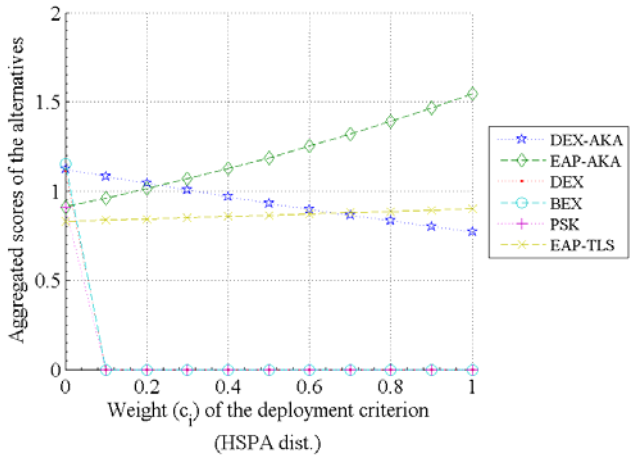


Figure 65 - Perturbation of the weight of deployment criterion in case of distributed reference scenario with HSDPA/UMTS access.

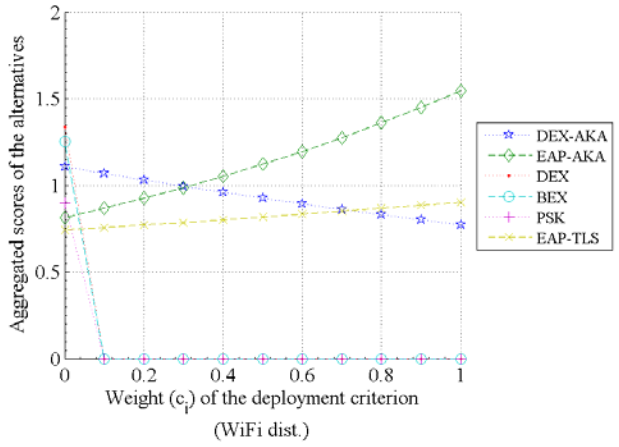


Figure 66 - Perturbation of the weight of deployment criterion in case of distributed reference scenario with Wi-Fi access.

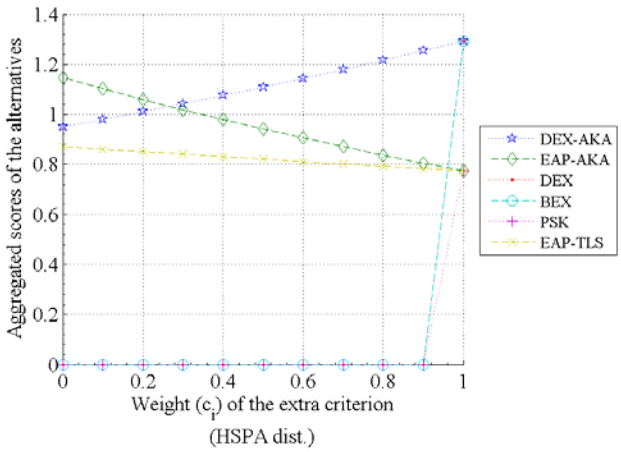


Figure 67 - Perturbation of the weight of extra functionalities criterion in case of distributed reference scenario with HSDPA/UMTS access.

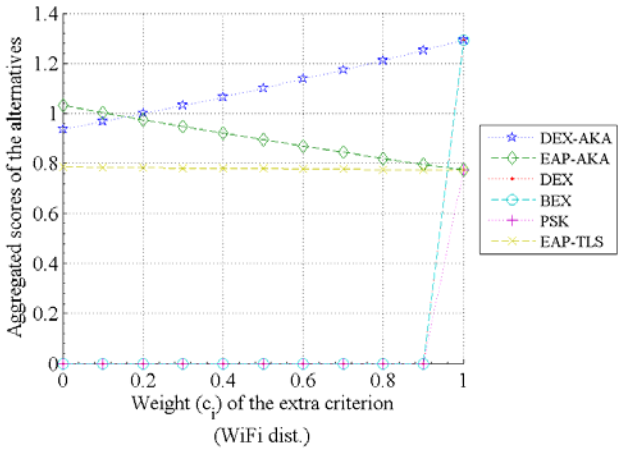


Figure 68 - Perturbation of the weight of extra functionalities criterion in case of distributed reference scenario with Wi-Fi access.

2.5.2.6 Applicability of the results

The results show that HIP-DEX-AKA method should only be used for UEs with very low computational and memory resources, and requiring the most important security features. However, the security of the original IKEv2 EAP-AKA method is a bit stronger due to for e.g., the support of perfect forward secrecy. IKEv2 EAP-AKA should be the applied method in use cases where there are no extra requirements regarding performance and multipath capabilities.

These multi-criteria decision processes try to decrease the subjectivity of decisions, and, at least for the chosen set of criteria they provide formal steps for the decision makers to think about important aspects of the compared alternatives. Subjectivity cannot be eliminated when dealing with orthogonal criteria. But by fixing realistic assumptions, the credibility of why choosing a given authentication method can be highly supported.

2.5.2.7 Partners involved

BME-MIK will coordinate this evaluation task. Multi-criteria decision making and ranking of alternatives will be more efficient if it reflects the thoughts of more decision makers. CWC will support this task by common discussions, and giving the opinion of researchers on different questions which will arise during the evaluation.

2.6 Support for user cooperation

Relaying techniques are considered as an alternative solution to enhance capacity for the cell network, to extend coverage in specific locations, to increase throughput in hotspots or to overcome excessive shadowing. It gives important advantages such as ease of deployment and reduced deployment cost compared to deploying regular Base Station (BS).

2.6.1 Performance evaluation of mobile relaying and its management

2.6.1.1 Objectives

One of the key expectations for the future wireless system is to provide ubiquitous high data rate coverage in the most cost-effective manner. During the recent years, a large part of the research has focused on Orthogonal Frequency Division Multiple Access (OFDMA) transmission technology, a very promising candidate for the physical layer in next generation cellular system, due to its inherent robustness against frequency–selective fading and its capacity for achieving high spectral efficiency. In OFDMA, each subcarrier can be allocated to a different user which can best exploit the current channel condition, hence maximizing the achievable capacity. But with the traditional cellular architecture, increasing the capacity along with the coverage would require the deployment of a large number of Base Stations (BS), which turns out to be a cost-wise inefficient solution to service providers. However, introducing Relay Stations (RS) in each cell can alleviate this problem since the RS can forward high data rates in remote areas of the cell while keeping a low cost of infrastructure.

In the literature two different types of relaying network architecture have been investigated as fixed relay station (FRS) and mobile relay station (MRS) as shown in Figure 69. The FRSs are part of the network infrastructure, thus where and how much FRSs will be deployed in a cell will be processed while the network planning, design and deployment process by operators. Compared to FRS, MRS can be flexible employed in a wireless cellular network. MRSs are effectively a moving aspect of FRSs. The goal of employing MRSs is not to replace FRSs, but rather to act as a complementary solution. In general, there are mainly two different types of scenarios MRSs employed in the wireless cellular networks. One is MRSs fitted on moving vehicle, e.g. trains, buses, cars, etc. to cover areas/UEs in/on/outside the vehicle. The other is the non-active Mobile Station (MS) (i.e. in idle state) acting as MRS to relay the signals of the active MS to BS.

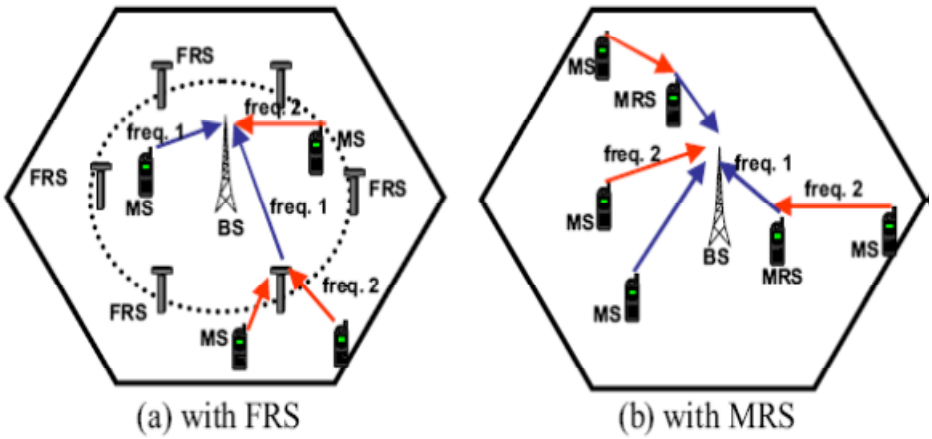


Figure 69 - Cellular Networks with Relaying.

In these relay-enhanced networks, potential gain in capacity and coverage is highly dependent on the radio resource management (RRM) strategy, a topic which draws more and more attention of the research community. How to perform RRM in such a complex environment is a big challenge and not clear at all. Increased number of links makes the resource allocation problem difficult to tackle.

Furthermore, cellular networks are overloaded with mobile data traffic due to the rapid growth of mobile broadband subscriptions. The combination of smartphones such as iPhones, netbooks and 3G/4G mobile networks are rapidly growing in very large numbers and as a result, this has created an exceptional demand for ubiquitous connectivity and quality of rich digital content and applications. To meet the requirements of future data-rich applications and terminals with improved multimedia, future wireless networks are expected to combine multiple access technologies and as a result mobile broadband operators are including WLANs like Wi-Fi as an alternative access network technology. This enables solutions to offload traffic from the primary access technology to the Wi-Fi access when applicable so as to provide extra capacity and improve overall performance. In these scenarios, access network selection and resource allocation is also problems that must be overcame.

In this study, we have examined three different scenarios that handle the problems above. In the first two scenarios, we have studied on the resource allocation problem for LTE-based

wireless networks which is enhanced by MRSs. The MRSs' that we have used are MSs acting as relays for other users. In the third scenario, we have focused on the Wi-Fi offloading problem that is addressing to overcome the mobile network congestion by offloading a portion of mobile data traffic to complementary wireless access networks using Wi-Fi. These scenarios are explained in detail as follows.

2.6.1.2 Validation scenarios

2.6.1.2.1 Scenario I

In the first scenario, single cell downlink LTE-based relay-enhanced network topology is used as shown in Figure 70. The BS is located in the centre of the cell and users are distributed uniformly around it. The cell area is divided into two zones; inner $(0 - 2R/3)$ and outer $(2R/3 - R)$ zones where R is the radius of the cell. The users in the inner zone are allowed to communicate with the BS directly and the users those who are in the outer zone (cell-edge users) can communicate with the BS either directly or over another user (relaying).

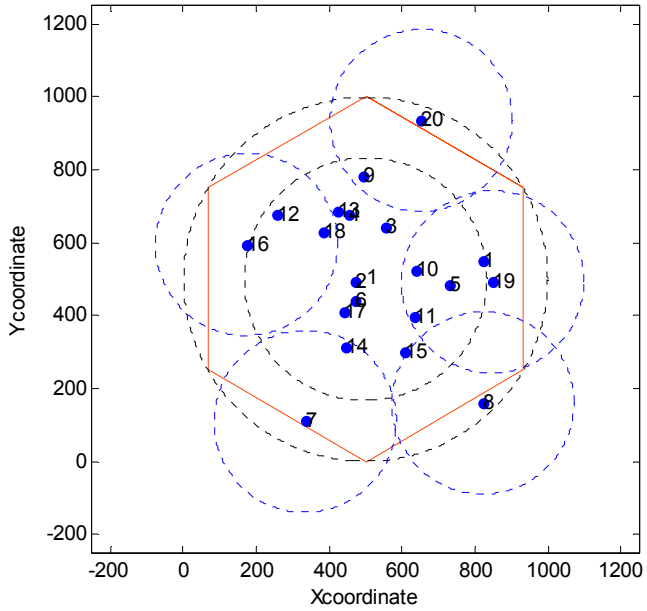


Figure 70 - Network topology of Scenario I.

The cell edge users can determine their relay candidates using their coverage areas whose radius is $R/2$ in this study. The user set which is in the coverage area of a cell edge user is the relay candidate set of this cell edge user. For example, the relay candidate set is {1,5,10,11} for the cell-edge user 19, as seen in Figure 70. Among these relay candidates, one relay is chosen by using the Minimum Total Path loss (MTP) Selection as follows;

$$r_s = \arg\min_{all\ r \in \Lambda}(PL_{r1} + PL_{r2}) \qquad (5)$$

Where, Λ denotes the set of candidate relaying nodes, PL_{r1} and PL_{r2} denote the pathlosses in dB associated with the first (between the base station to the candidate relaying node) and the second (between the candidate relaying node to the relayed node) hops, respectively. The selected relay candidates are sent to the BS by the cell-edge users and BS uses this information for the resource allocation.

