QoS-aware Mobility Management in IP-based Communication Networks

Analysis, Design and Implementaion

Dissertation Zur Erlangung des
akademischen Grades Doktoringenieur (Dr.-Ing.)

vorgelegt der Fakultät für Informatik und Automatisierung
der Technischen Universität Ilmenau

von Dipl.-Ing. Esam Alnasouri

1. Gutachter: Prof. Dr.-Ing. habil. A. Mitschele-Thiel
2. Gutachter Prof. Dr. rer. nat. habil. J. Seitz
3. Gutachter PD Dr. O. Waldhorst

Tag der Einreichung: 06.05.2012
Tag der wissenschaftlichen Aussprache: 10.10.2012

Urn:nbn:de:gbv:ilm1-2012000342

QoS-aware Mobility Management in IP-based Communication Networks

Analysis, Design and Implementation

Dipl.-Ing. Esam Alnasouri

Acknowledgement

I would like first to express that I am indebted to all those who gave me the possibility to complete this thesis. I want to thank the Integrated Communication Systems (ICS) research group for giving me permission to commence this thesis in the first instance, to do the necessary research work and to use the research group laboratories and data. I have further to thank ICS members who did efforts in discussions and reviews of my work. Especially, I would like to show my appreciation to my colleague Ausama Yousef for his encouragement and help.

I am deeply indebted to my supervisor Prof. Dr.-Ing. habil. Andreas Mitschele-Thiel (Doktorvater) from the Ilmenau University of Technology. His continuous support as well as stimulating suggestions and encouragement significantly assisted me in improving my scientific research, writing skills and successfully completing my thesis. I would like also to thank Prof. Jochen Seitz and PD Dr. Oliver Waldhorst for being co-supervisor of the thesis. Especially, I would like to express my gratitude to my colleague Dr. Ali Diab for his support and encouragement.

I want to thank my parents and my brothers who have always believed in me and have supported all my decisions vorbehaltlos and my scientific activities. I must not forget the people who encouraged me in my professional life and especially Prof. Salman Ajib.

Finally, I want to thank the person who shared this adventure more closely with me, which was my constant motivation, the one with whom I lived the best years of my life, my love Ourouba. I further want to give my special thanks to my children, Yousef and Salma, whose smile gave me the power to go ahead.

Abstract

Ubiquitous access to information anywhere, anytime and anyhow is an important feature of future all-IP mobile communication networks, which will interconnect various systems and be more dynamic and flexible. The deployment of these networks, however, requires overcoming many challenges. One of the main challenges of interest for this work is how to provide Quality of Service (QoS) guarantees in such highly dynamic mobile environments.

As known, mobility of Mobile Nodes (MNs) affects the QoS in mobile networks since QoS parameters are made for end-to-end communications. Therefore, it is a challenge to develop new solutions capable of supporting seamless mobility while simultaneously providing QoS guarantees after handoffs. Addressing this challenge is the main objective of this dissertation, which provides a comprehensive overview of mobility management solutions and QoS mechanisms in IP-based networks followed by an insight into how mobility management and QoS solutions can be coupled with each other. Following the highlight of the state of art along with the pros and cons of existing approaches, the dissertation concludes that hybrid strategies are promising and can be further developed to achieve solutions that are capable of simultaneously supporting mobility and QoS, simple from the implementation point of view, efficient and applicable to future all-IP mobile communication networks.

Based on this, the dissertation proposes a new hybrid proposal named QoS-aware Mobile IP Fast Authentication Protocol (QoMIFA). Our proposal integrates MIFA as a mobility management protocol with RSVP as a QoS reservation protocol. MIFA is selected due to its capability of the provision of fast, secure and robust handoffs, while RSVP is chosen because it presents the standard solution used to support QoS in existing IP-based networks. The hybrid architecture is retained by introducing a new object, called “mobility object”, to RSVP in order to encapsulate MIFA control messages.

Following the specification of the new proposal, the dissertation also evaluates its performance compared to the well-known Simple QoS signaling protocol (Simple QoS) by means of simulation studies modeled using the Network Simulator 2 (NS2). The evaluation comprises the investigation of the impact of network load and MN speed. The performance measures we are interested in studying comprise the resource reservation latency, number of dropped packets per handoff, number of packets sent as best-effort per handoff until the reservation is accomplished and probability of dropping sessions. Our simulation results show that QoMIFA is capable of achieving fast and smooth handoffs in addition to its capability of quickly reserving resources after handoffs. Considering the impact of network load, QoMIFA outperforms Simple QoS in all studied scenarios (low- , middle- and high-loaded scenarios). With respect to the impact of MN speed, it can be observed that the impact of ping-pong effects is seen with both protocols and results in higher resource reservation latency, more dropped packets per handoff and more best-effort packets per handoff at low speeds than at higher ones. The worst impact of ping-pong effects is seen at a speed of 3 km/h when employing QoMIFA and Simple QoS, respectively. However, QoMIFA remains performing significantly better than Simple QoS under all studied MN speeds and can even properly serve MNs moving at high speeds.