Resource allocation is performed by the BS which has the channel knowledge between the BS-MSs and MS-RSs. Half duplex (HD) relaying is used so resource allocation is performed in two time slots as shown in Figure 71. In the first time slot, BS sends data to MSs and RSs and in the second time slot MSs receive information from RSs and BS. Sub-channels are allocated to the BS-MS, BS-RS links at time slot one and RS-MS, BS-MS links at time slot two.

Maximum Channel Quality Indicator (CQI) scheduling is used in order to allocate the resources to the users. The sub-channel is given to the link which has the max CQI value. In order to obtain a fairness between the users a threshold data rate value (R_{th}) is determined

and the user who reached this data rate value is removed from the system and called satisfied user.

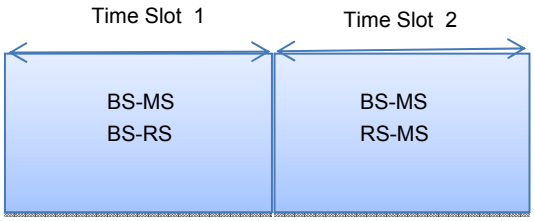


Figure 71 – Resource allocation.

In the simulation of Scenario I, we compared two conditions that use relay and do not use relay to communicate with the BS. Thus, the resource allocation algorithms are explained in detail for these conditions; with and without relay cases as follows.

Algorithm1-Without Relay Case (Only Macro Users)

Let K, N, R, R_{th}, U be the number of total users, number of subchannels, total data rate of each user, threshold data rate value that each user has to get and unsatisfied user set, respectively.

Initially, $R_k = 0$, $\forall k$, $U = \{1,2, \dots, K\}$

$n = 1$;

while $U \neq \emptyset$ do

- Determine the CQI values for $\forall k \in U$ from the LTE CQI table by using the SNR values.
- Find the user k' that has the maximum CQI value for the subchannel n. $k' = \arg\max_{k \in U}(CQI_{k,n})$
- Update the data rate $R_{k'}$ using CQI table.
- If $R_{k'} \geq R_{th}$, user k' is satisfied, so remove it from the unsatisfied user set $U \leftarrow U \setminus \{k'\}$.
- Increase n by 1 until $n = N$.

end while

Algorithm2-With Relay Case (Macro+Relay Users)

Let N_1, N_2 be number of subchannels at time slot 1 and slot2, respectively.

Initialization

$R_k = 0$, $\forall k$, $U = \{1,2, \dots, K\}$, $N_1 = \{1,2, \dots, N\}$, $N_2 = \{1,2, \dots, N\}$

Step1

$n = 1$;

while $U \neq \emptyset$ do

- Determine the CQI values for $\forall k \in U$ from the LTE CQI table by using the SNR values of the users. If k is the outer user and using relay to communicate with BS the CQI value is the minimum of the two hop links (BS-RS and RS-MS) such as; $CQI_{k,n} = \min(CQI_{BS-RS}, CQI_{RS-MS})$
- Find the user k' that has the maximum CQI value for the subchannel n. $k' = \arg\max_{k \in U}(CQI_{k,n})$
- Update the total rate of the user $k', R_{k'}$ using CQI table.
- If $R_{k'} \geq R_{th}$, user k' is satisfied, so remove it from the unsatisfied user set, $U \leftarrow U \setminus \{k'\}$.

- If user $k^{'}$ is a user that communicating with BS over a relay, remove the sub-channel n from the second time slot subchannel set. Since, we assume that the RS transmit and receive at the same subchannel, $N_2 \leftarrow N_2 \setminus \{n\}$
- Increase n by 1 until $n = N$.

end while

Step2

while $U \neq \emptyset$ and $N_2 \neq \emptyset$ do

- Select a sub-channel n_j' through the remaining sub-channels of N_2.
$$n_j' = \arg\min_{n_j \in N_2} n_j$$
- Determine the CQI values for $\forall k \in U$ from the CQI table by using the SNR values of the users.
- Find the user $k^{'}$ that has the maximum CQI value for the sub-channel n_j'.
$k' = \arg\max_{k \in U}(CQI_{k,n_j'})$
- Update the $R_{k'}$ using CQI table.
- If $R_{k'} \geq R_{th}$, user k' is satisfied, so remove it from the unsatisfied user set $U \leftarrow U \setminus \{k'\}$.
- $N_2 \leftarrow N_2 \setminus \{n_j'\}$.

end while

2.6.1.2.2 Scenario II

This scenario is very similar to Scenario I. However, in this scenario the users which are in the inner zone are classified as active and inactive users. Only inactive users can be a relay candidate for the users at the outer zone. For example, as shown in Figure 72, the active users {10,11} cannot be the relay candidates of the cell-edge user 19 at all. These active users only communicate with the BS for themselves. Inactive users {1,5} can be the relay candidate of the cell-edge user 19. Among these relay candidates, one relay is chosen again by using MTP Selection as defined in Equation (5). The resource allocation part is the same with the Scenario I.

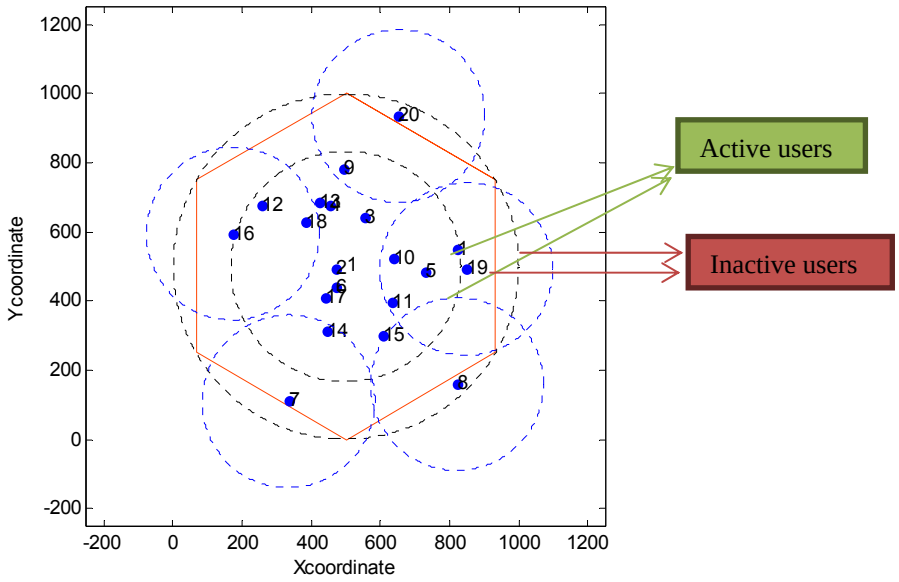


Figure 72 - Network topology of Scenario II.

2.6.1.2.3 Scenario III: Wi-Fi Offloading

This scenario addresses Wi-Fi offloading as a solution to the exploding future growth of mobile broadband data traffic in the deployed LTE networks thereby using Wi-Fi as an alternative access network technology. The reason why traffic offloading by Wi-Fi is considered to be a viable solution of mobile data traffic explosion is that because there is a lot of unlicensed Wi-Fi spectrum already existing with very large number of compatible devices in which operators can make use of. We have examined this scenario with relaying and without relaying cases as given below.

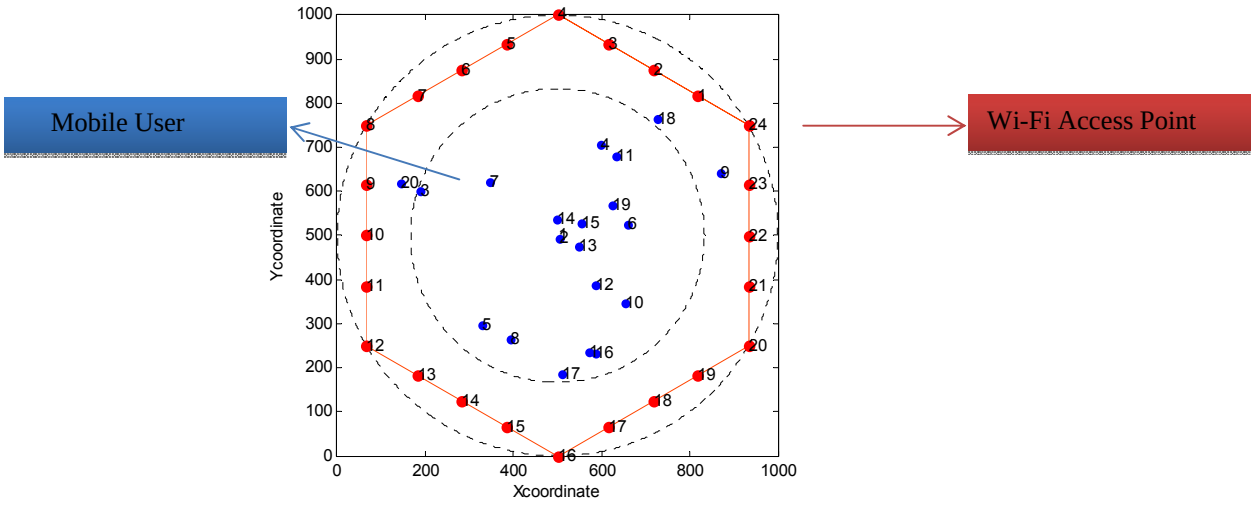
2.6.1.2.3.1 Wi-Fi-Offloading without relaying concept

In this scenario, different number of Wi-Fi points are located at the cell edges or uniformly distributed at the outer zone as shown in Figure 73-a and Figure 73-b, respectively. Next Generation broadband wireless heterogeneous networks are characterized by the coexistence of multi-access wireless networks utilizing different access network technologies which complement each other in terms of offered bandwidth and operational costs e.g. LTE, Wi-Fi etc. In such multi-access wireless access networks, network discovery and access selection are the fundamental problems. As its name indicates, access selection refers to the process of deciding over which access network to connect at any point in time.

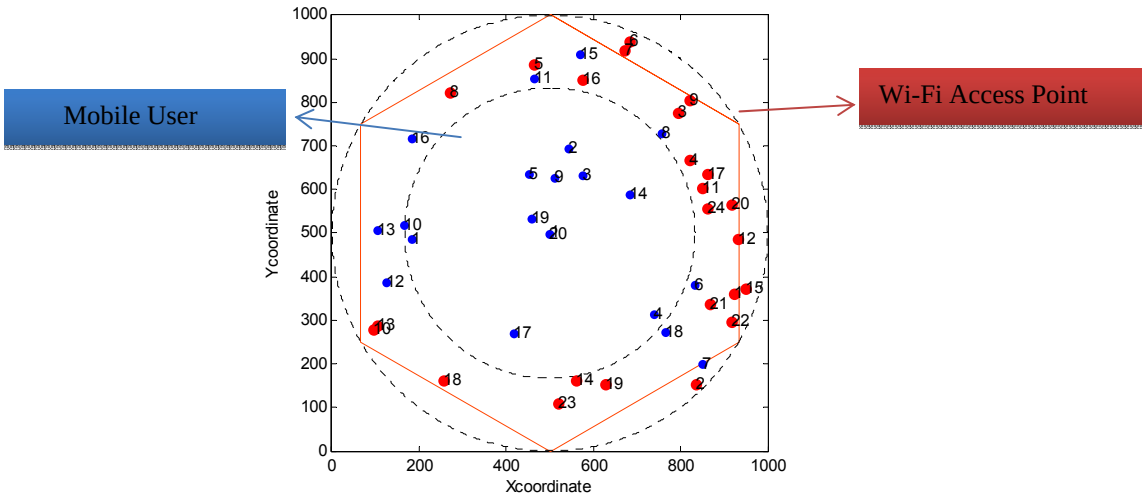
In this study, the SNR-based network selection algorithm is studied. This algorithm may formally be described as:

$$network_k = \begin{cases} WiFi, & SNR_{WiFi,k} \geq SNR_{th} \\ LTE, & o.w \end{cases} \quad (6)$$

A user k thus selects Wi-Fi if the SNR from the best Wi-Fi point, which has the maximum SNR with user k, equals or exceeds the threshold SNR_{th}.



(a)



(b)

Figure 73 - Network topology of Scenario III-A.

After the selection of all users' access networks, resource allocation is performed for Wi-Fi and LTE users. For the LTE users, Algorithm 1 is used to allocate the sub-channels among them and for the Wi-Fi users; Algorithm 3 is used that is defined below;

Algorithm 3-(Resource Allocation for Wi-Fi Users)

Let K_w, N, R, R_{th}, U be the number of total Wi-Fi users, number of subcarriers, data rate of the Wi-Fi users, threshold data rate value that each user has to get and unsatisfied user set, respectively.

Initially, $R_k = 0$, $\forall k$, $U = \{1,2,\dots,K_w\}$

$n = 1$;

while $U \neq \emptyset$ do

- Determine the Modulation and Coding Scheme (MCS) values for $\forall k \in U$ from the MCS table of 802.11n standard by using the SNR values of the users.
- Find the user k' that has the maximum MCS value for the sub-channel n.
  $k' = \arg\max_{k \in U}(MCS_{k,n})$
- Update the $R_{k'}$ using MCS table.
- If $R_{k'} \geq R_{th}$, user k' is satisfied, so remove it from the unsatisfied user set $U \leftarrow U \setminus \{k'\}$.
- Increase n by 1 until $n = N$.

end while

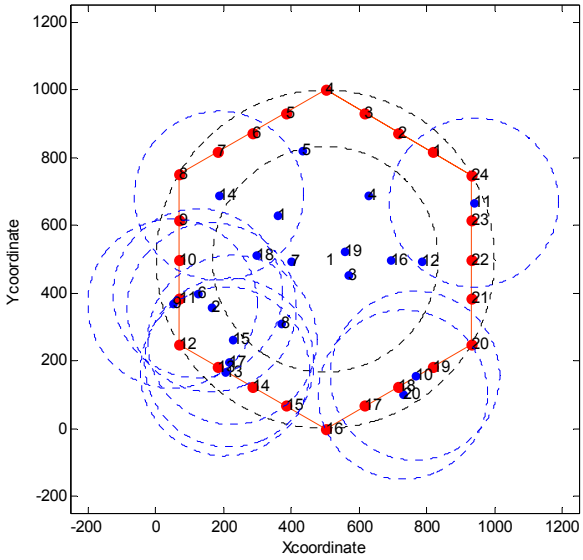


Figure 74 - Network topology of Scenario III-B.

2.6.1.2.3.2 Wi-Fi-Offloading with relaying concept

This scenario is the expanded version of Scenario III-A. The difference is LTE users at the outer zone can also use mobile relays if their direct link is not good as given in Figure 74. The network selection and resource allocation is also performed independently in this scenario. The network selection procedure is the same as given in Scenario III-A. Algorithm 2 and Algorithm 3 is used for the resource allocation of LTE and Wi-Fi users, respectively.

2.6.1.3 Validation tools

MATLAB test environment will be used to simulate the proposed mobile relaying algorithm. The proposed algorithm will use the real LTE network parameters.

2.6.1.4 Expected results

Our algorithm will primarily investigate edge-user throughput improvements over possible deployment of relay nodes in a heterogeneous environment and its impact on the core network performance. The signalling overhead of relay deployment is not considered as significant compared to data traffic in the backhaul or core network. It is estimated that core data traffic requirements will increase to 130 Gbps (Current core bandwidth requirements are less than 40 Gbps) in Europe [24].

2.6.1.5 Results

The scenarios explained in detail in the Section II are simulated using MATLAB and simulation results are obtained. The simulation parameters that we have used in the simulations are given in Table 1 and Table 2 for LTE and Wi-Fi users, respectively.

Table 9 - LTE users' system parameters

Parameter	Value
Frequency	2 GHz
Bandwidth	20MHz
Thermal Noise Density	-134.89dBm/Hz
nTX x nRX antennas	1 x 1
eNodeB TX power	49dBm
UE as relay TX power	23dBm
Cell radius	500m
Pathloss model	128.1 +37.6* log10 (d)
Shadowing model	Lognormal distribution, μ=0, σ=10(dB)
Multipath model	Extended Pedestrian A (EPA)
UEs position	Uniformly distributed in the cell area.

Table 10 - Wi-Fi users' system parameters

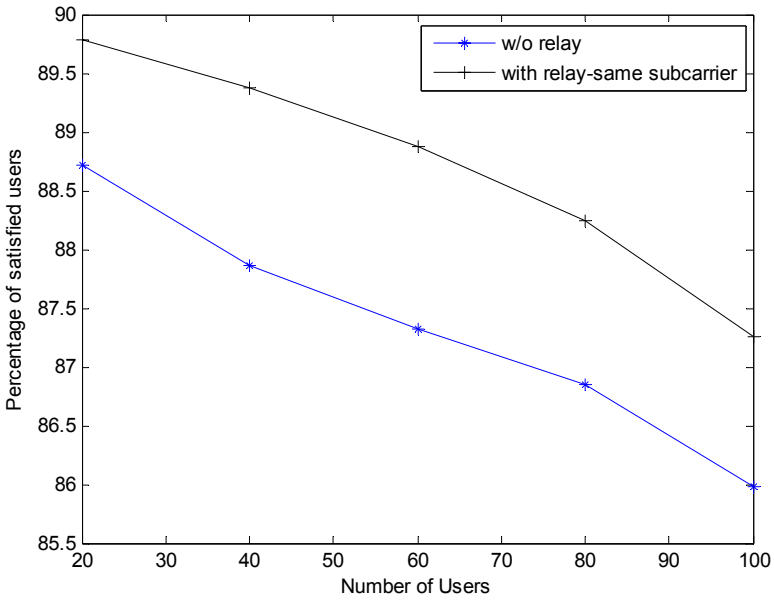
Parameter	Value
Frequency (fc)	2.4GHz
Bandwidth	20MHz
Thermal Noise Density	-134.89dBm/Hz

nTX x nRX antennas	1 x 1
Wi-Fi TX power	20dBm
Path loss model IEEE802.11 TGn channel model D standard NLOS	20*log10(4*pi*fc*d_BP/3e8) + 35*log10(d/d_BP) d_BP : 10 (breakpoint distance in meters)
Shadowing model	Lognormal distribution, μ=0, σ=5 (dB)
Multipath model	IEEE 802.11 TGn channel model D
Number of Wi-Fi points	12,24,36

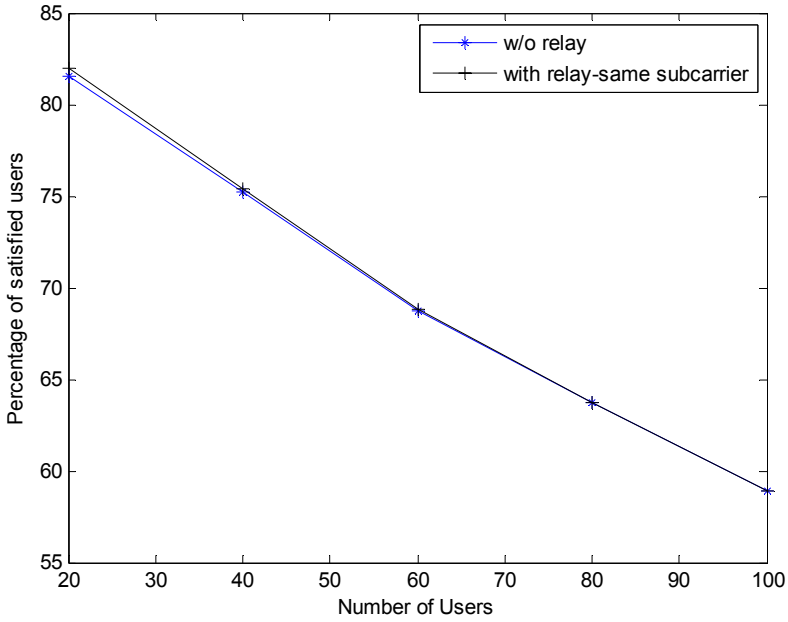
Performance results for Scenario I:

First of all, we obtained the results for Scenario I which we compared two cases as relay case and without relay case. The simulation results are obtained for different R_{th} values. In Figure 75-a and Figure 75-b the percentage of satisfied users are obtained for $R_{th} = 168kbps$ and $R_{th} = 933kbps$, respectively. For the low data rate threshold value as seen in Figure 75-a, the number of satisfied users are increased by using relays. When we increased the R_{th} value to $933kbps$ this difference between satisfied users is getting closer since the number of subchannels allocated to relayed users is decreasing.

Figure 76 and Figure 77 show the sum rate of all users and sum rate of cell edge users, respectively. As expected, when the relays are used in the system not only total data rate but also cell edge data rate is being increased. The effect of relaying is seen at the cell edge users clearly since these users can exploit the other users as relays. Again, we can see from these figures that increasing the threshold data rate is extinguishing the benefits of relaying since the number of sub-channels that will be used by the relayed users is decreasing.

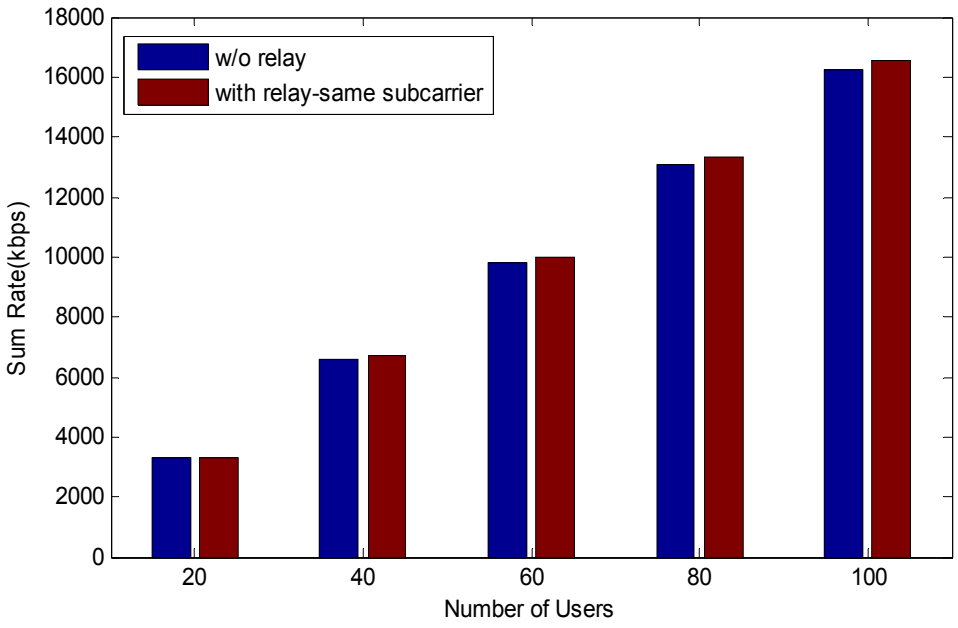


(a)

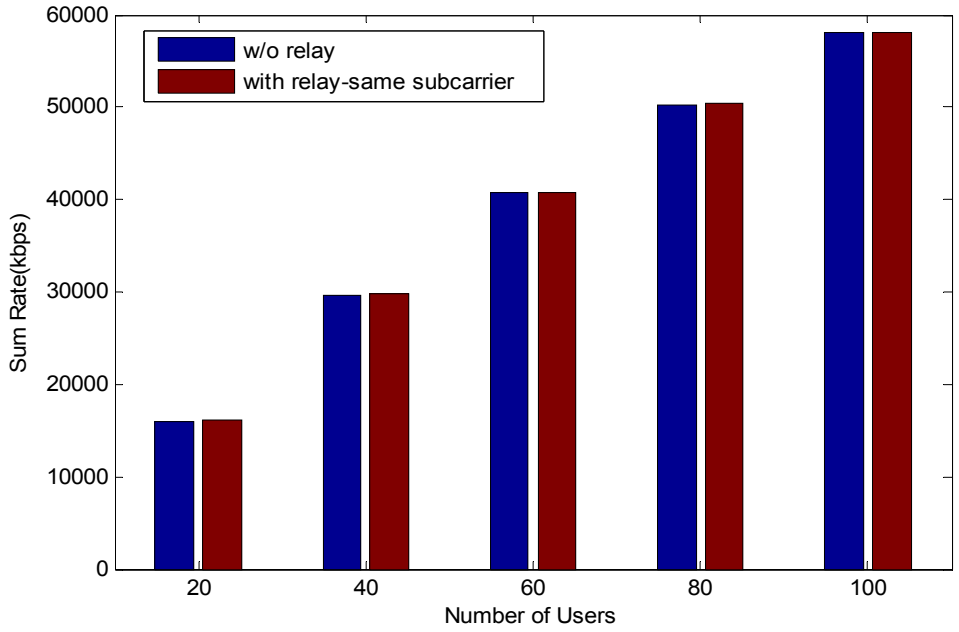


(b)

Figure 75 - Percentage of satisfied users vs Number of Users for Scenario I a) R_{th}=168kbps, b) R_{th}=933kbps