Following the simulative evaluation, the dissertation estimates the signaling cost of both studied protocols with respect to the location update and packet delivery cost. Our results show that QoMIFA achieves the above mentioned performance improvements at the cost of greater location update cost and slightly higher packet delivery cost than Simple QoS.

Zusammenfassung

Der allgegenwärtige Zugang zu Informationen, jederzeit und überall, ist ein wichtiges Merkmal künftiger All-IP-Mobilfunktnetze, die verschiedene Systeme miteinander verbinden, dabei dynamischer und flexibler sein werden. Der Einsatz dieser Netze erfordert es jedoch, viele Herausforderungen zu überwinden. Eine der wichtigsten im Rahmen dieser Arbeit, ist die Frage, wie Quality of Service (QoS) Eigenschaften in solchen hoch dynamischen, mobilen Umgebungen zu garantieren sind.
Bekanntermaßen beeinflusst die Mobilität von Mobilknoten (MN) die Dienstgüte in mobilen Netzen, da QoS-Parameters für die Ende-zu-Ende-Kommunikation vereinbart werden. Daher müssen Lösungen entwickelt werden, die nahtlose Mobilität, bei gleichzeitigen QoS-Garantien nach Handoffs, unterstützen. Diese Herausforderung ist das Hauptziel der vorliegenden Dissertation, die einen umfassenden Überblick über die bestehenden Mobilitäts- und QoS-Managment-Lösungen in IP-basierten Netzen liefert, gefolgt von einem Einblick in Methoden zur Kopplung von Mobilitätsmanagement und QoS-Lösungen. Nach Betrachtung der Vor- und Nachteile bestehender Ansätze, kommt die Dissertation zu dem Schluss, dass hybride Strategien vielversprechend sind und zu praktikablen Lösungen weiterentwickelt werden können, die sowohl Mobilitäts- als auch QoS-Anforderungen auf effiziente Weise, in allen zukünftigen IP-Mobilfunknetzen erfüllen können.
Auf dieser Grundlage schlägt die Dissertation ein neues Hybrid-Protokoll, genannt "QoS-aware Mobile IP Fast Authentication Protocol" (QoMIFA), vor. Unser Vorschlag integriert MIFA als Mobilitäts-Management-Protokoll mit RSVP als QoS Reservierungsprotokoll. MIFA wird aufgrund seiner Fähigkeit zu schnellen, sicheren und robusten Handoffs gewählt. RSVP hingegen dient als Standardlösung zur Bereitstellung von QoS in bestehenden IP-basierten Netzen. Unter Einhaltung der Hybrid-Architektur wird RSVP um ein neues Objekt, genannt "Mobility Object" erweitert, welches MIFA-Kontrollnachrichten kapselt.
Nach der Spezifikation des neuen Vorschlags, bewertet die Dissertation auch seine Leistung im Vergleich zu dem bekannten "Simple QoS Signaling Protocol" (Simple QoS), mittels Simulationsstudien, modelliert mit dem "Network Simluator 2" (NS2). In der Auswertung werden der Einflusses der Netzwerklast und der Geschwindigkeit des Mobilknotens untersucht. Die hierzu verwendeten Leistungsparameter umfassen die Ressourcen-Reservierungs-Latenz, die Anzahl verlorener Pakete pro Handoff, die Anzahl der, vor Abschluss der Reservierung, mit Best-Effort-Eigenschaften übertragenen Pakete pro Handoff und die Wahrscheinlichkeit von Verbindungsabbrüchen. Unsere mittels Simulation erzielten Ergebnisse zeigen, dass QoMIFA schnelle und nahtlose Handoffs mit schneller Ressourcenreservierung nach Handoffs kombinieren kann. Unter Berücksichtigung des Einflusses der Netzwerklast, ist nachweisbar, dass QoMIFA eine besser Leistung als Simple QoS in allen untersuchten Szenarien mit geringer, mittlerer und hoher Last erreicht. Bei Betrachtung des Einflusses der Bewegungsgeschwindigkeit des Mobilknotens auf die Leistung, lassen sich unter beiden Protokollen Ping-Pong-Effekte beobachten, welche zu höheren Ressourcen-Reservierungs-Latenzen, mehr verlorenen Paketen und mehr Best-Effort-Paketen pro Handoff bei geringeren Geschwindigkeiten führen. Der stärkste Einfluss dieser Pinp-Pong-Effekte ist jeweils bei 3 km/h zu beobachten. Allerdings verhält sich QoMIFA unter allen untersuchten Bewegungsgeschwindigkeiten besser als Simple QoS und kann Mobilknoten auch bei hohen Geschwindigkeiten bedienen. In Anschluss an die simulationsgestützte Evaluierung, schätzt die Dissertation die Signalisierungskosten beider Protokolle unter Betrachtung der Kosten für Ortslokalisierung und Paketzustellung. Im Ergebnis erreicht QoMIFA die zuvor genannten Leistungsverbesserungen auf Kosten von größeren Ortslokalisierungskosten und leicht höherer Paketzustellungskosten.