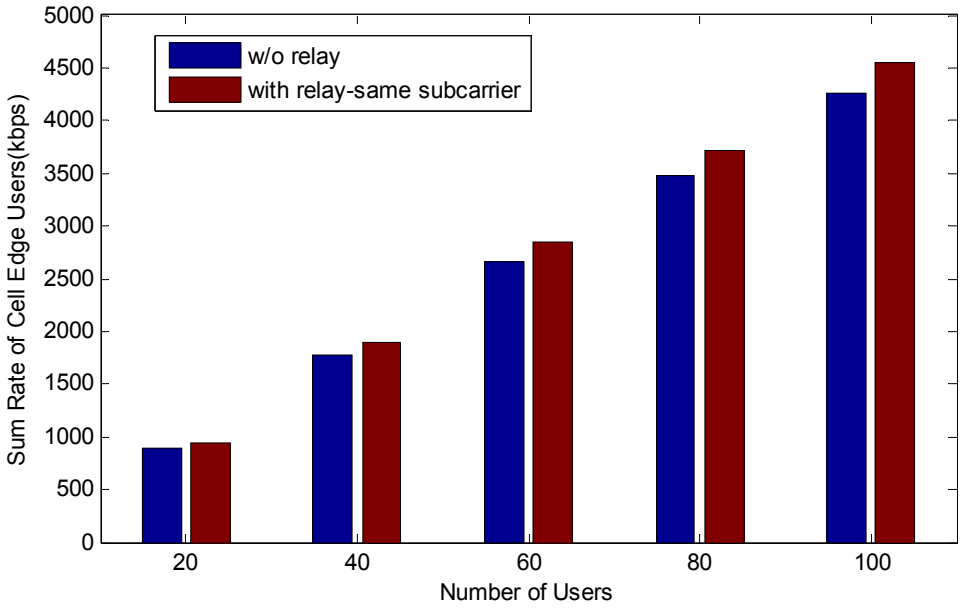


(a)

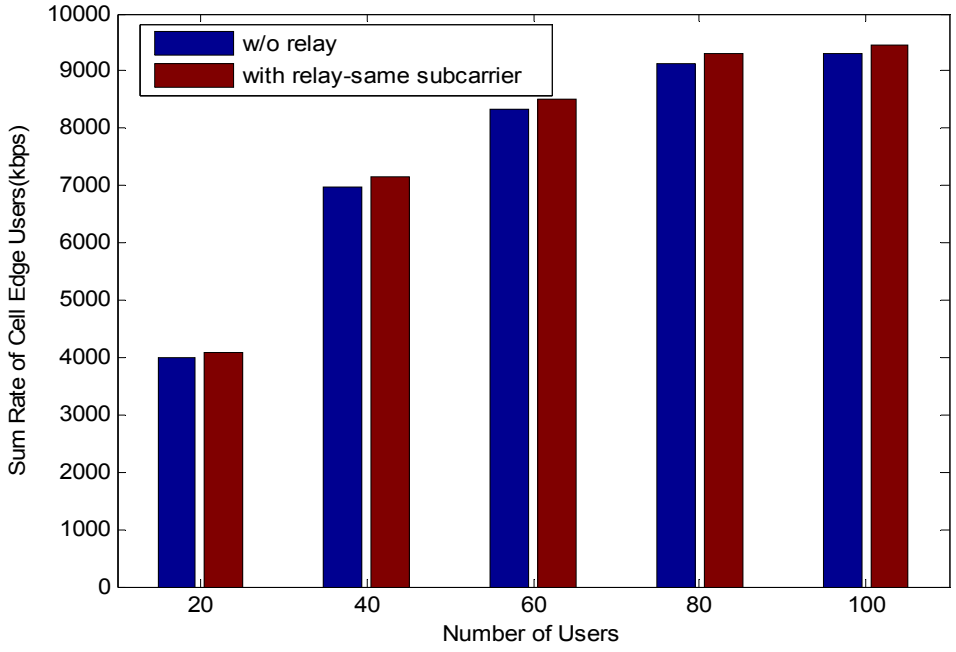


(b)

Figure 76 - Sum Rate vs Number of Users for Scenario I
a) $R_{th} = 168kbps$, **b)** $R_{th} = 933kbps$



(a)



(b)

Figure 77 - Sum Rate of celledge users vs Number of Users for Scenario I

a) $R_{th} = 168kbps$, **b)** $R_{th} = 933kbps$

Performance results for Scenario II:

In Scenario II, not all inner users can be used as relay by the outer users. The outer users can only use the inactive inner users as relays. We obtained some results by selecting the active inner user percentage as 50% and 70%. The results for 50% are shown in Figure 78, Figure 79 and Figure 80, and the results for 70% are given in Figure 81, Figure 82 and Figure 83. From all these figures, the superiority of the relaying case compared to without relay case is revealed.

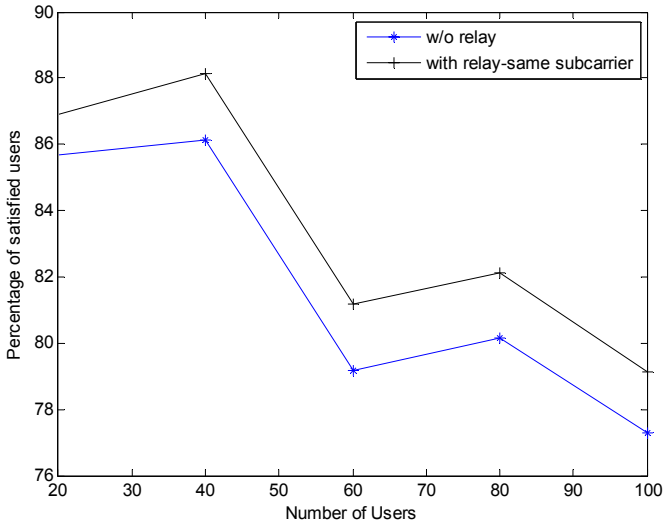


Figure 78 - Percentage of satisfied users vs Number of Users for Scenario II

$R_{th} = 168kbps$**, Active inner user percentage=50%**

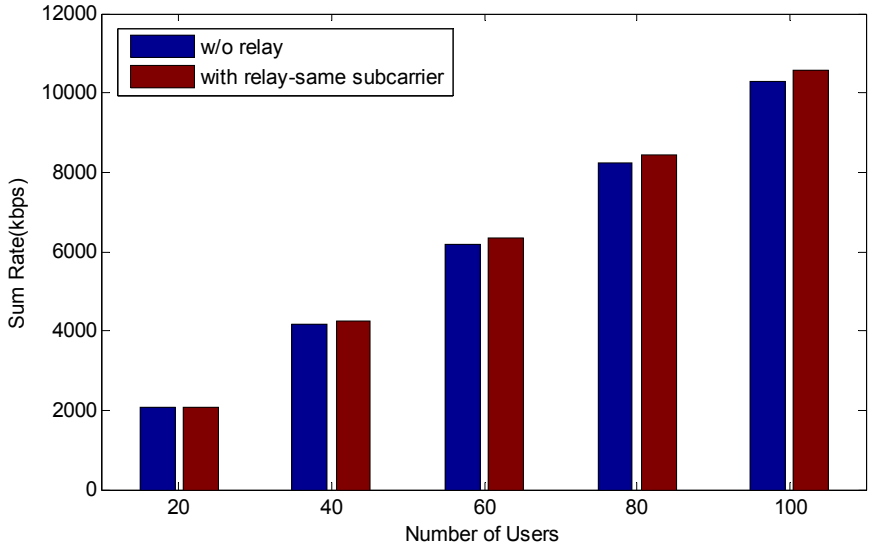


Figure 79 - Sum Rate vs Number of Users for Scenario II
$R_{th} = 168kbps$, Active inner user percentage=50%

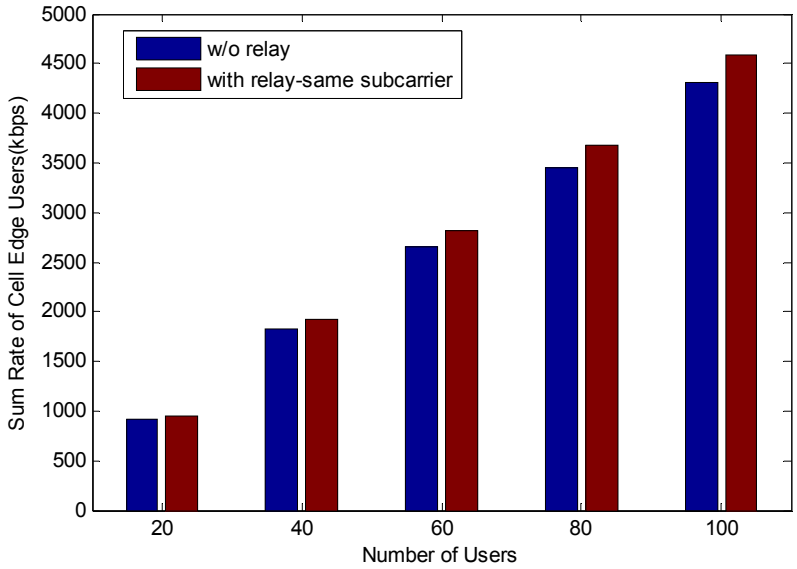


Figure 80 - Sum Rate of cell edge users vs Number of Users for Scenario II
$R_{th} = 168kbps$, Active inner user percentage=50%

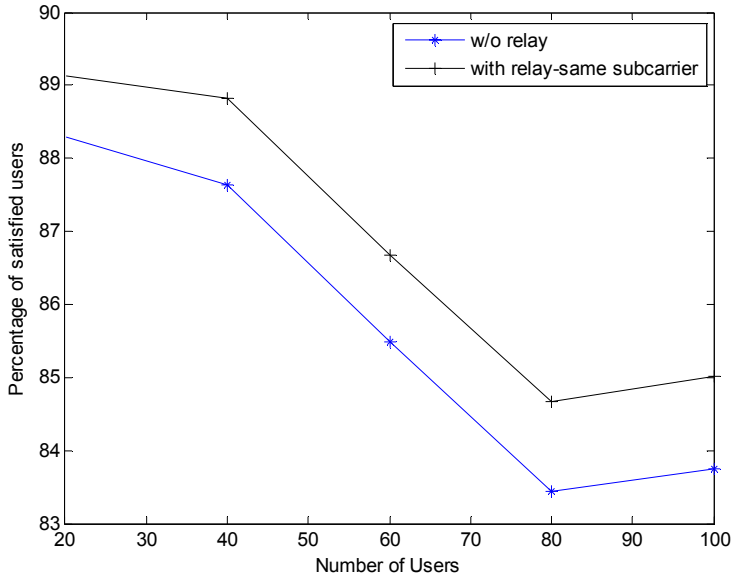


Figure 81 - Percentage of satisfied users vs Number of Users for Scenario II $R_{th} = 168kbps$, Active inner user percentage=70%

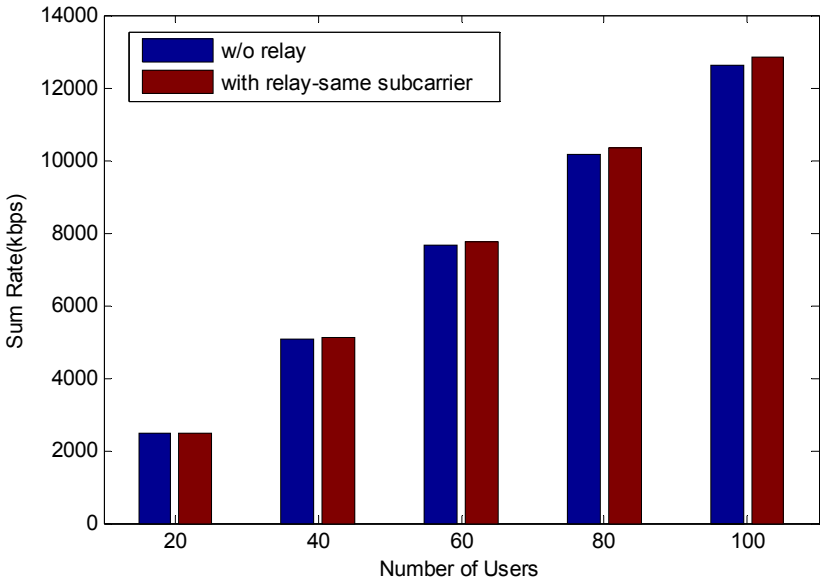


Figure 82 - Sum Rate vs Number of Users for Scenario II $R_{th} = 168kbps$, Active inner user percentage=70%

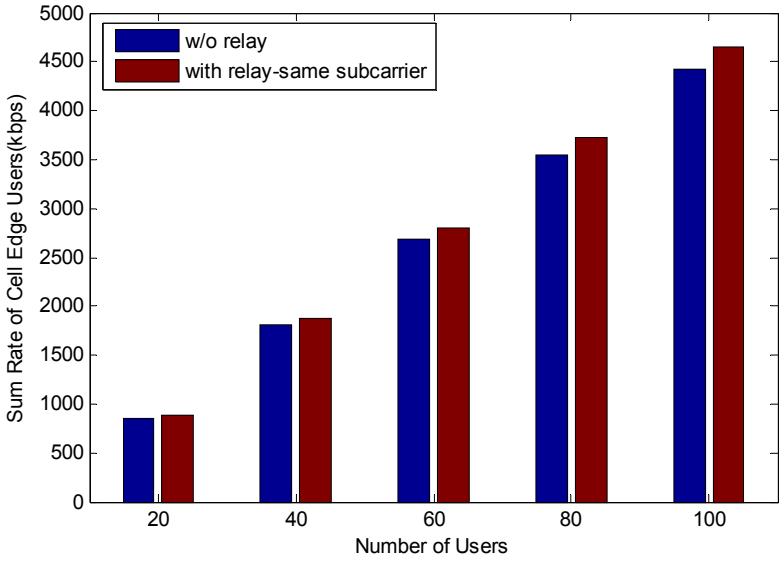
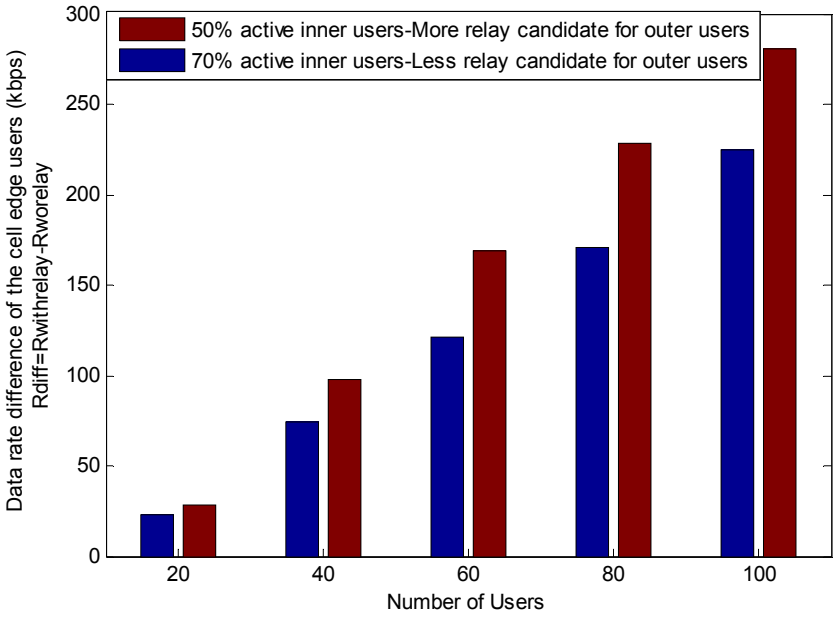
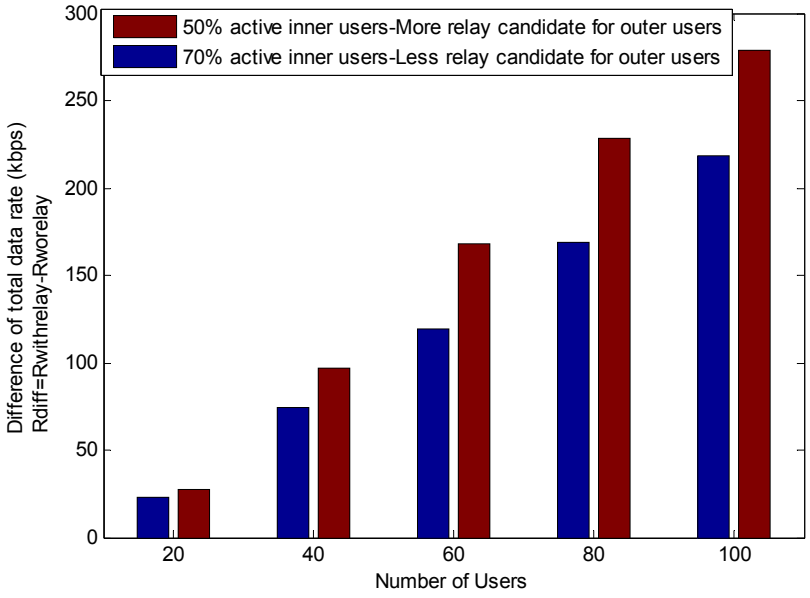


Figure 83 - Sum Rate of cell edge users vs Number of Users for Scenario II $R_{th} = 168kbps$, Active inner user percentage=70%

We also examined the effect of increasing the active inner user percentage from 50% to 70%. We have seen from the Figure 79 and Figure 82 that the total data rate is higher for 70% active inner user case. The reason is clear that the number of users close to the BS is higher in that case. Figure 83 gives us the data rate differences between the case with relay and without relay, which is an indicator of the data rate increment when relaying is used. Figure 84-a and Figure 84-b show the cell edge data rate increment and total data rate increment values, respectively. As expected, when the active inner user percentage is 50% either the celledge data rate increment or the total data rate increment is higher, since the number of relay candidates is increasing.



(a)



(b)

Figure 84 - Comparison of data rate increments for 50% and 70% active user cases

$$R_{th} = 168kbps$$

Performance results for Scenario III:

Finally, we obtained the results for the Scenario III, which focused on the Wi-Fi offloading. In this scenario, the SNR threshold value which is used to select access network is set to $0.8dB$. First of all, we located different number of Wi-Fi points at the cell edge in order to understand the effect of number of Wi-Fi points on the system performance. In this case (Scenario III-A), the outer users are not allowed to communicate over other users (relaying). In Figure 85, Figure 86 and Figure 87 the percentage of satisfied users, sum rate of the system and sum rate of the cell edge rates for 12, 24 and 36 Wi-Fi points are compared, respectively. The simulation results show that increasing the number of Wi-Fi points in the system has a positive effect on the system performance. It has also been showed in Figure 88, the number of LTE and Wi-Fi users for different number of Wi-Fi points. As expected, the Wi-Fi user ratio is increasing by adding more Wi-Fi points on the cell.

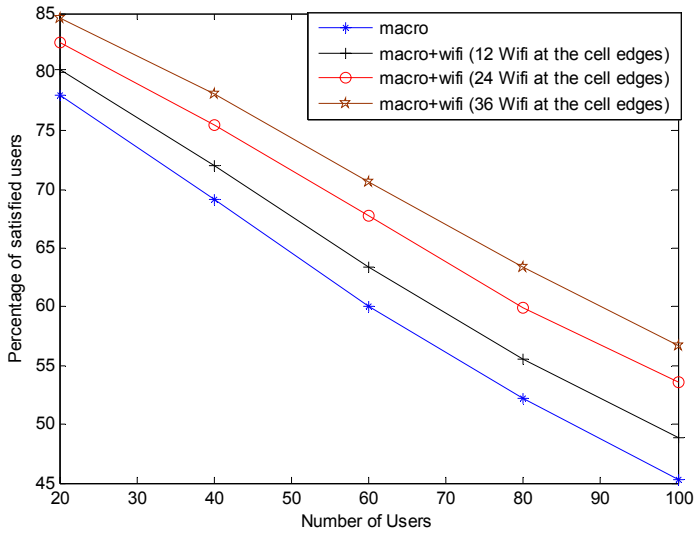


Figure 85 - Percentage of satisfied users vs Number of Users for Scenario III-A

$$R_{th} = 933kbps$$

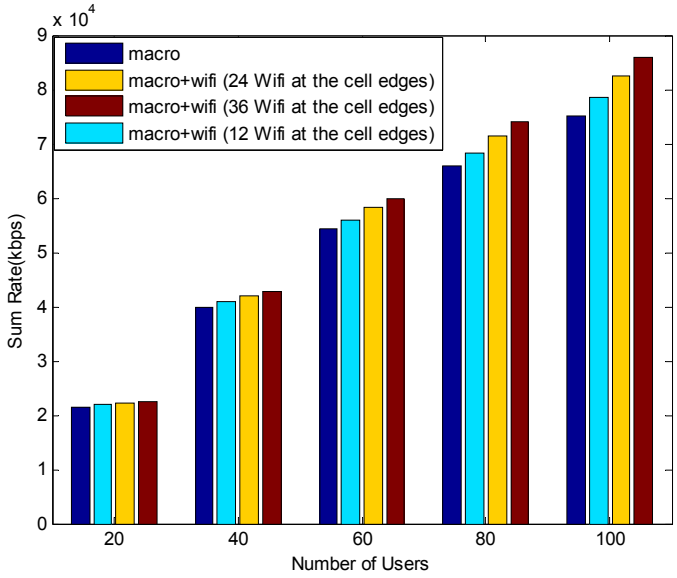


Figure 86 - Sum Rate vs Number of Users for Scenario III-A

$$R_{th} = 933kbps$$

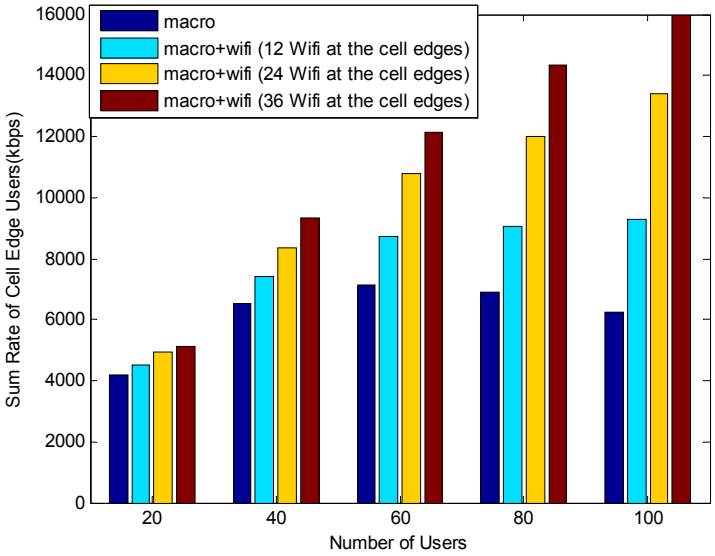
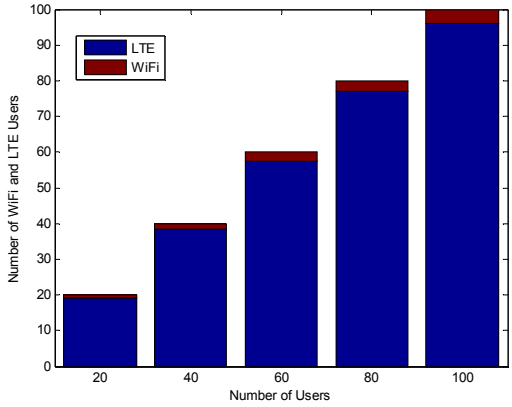
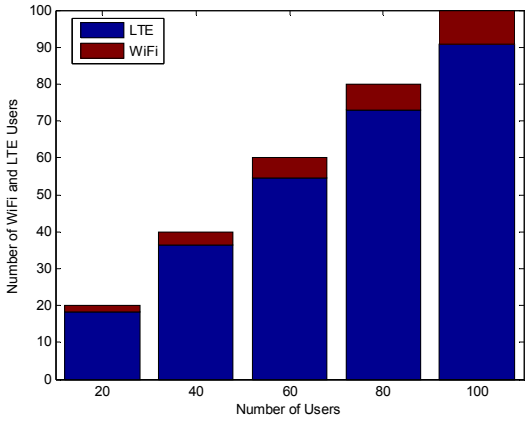


Figure 87 - Sum Rate of celledge users vs Number of Users for Scenario III-A

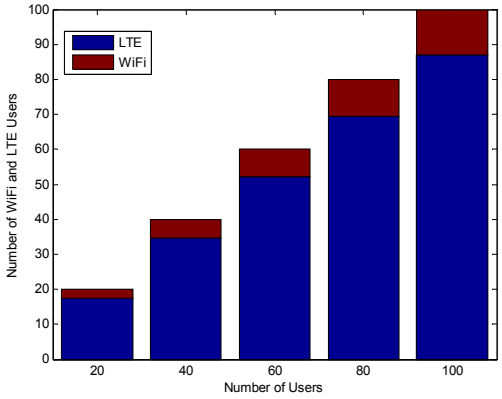
$$R_{th} = 933kbps$$



(a)



(b)



(c)

Figure 88 - Number of LTE or Wi-Fi users vs Number of Users for Scenario III-A
a) 12, b) 24, c) 36 Wi-Fi points at the cell edges

In scenario III-A, the Wi-Fi locations are also located at the outer zone uniformly and compared with the case which Wi-Fi points are located at the celledge. The results are obtained for 36 Wi-Fi points and R_{th} is set to $933kbps$. From Figure 89, Figure 90 and Figure 91, we can observe that the case which Wi-Fi APs are located at the cell edges increase slightly the percentage of satisfied users, sum data rate and sum rate of celledge users, respectively.