Outlines

Abbreviations

2G	Second generation mobile communication networks
3G	Third generation mobile communication networks
4G	Fourth generation mobile communication networks
ADSPEC	ADvertisement SPECification object
AF	Assured Forwarding
Agnt_Adv	Agent Advertisement
Agnt_Sol	Agent Solicitation
ANG	Access Network Gateway
AP	Access Point
A-QoS	Assessed QoS
AR	Access Router
BA	Binding Acknowledgement
BS	Base Station
BU	Binding Update
CIP	Cellular IP
CN	Corresponding Node
CoA	Care of Address
DA	Domain Address
DCoA	Domain Care of Address
DiffServ	Differentiated Services
ECS	Eager Cell Switching
EF	Expedited Forwarding
E-LSP	EXP-Inferred-PSC LSP
ERO	Explicit Route Object
EXP	Experimental-bits field
FA	Foreign Agent
FEC	Forwarding Equivalence Class
FILTER_SPEC	FILTER SPECification object
FLOWSPEC	FLOW SPECification object
GCoA	Global Care of Address
GFA	Gateway Foreign Agent
GIST	General Internet Signaling Transport layer
GSM	Global Standard for Mobile communication
HA	Home Agent

HA_Ack	HA Acknowledgement
HA_Not	HA Notification
HAWAII	Handoff-aware Wireless Access Internet Infrastructure
HMIPv6	Hierarchical Mobile IP version 6
HMRSVP	Hierarchical Mobile ReSource reserVation Protocol
HMSIP	Hierarchical Mobile Session Initiation Protocol
HoA	Home Address
IETF	Internet Engineering Task Force
Int_Ack	Initial Acknowledgement
IntServ	Integrated Services
IPv4	Internet Protocol version 4
IPv6	Internet Protocol version 6
I-QoS	Intrinsic-QoS
I-TCP	Indirect- Transmission Control Protocol
L2-LD	Layer 2- Link Down
L2-LU	Layer 2- Link Up
L2-trigger	Layer 2- trigger
L3-FHR	Layer 3- Frequent Handoff Region
LCoA	Local Care of Address
LCS	Lazy Cell Switching
L-LSP	Label-only-Inferred-PSC LSP
LRSVP	Localized ReSource reserVation Protocol
LSP	Label Switched Path
LSR	Label Switching Router
LTE	Long Term Evolution
M_P_Ack	Movement Probability Acknowledgement
M_P_Not	Movement Probability Notification
Mem_Join_Resp	Member Join Response
Mem_Join_Rqst	Member Join Request
MF	Multi-Field
MIFAv4	Mobile IP Fast Authentication protocol version 4
MIPRR	Mobile IP Regional Registration
MIPv4	Mobile IP version 4
MIPv6	Mobile IP version 6
MMSP	Mobile Multimedia Streaming Protocol

MN	Mobile Node
MR	Mobile Receiver
MRI	Message Routing Information
MRSVP	Mobile ReSource reserVation Protocol
MS	Mobile Sender
MSPEC	Mobility SPECification
M-TCP	Mobile-TCP
NS2	Network Simulator 2
NSIS	Next Steps In Signaling
NSIS RMD	NSIS Resource Management in DiffServ
NSLP	Next Steps In Signaling (NSIS) Signaling Layer Protocol
NTLP	Next Steps In Signaling (NSIS) Transport Layer Protocol
OA	Ordered Aggregate
OPWA	One Pass With Advertising
OTcl	Object-oriented Tool control language
PATHErr	PATH Error
PBA	Proxy Binding Acknowledgement
PBU	Proxy Binding Update
PFA_Ack	Previous Foreign Agent Acknowledgement
PFA_Not	Previous Foreign Agent Notification
PHB	Per-Hob Behavior
PHOP	Previous hop object
P-QoS	Perceived-QoS
Pr_Rt_Adv	Proxy Router Advertisement
Pr_Rt_Sol	Proxy Router Solicitation
PSC	PHB Scheduling Class
PSNR	Peak Signal-to-Noise Ratio
QM	QoMIFA flag
QNE	QoS-NSLP Entity
QoMIFA	QoS-aware Mobile IP Fast Authentication protocol
QoS	Quality of Service
QSPEC	QoS SPECification object
R^2CP	Radial Reception Control Protocol
RA	Router Advertisement
Receiver MSPEC	Receiver Mobility SPECification

Reg_Rply	Registration Reply
Reg_Rqst	Registration Request
remote Agnt_Adv	remote Agent Advertisement
remote Agnt_Sol	remote Agent Solicitation
RFA	Regional Foreign Agent
RMD	Resource Management in DiffServ
RNCR	Receiver Nearest Common Router
RRO	Record Route Object
RTT	Round Trip Time
RSVP	ReSource reserVation Protocol
SA	Security Association
SENDER_TEMPLATE	Sender template object
SENDER_TSPEC	Sender traffic specification object
Simple QoS	Simple QoS signaling protocol
SIP	Session Initiation Protocol
SLA	Service Level Agreement
SLS	Service Level Specification
SNCR	Sender Nearest Common Router
TCP	Transmission Control Protocol
TCP-R	Transmission Control Protocol-Redirection
TE	Traffic Engineering
ToS	Type of Service
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunication System
WIMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Networks
WLRP	Wireless Lightweight Reservation Protocol

Chapter 1: Introduction

Ubiquitous access to information anywhere, anytime and anyhow is a main feature of 4th Generation (4G) mobile communication networks, which will interconnect heterogeneous systems, e.g. GSM[1], UMTS[2], LTE[3], etc., via a common IP-core. 4G networks are therefore, called all-IP and expected to be more dynamic and flexible, provide higher bandwidth at lower costs, serve fixed and mobile subscribers even those moving at high speeds, etc. However, to reach the ambitious goals of all-IP networks, many challenges must be overcome. One of the essential challenges is how to provide Quality of Service (QoS) guarantees in such highly dynamic mobile environments. Overcoming this challenge is the main focus of this dissertation, which will provide a comprehensive analysis of the problem and propose a suitable solution capable of providing QoS guarantees for mobiles moving inside such access networks.

This chapter introduces the dissertation and provides a detailed analysis of the problem statements addressed in this dissertation, followed by the objectives and contributions of this work. Chapter 1 is structured as follows: section 1.1 briefly overviews some basic principles and the architecture of all-IP mobile communication networks. Section 1.2 discusses the problem statements. The objectives and main contributions of this work are listed in section 1.3. Finally, the dissertation outline is presented in section 1.4.

1.1 All-IP Mobile Communication Networks

As known, the Internet is a collection of networks (termed as subnets as well), eachcomprises a set of interconnected hosts.Each host is identified by means of a unique IP address, which consists of two parts. The first part stands for the subnet, while the second identifies the host inside this subnet [Hal96]. The interconnection between subnets is done by means of routers operating specific routing protocol, e.g. the Internet Protocol version 4 (IPv4) [Pos81] or version 6 (IPv6) [DHi98]. Wireless IP-based access networks are constructed using principles similar to those of the Internet. The difference is that each access network consists of a set of Access Points/Base Stations (APs/BSs) serving mobile devices moving around. Furthermore, each access network interconnects with the global Internet via specific gateway(s). Notice that mobile devices are IP-based as well and should, therefore, be assigned a topology-correct IP address.

The tremendous success of the Internet, the demand for being always-on regardless of users' locations, the wide range of new up-coming applications requiring high bandwidth, etc. are the main drivers towards an all-IP architecture [Dia10]. The new architecture is defined as a collection of entities that provide a set of capabilities for the provision of IP services to users based on IP technology where various access systems can be connected [3GPP-TR], see Figure 1.1.

Achieving the ambitious goals of all-IP architecture requires overcoming several challenges, e.g. support of seamless mobility, guarantee of QoS for mobiles moving around, guarantee of a secure communication and data exchange, etc.

---

[1] Global Standard for Mobile communication (GSM) is the European standard of 2nd generation (2G) mobile communication networks [ESTI].