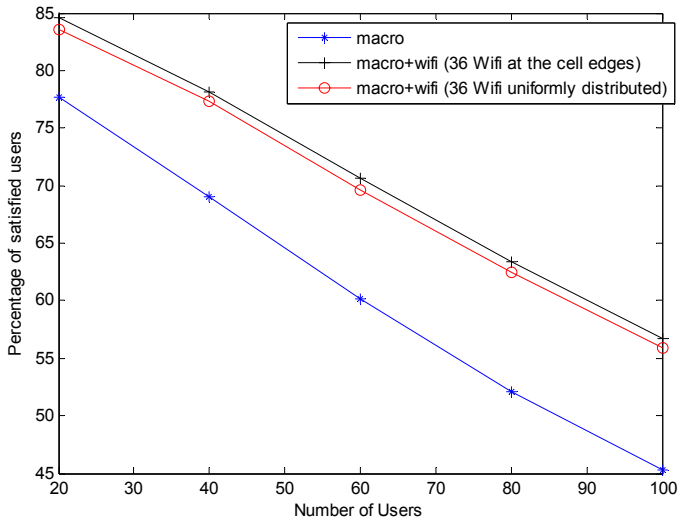


Figure 89 - Percentage of satisfied users vs Number of Users for Scenario III-A

$R_{th} = 933kbps$

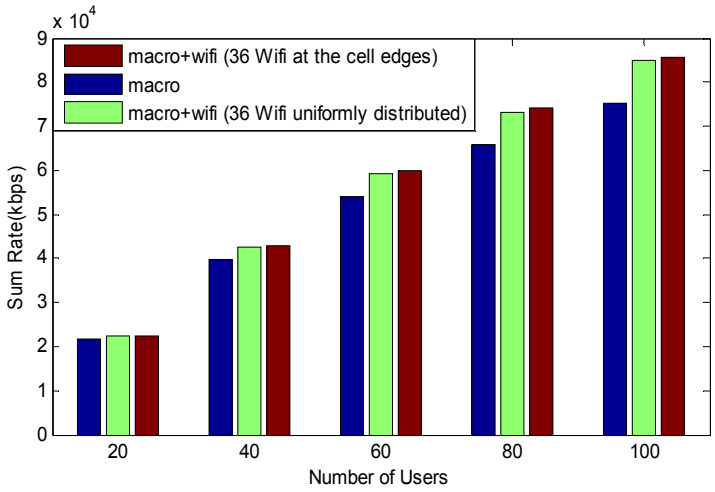


Figure 90 - Sum Rate vs Number of Users for Scenario III-A

$\boldsymbol{R_{th} = 933kbps}$

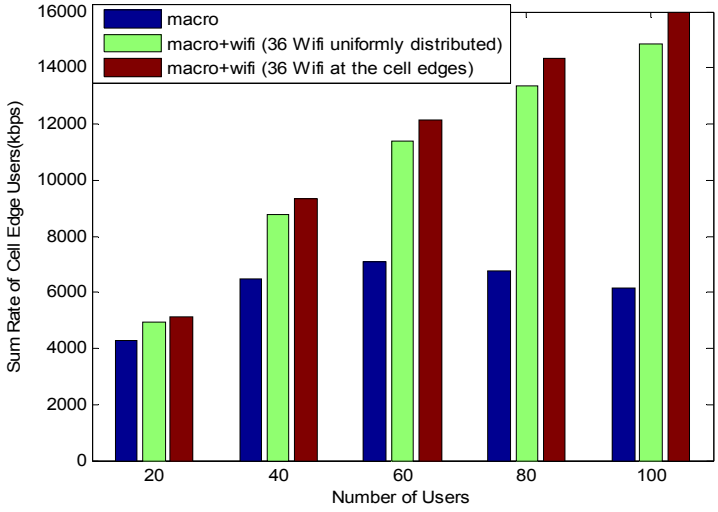


Figure 91 - Sum Rate of celledge users vs Number of Users for Scenario III-A

$\boldsymbol{R_{th} = 933kbps}$

We have extended the Scenario III-A and let the outer users to communicate with the BS over a mobile relay in Scenario III-B. Simulation results for this scenario is also obtained for $R_{th} = 168kbps$ and $R_{th} = 933kbps$. Firstly, R_{th} is set to $168kbps$ and the simulation results are obtained from Figure 92 to Figure 97 for different number of Wi-Fi points. In these figures, four different conditions are examined. In the first condition (macro), users are connected only macro BS for communication, in the second condition (macro+relay) outer users can also use other users for communication, in the third condition (macro+wi-fi) users can connect the BS or Wi-Fi access points and in the fourth condition (macro+wi-fi+relay) users can access macro or Wi-Fi and also outer macro users can use mobile relays if their direct links are not good enough. In Figure 92, Figure 93 and Figure 94, percentage of satisfied users, sum data rate and celledge data rates are compared for 12 Wi-Fi points, respectively. Macro+relay+wi-fi condition is the best among all and macro+relay condition is better than macro+Wi-Fi condition. However, when we increased the number of Wi-Fi points from 12 to 36, macro+wi-fi condition has a better performance than macro+relay condition as seen in Figure 95, Figure 96 and Figure 97. Macro+wi-fi+relay condition has also the best performance for 36 Wi-Fi points. Secondly, R_{th} is set to $933kbps$ and for 36 Wi-Fi points the simulation is repeated. It is seen from Figure 98, Figure 99 and Figure 100, macro+relay condition is getting worse for the high data rate threshold which has also been shown with

the results of Scenario I. This result affects the performance of macro+wi-fi+relay condition since the relaying is not making contribution.

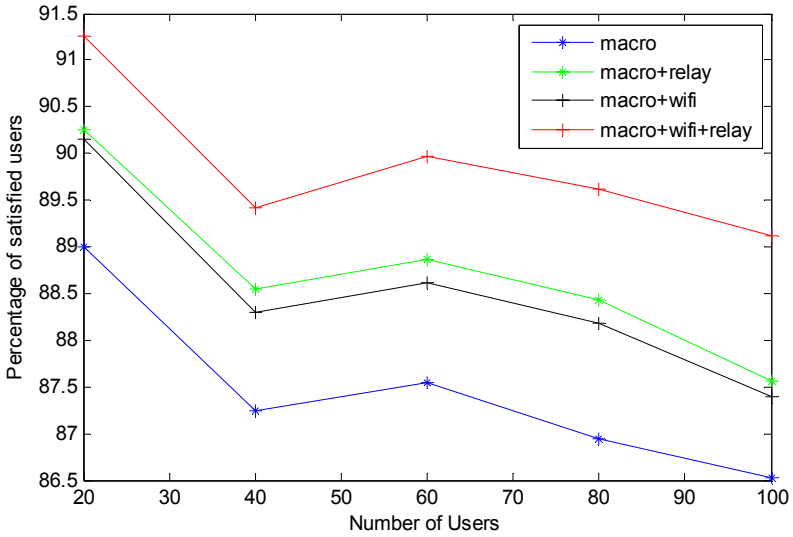


Figure 92 - Percentage of satisfied users vs Number of Users for Scenario III-B, $R_{th} = 168kbps$, AP=12

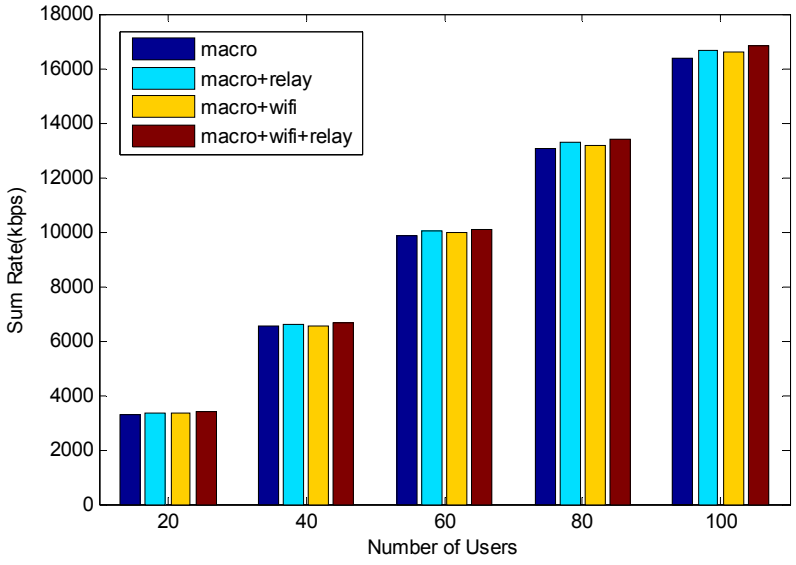


Figure 93 - Sum Rate vs Number of Users for Scenario III-B $R_{th} = 168kbps$, AP=12

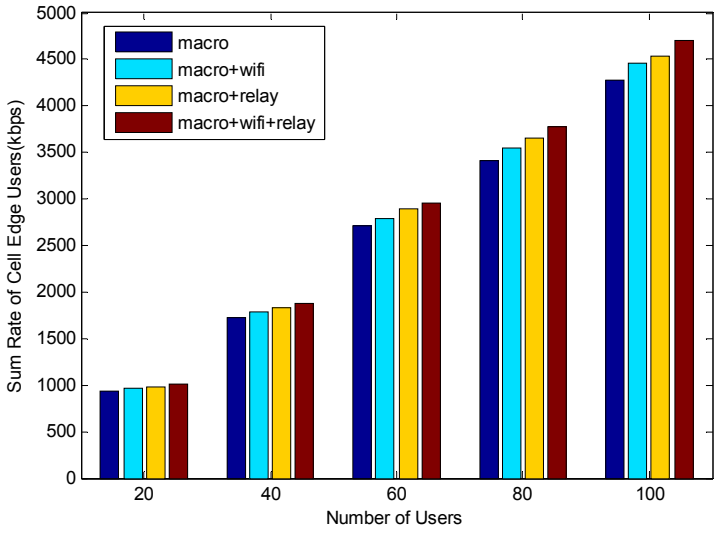


Figure 94 - Sum Rate of celledge users vs Number of Users for Scenario III-B $R_{th} = 168kbps$, AP=12

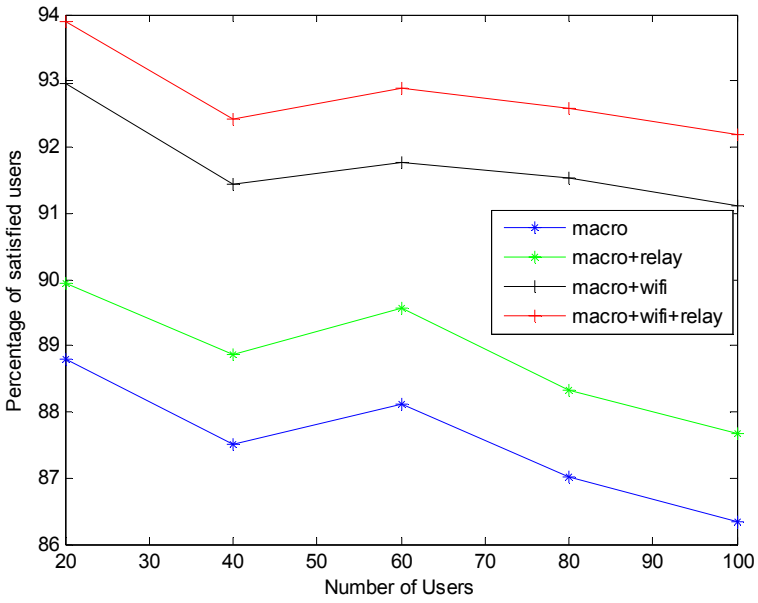


Figure 95 - Percentage of satisfied users vs. number of users for Scenario III-B, $R_{th} = 168kbps$, AP=36

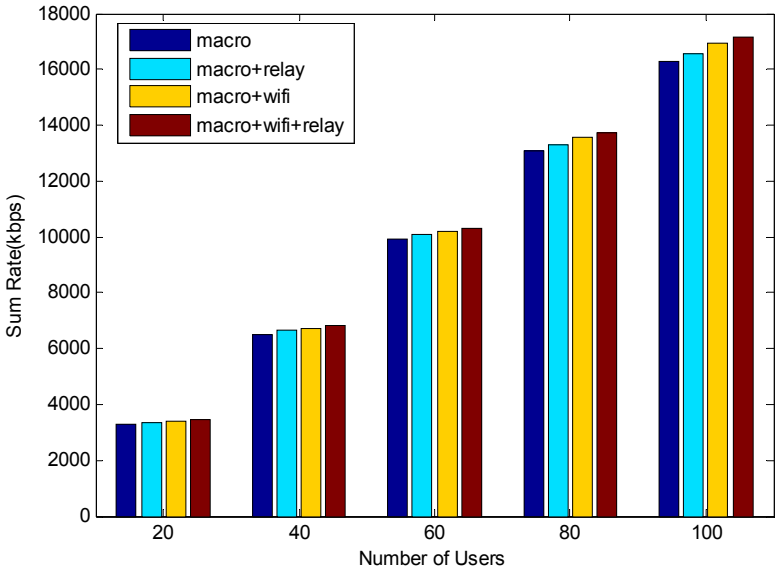


Figure 96 - Sum Rate vs Number of Users for Scenario III-B
$R_{th} = 168kbps$, AP=36

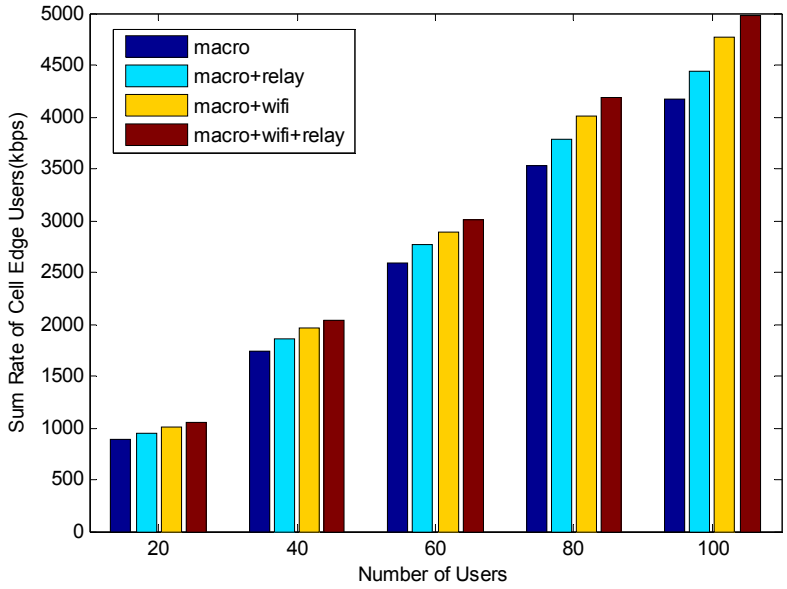


Figure 97 - Sum Rate of celledge users vs Number of Users for Scenario III-B
$R_{th} = 168kbps$, AP=36

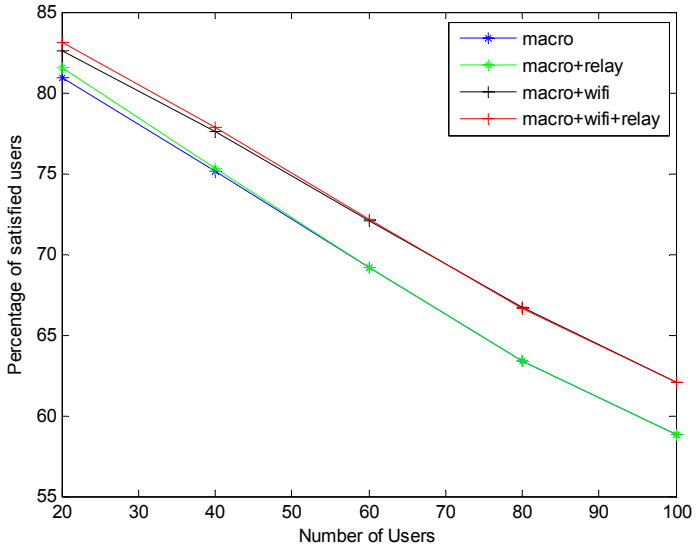


Figure 98 - Percentage of satisfied users vs Number of Users for Scenario III-B, $R_{th} = 933kbps$, AP=36

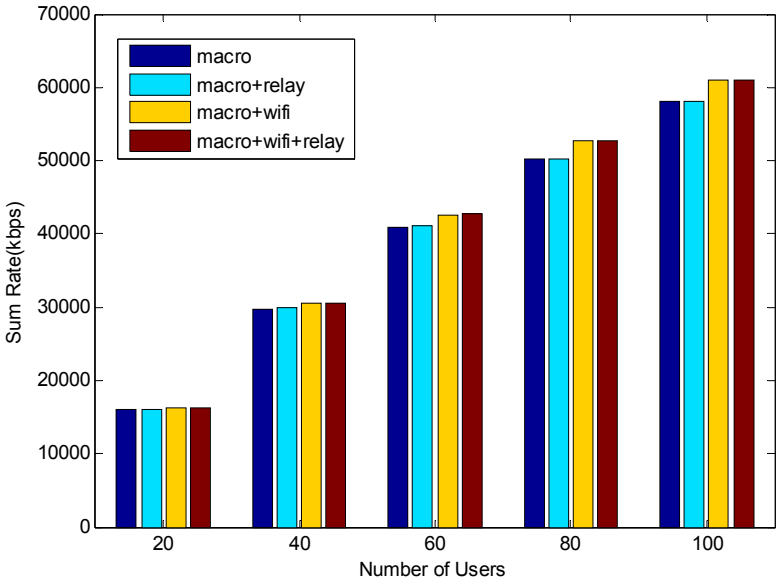


Figure 99 - Sum Rate vs Number of Users for Scenario III-B $R_{th} = 933kbps$, AP=36

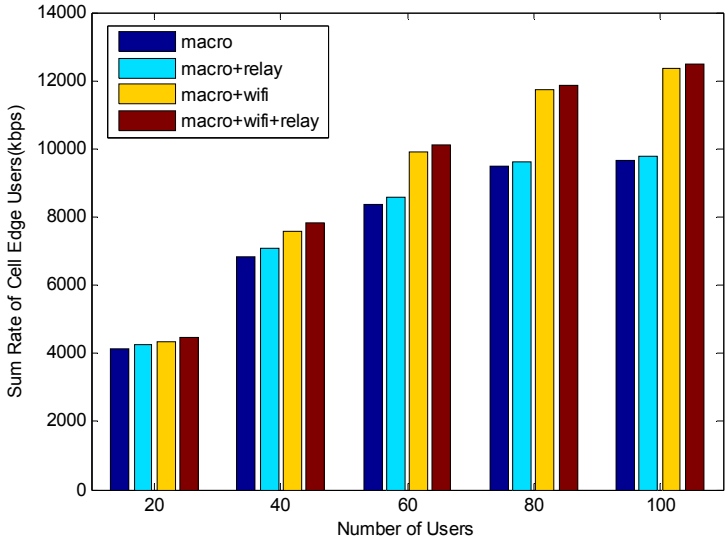


Figure 100 - Sum Rate of celledge users vs Number of Users for Scenario III-B $R_{th} = 933kbps$, AP=36

2.6.1.6 Applicability of the results

- Relaying in LTE-Advanced E-UTRAN Architecture:

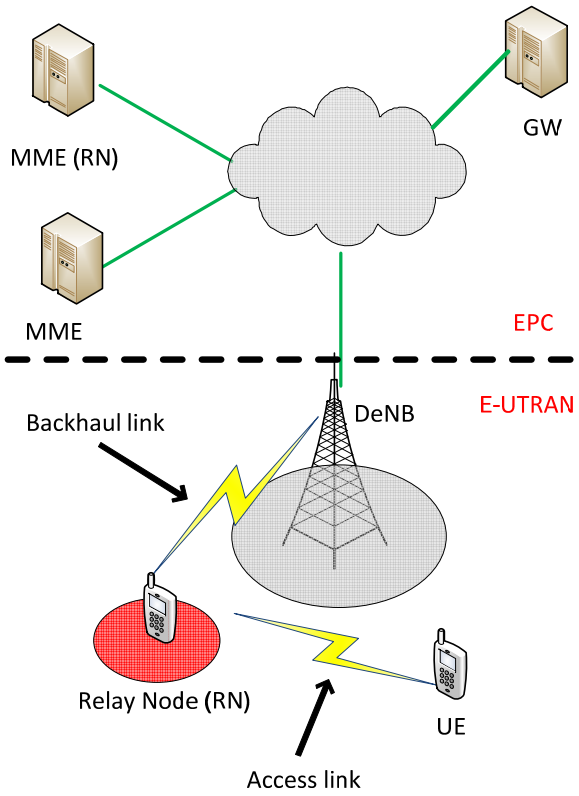


Figure 101 - Evolved Packet Core (EPC) and Evolved UMTS Terrestrial Radio Access Network (E-UTRAN).

Depending on the relaying strategy, a relay node (RN) may i) Control its own cell structure ii) Be part of the donor cell.

In the above architecture, the relay node (RN) (which is also a UE as well in our configuration) is seen as a new cell under Donor eNB (DeNB). Under S1 interface, the DeNB appears as an MME and under X2 interface, DeNB appears as an eNB to the RN. Therefore, DeNB hides the relay node (RN) that serves the UE from MMEs/GWs by providing the proxy functionalities.

In fact, DeNB acts as a gateway for RN. It creates sessions for RN and manages the EPS virtual connection for the RN, i.e. provides a transport service with specific QoS attributes. The functionality of MME (RN) is supported by MMEs.

The evolved packet core network will contain control planes MME and MME (RN) with S1 control plane (S1-c) traffic and user plane gateways with S1 user plane (S1-u) traffic.

The presented mobile relaying solution is “backhaul capacity improvement” in WP1.

- Impact of mobile relaying on EPC architecture

We envision two cases of deployment for the mobile relay assisted communication for EPC architecture. First, MME (RN) (or MME directly) and DeNB cooperation will be required. In this case, MME will store the location information of the UE and it will choose the appropriate relay for UE.

In the second case, DeNB will initiate relay signalling with target UE and the relay UE. In this case, DeNB will handle all coordination. This will simplify the load on EPC and will also increase the complexity of DeNB.

2.6.1.7 Partners involved

Turk Telekom and AVEA cooperatively contribute for simulation and design architecture of mobile relaying.

3. Conclusions

This document contains the validation results of technologies that are connected to mobility management. Mobility management not only includes handover execution mechanisms but all mobility related functions which contribute to mobility management, point-of-access selection of devices or flows and which target to achieve better resource utilization of the network.

Table 11 summarizes the challenges this work is concerned with, describes the main criteria behind the challenges, and enumerates that which technology is concerned with which challenges.

Table 11 – Challenge-technology mapping in mobility management

Challenges	Criteria behind challenges	Technologies
Keep signaling under certain levels (High)	Minimize signaling overhead due to increased mobility Minimize the number of handover events Minimize network scan phase Reduce signaling due to IP-mobility Minimize signaling overhead due to initial authentication	DMA with GTP mobile relaying ANDSF, 802.21 MIH PMIP-NEMO, SCTP, NMIP, UFA-SIP, UFA-HIP, PMIP-RO HIP DEX-AKA authentication
Improve UE multiple network access (High)	Provide optimized set of rules to the UE, operator-managed access selection	ANDSF, 802.21 MIH, Wi-Fi offload
Seamless interworking with different RANs (High)	Minimize real-time service interruption delay (e.g. CSFB LTE->3G)	**not covered**
Unoptimized routing due to anchoring (H)	Minimize path length (tackle with centralized traffic anchors due to mobility management)	PMIP-RO, UFA-HIP, UFA-SIP, DMA with GTP, SCTP, NMIP, MPTCP
Inter P-GW mobility/load balancing support (High)	Minimize data path length due to anchoring Support inter-GW mobility Better distribute userdata traffic in the network	DMA with GTP, UFA-HIP, UFA-SIP, PMIP-RO, PMIP-NEMO,
Dynamic mobility anchoring (High)	Reduce signaling due to IP-mobility management. This is part of challenge "keep signaling under certain level". Instead of always-on nature introduce something less costly in terms of active mobility contexts.	PMIP-RO, DMA with GTP, SCTP, NMIP
Reduce security setup overhead (for IPsec SA establishment) (Medium)	Minimize computational cost, memory requirement on UE and GW, AAA, HSS. Minimize authentication delay when the UE roams to a non-3GPP, unmanaged access network	HIP DEX-AKA authentication
Optimize paging and location updates (High)	Improve paging and LU procedures for multiple gateways, e.g., Minimize paging delay of IDLE mode management procedures. Provide always reachability to moving UEs, in case of DMA. These could also be part of the challenge: " keep signalling under certain level"	**not covered**
Improve the support for moving networks (High)	Minimize signaling overhead due to moving networks. Part of keep signaling under certain levels	PMIP-NEMO, HIP-NEMO
Support for user cooperation (High)	Increase LTE-A coverage and cell edge throughput, decrease power consumption (UE and BS)	mobile relaying

The validation of the proposed techologies has been split to seven topics, i.e.,

1) interface selection, access network discovery and selection
2) traffic offload
3) dynamic mobility anchoring
4) terminal-based mobility management

5) flat and distributed mobility management
6) user access authorization
7) user cooperation

The main results and future work for the validation topics are summarized in the followings.

1) This topic deals with decision and handover preparation methods for efficient load balancing and flow mapping, and the validation results will be introduced in the next version of the deliverable.

   The interest of these solutions is a gain in energy consumption and an improvement of the HO process for Wi-Fi scan. The major drawback is a need to add applications in both UE and POA (eNodeB and Wi-Fi access points) to manage this new feature.

2) This topic focuses on improving UE's multiple network access capability through operator-managed Wi-Fi offloading:

   The expected gains with operator-managed Wi-Fi technology described in Section 2.1 have been proven by the prototype that has been built during the project. We have been able to demonstrate off-load of broadband traffic from wide area radio network to Wi-Fi, provide better indoor coverage for Wi-Fi enabled devices and provide operator services tied to mobile subscription also over Wi-Fi

   Possible future functions to make this solution even more useful are 1) seamless handover, 2) handover between Wi-Fi access points, 3) forced handover and 4) load balancing. Some of the above mentioned works are discussed in different standardization groups as well.

3) This validation topic analyzed whether it worth to deal with the introduction of DMA principles for GTP-based mobility, what is the gain in selecting always the closest/cheapest distributed PGW for new traffic flows or keeping them anchored to the initial PGW.
   The results have shown that at least in cases where a GW serves a "small" amount of cells, optimization of local GW change procedures should be considered. The estimations obtained by using the traffic model is that for Dynamic Mobility Anchoring almost only one GW is used per flow whereas for the 3GPP case a fast moving UE would pass 3 to 8 GWs (depending on the number of cells per GW). This leads to the following conclusion: Only for fast moving terminals the problem of GW changes/routing optimization need to be considered (e.g. for transport systems). For these highly mobile scenarios it is worth investigating how dynamic mobility anchoring principles can be applied to the EPC.

4) This topic deals with terminal-based mobility management. The considered technologies were SCTP, NMIP and MPTCP. These technologies do not need network deployment, the necessary functionalities are implemented in the UEs and correspondent nodes.

   Regarding SCTP protocol the following conclusions can be stated. The main objectives of this work is 1) to have a robust handover mechanism which provides seamless connectivity across changes in the network by preserving communication, 2) to have a comprehensive mobility solution that addresses both change in the host's IP address and the problem of long network disconnections. In MEVICO project, SCTP is one of the terminal-based mobility protocols which do not need infrastructure in the network and are anchorless. For the initial reachability of UEs, some support from the infrastructure is required. Optimized routing and flow mobility is provided by them for the supported protocols. SCTP provides connection oriented reliable service and congestion control services like TCP. It also provides multi-streaming and multi-homing features that provide resiliency in case of path failure.

   Possible future works for SCTP are 1) integration of session layer into mobile devices (Android devices for instance), 2) conduct performance tests on mobile devices, these tests could be to measure the handover delay, the time it takes to resume the session after regaining network connection etc., 3) study the possibility of moving this session layer into transport layer in order to reduce the overhead of extra buffers and the

possible delay caused by them and 4) charging policies and gateway selection from the operators' point of view can be studied

For the NMIP protocol, the main advantage is its efficiency. As it is based on TCP and uses most of its features, this allows benefiting of all optimizations done on TCP. The difference with TCP lies only on the management of the IP address change that permits to realize fast HO without breaking the TCP connections. There are two drawbacks, the first one is the problem with the traversal of NATs and firewalls that should be improved; the second one is it should be widespread to be really useful. The two end hosts shall implement NMIP to have its features used.

MPTCP is the only one of these three protocols to manage the multi paths natively. This allows having seamless HO on one interface by the continuing use of the other interface(s). The main disadvantage is its performance, especially when the one interface has largely better throughput than the other. MPTCP implementation is quite recent, and it should improve in next years. Furthermore it has the same drawback as the two other protocols,i.e., its use is too marginal to have a real use of it.

5) This topic deals with flat and distributed mobility management protocols or their extensions aiming to provide better scalability of the network, better ressource utilization due to mobility management. All the considered technologies require functions in the network. The considered technologies were UFA-SIP, PMIP-RO and PMIP-NEMO. The UFA-SIP and PMIP-extensions are applicable in three distcinct architecture options. All of them except PMIP-NEMO solve unoptimized routing. The UFA technologies imply drastic changes to the network functions, but provide complete mobility management from handover initiation through HO preparation and ressource allocation, to HO execution, and release of resources. PMIP-RO solves route optimization and extends PMIP-based tunneling options of 3GPP. PMIP-NEMO extends PMIP with support for moving networks and requires few changes to 3GPP EPC. The following results have been achieved for the considered technlogies.

   UFA-SIP presents good performances compared to existing solutions. For SIP-based applications, UFA-SIP presents better handover delays than the use of simple SIP for mobility management within EPS+IMS architecture. For non-SIP based applications transported over SCTP in the user plane, UFA-SIP presents better handover delays and performances than the use of m-SCTP. These properties are thanks to an optimized handover procedure and UFA flat aspects.

   Around PMIPv6, two contributions, namely PMIP-RO and PMIP-NEMO, have been proposed, implemented, and evaluated.

   PMIP-RO is an extension of PMIPv6 which enable routing optimization as well as localized routing. It introduces intermediate data anchors (IAs) within or outside the EPC to offload the LMA while being able to apply common on traffic services such as traffic shaping, content filtering, lawful interception, caching, etc. When IAs are located outside of the EPC, PMIP-RO is able to achieve traffic offloading following the architecture described in LIPA, i.e., one L-GW acting as IA in the local network and possibly the MAG functionality on HeNB for instance.

   The performance evaluation has highlighted the efficiency of the optimization procedure as well as the high gain in throughput optimized data flows may found on alternative routing paths.

   PMIP-NEMO is an extension of PMIPv6 which improve the interaction between the PMIPv6 architecture (and operations) with the DHCPv6 infrastructure in order to enable the support of moving networks. The support of moving networks is specifically ensured through the support of delegation of prefixes, specified in DHCPv6, in the PMIPv6's user profiles. Channels of communications between the DHCPv6 server and the MAGs and LMA have been specified to handle those delegated prefixes and the required modification of routing tables and rules. The implementation has been validated in a real testbed.

   For both technologies (PMIP-RO and PMIP-NEMO), further integration in 3GPP procedure has to be planned. QoS enforcement and charging (BBERF and PCRF elements) are two main aspects that have very specific procedures that are not embedded in PMIPv6.

6) User access authorization deals with reduction of security setup overhead for untrusted non-3GPP access.

   The performance evaluation and suitability analysis of HIP-DEX-AKA protocol have shown that this L3 user access service is applicable for UEs with very low computational and memory resources, and requiring high security. However, the security of the IKEv2 EAP-AKA method applied in case of untrusted non-3GPP access is a bit stronger due to the support of ephemeral Diffie-Hellman protocol that provides perfect forward secrecy. IKEv2 EAP-AKA should be the applied method in use cases where there are no extra requirements regarding performance and multipath capabilities. HIP provides a secure overlay for UEs. It is important to notice that on UEs that require HIP/IPsec transport, all of the applications should be protected to avoid threats caused by applications that could by-pass the IPsec firewall.

   Possible future work for HIP DEX-AKA method is 1) to involve the AAA proxyes and server in the process of getting the authentiation vectors from the HSS to decrease the load on the HSS, 2) extend the prototype with security policy database and security associaton management

7) User cooperation deals with better radio resource utilization by enabling relaying. User cooperation through mobile relaying targets the edge users whose data rates are usually the lowest among other users. In this study we simulated the expected throughput gain of edge users capable of mobile relaying within a heterogeneous network. In the simulated heterogeneous network the edge users might have access to user-relay as well as Wi-Fi AP.
   Three different scenarios are simulated from simple to a more complex scenario. The performance of the system is compared with the standard case where no user-relay exists. The simulation results prove that the edge user performance increases with the proposed system for all scenarios. Moreover, the system performs better as the number of users increases since more user-relays become available.
   Though the system is shown to increases the throughput, the impact on the LTE-EPC has to be investigated as a future work. In the current system, the control of mobile relaying is assumed to be handled by MME. Another aspect to be studied is security since the system enables forwarding of other users' data. Additional security protocols will be necessary for secure user-relay assisted communications.

4. References

[1] ALU (ed.), “D2.1 Advanced EPC architecture for smart traffic steering”, public deliverable, MEVICO CELTIC/CP7-011, 14.11.2011.

[2] Jani Pellikka (ed.), “D2.2 Architectural EPC extensions for supporting heterogeneous mobility schemes, public deliverable”, MEVICO CELTIC/CP7-011, 10.07.2011.

[3] IETF RFC 4960; "Stream Control Transmission Protocol".

[4] IETF RFC 5061; "Stream Control Transmission Protocol (SCTP) Dynamic Address Reconfiguration".

[5] IETF RFC 3775; "IP Mobility Support in IPv6".

[6] M. Chang, M. Lee, H. Lee, "An enhancement of transport layer apporach to mobility support", LNCS, volume 3391, 2005.

[7] M. Honda et al, "SmSCTP: A Fast Transport Layer Handover Method Using Single Wireless Interface", ISCC 2007.

[8] M. Afif et al, "Radio aware SCTP extension for Handover in EGPRS", PIMRC 2006.

[9] Y.Qiao et al, "SCTP Performance Issue on Path Delay Differential", LNCS, 2007, Volume 4517/2007

[10] T.V. Lakshman, U.Madhow, "The Performance of TCP/IP for Networks with High Bandwidth-Delay Products and Random Loss", IEEE/ACM transactions on networking, VOL. 5, NO.3, June 1997.

[11] X.Wu, et al, "TCP Handoff: A practical enhancement for heterogenous mobile environments", in Proc. of ICC 2007.

[12] IETF RFC 3261, "SIP: Session Initiation Protocol".

[13] Y.Chen, K.Chiu, R.Hwang, "SmSCTP: SIP based MSCTP scheme for session mobility over WLAN/3G heterogenous networks"

[14] G. De Marco, S. Loreto, L.Barolli "Performance Analysis of IP Micro-mobility Protocols in Single and Simultaneous Movements", EUC Workshops 2005, LNCS 3823, pp. 443, 2005.

[15] IETF RFC 3550, "A Transport Protocol for Real-Time Applications".

[16] R. Fracchia et al, "A wise extension of SCTP for wireless networks", ICC 2005.

[17] G. Appenzeller et al, "Sizing router buffers", SIGCOMM 2004.

[18] "Network Simulator NS2", http://www.isi.edu/nsnam/ns/.

[19] S. Hemminger, “Network Emulation with NetEm”, linux.conf.au (LCA 2005), Camberra

[20] L. Bokor, Z. Faigl, “A Delegation-based HIP Signaling Scheme for the Ultra Flat Architecture.”, in Proceedings of the 2nd International Workshop on Security and Communication Networks (IWSCN 2010), Karlstad, Sweden, 2010, pp. 1-8.

[21] F. A. Lootsma, “Multi-criteria decision analysis via ratio and difference judgement”, ser. Applied Optimization. Dordrecht: Kluwer Academic, 1999, vol. 29.

[22] T. L. Saaty, “The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation”, McGraw-Hill, New York, St. Louis, San Francisco, 1980.

[23] 3GPP Release 10: website: http://3gpp.org/Release-10.

[24] “Architectural Considerations for Backhaul of 2G/3G and Long Term Evolution Networks’’ Cisco White paper.

[25] Z. Faigl, L. Bokor, P. M. Neves, K. Daoud, P. Herbelin: „Evaluation of two integrated signalling schemes for the Ultra Flat Architecture using SIP, IEEE 802.21, and HIP/PMIP protocols”, Computer Networks, © Elsevier B.V., ISSN: 1389-1286, DOI: doi:10.1016/j.comnet.2011.02.005, 2011.

[26] R. Kuntz, D. Sudhakar, R. Wakikawa, L. Zhang: “A Summary of Distributed Mobility Management”, IETF Internet Draft, draft-kuntz-dmm-summary-01, August 11, 2011.

[27] R. Wakikawa, R. Kuntz, Z. Zhu, L. Zhang: “Global HA to HA Protocol Specification”, IETF Internet Draft, draft-wakikawa-mext-global-haha-spec-02, September 2, 2011.

[28] L. Bokor, Sz. Nováczki, S. Imre: „A Complete HIP based Framework for Secure Micromobility”, 5th @WAS International Conference on Advances in Mobile

Computing and Multimedia, MoMM2007, ISBN 978-3-85403-230-4, pp. 111-122., Jakarta, Indonesia, 3-5 December 2007.

[29] L. Bokor, Sz. Nováczki, S. Imre: „Host Identity Protocol: The Enabler of Advanced Mobility Management Schemes", in Advanced Communication Protocol Technologies: Solutions, Methods, and Applications, Book edited by K.n Tarnay, G. Adamis and T. Dulai, Hershey: IGI Global, Information Science Reference, ISBN: 978-1-609-60732-6, pp. 247-272. 2011.

[30] T. Polishchuk, A. Gurtov. Improving TCP-friendliness and fairness for mHIP. In Infocommunications journal. 2011/I, 26-34.