[2] Universal Mobile Telecommunication System (UMTS) [3GPP] is the standard of 3rd generation mobile communication systems (3G).

[3] Long Term Evolution (LTE) [3GPP-L] is the development of 3GPP's UMTS towards an all-IP architecture. LTE enhances the spectrum efficiency and reduces the control and user plane latency. It also provides lowered costs for operators and users.

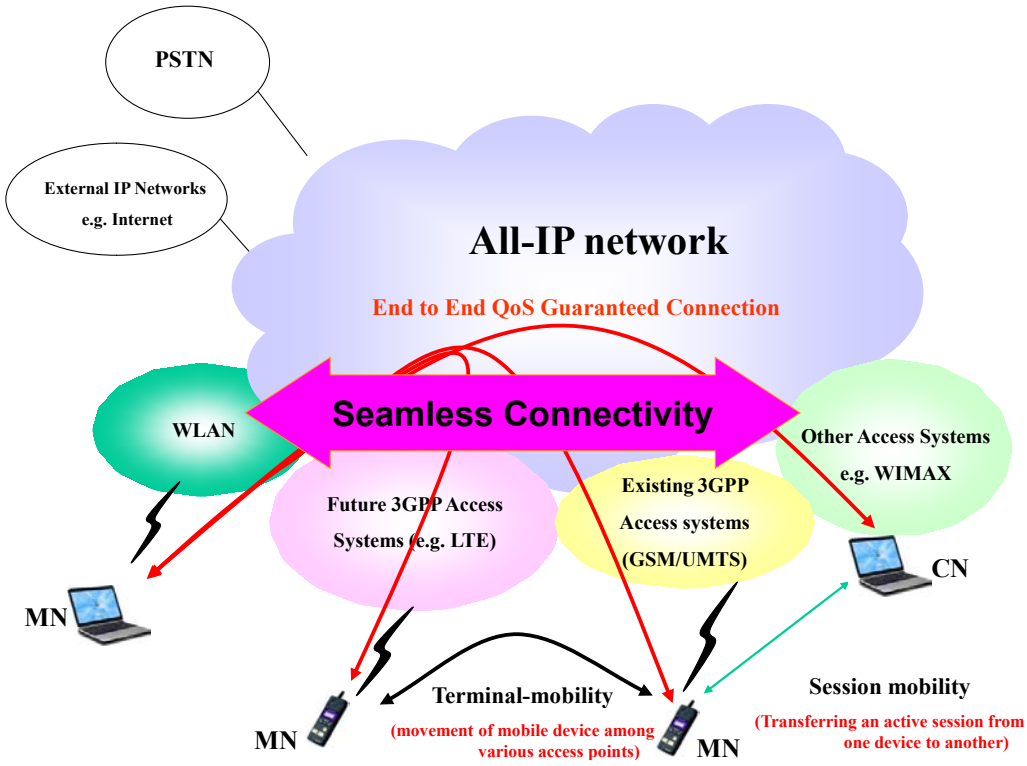


Figure 1.1: All-IP architecture [3GPP-TR]

1.2 Problem Statements

As a Mobile Node (MN) moves from one AP/BS to another, the responsibility of serving this MN will be transferred to the new AP/BS. The transfer of the responsibility from one AP/BS to another is referred to as a handoff. So as to satisfy real-time requirements during and after handoffs, three main issues should be considered.

1. First, handoff time should be minimized, so that ongoing applications do not suffer from any impairment due to the user mobility. This implies that there should be a check of resources availability and fast reservation of these resources, in case they are available, after or even during handoffs. Keep in mind that QoS parameters are made for an end-to-end communication. Many problems are encountered, therefore, due to handoffs, e.g. the availability of resources in the new location, fast and efficient resources reservation after handoffs, etc.
2. Resources that are no longer needed by the MN after the MN hands off must be released as fast as possible to be available for other users.

As mentioned in the previous section, the Internet will be a major part of future all-IP mobile communication networks and is assumed to deliver services at lower costs than currently obtained from today's mobile communication systems. Different radio access networks are interconnected with the Internet via gateways that represent IP routers. Moreover, MNs moving in the ranges of these radio access networks are, in principle, IP hosts. This implies that mobility management solutions and QoS techniques that should be employed in future all-IP mobile communication networks are IP-based. The IP address represents the point of attachment. This address is required to establish sessions between the MN and its communication partner. When the MN changes the point of attachment, a new topology-correct IP address will be assigned to it. Altering the IP address while the MN has active sessions may result in a communication disruption [Dia10]. In addition, if the MN operates services demanding reser-

vation of resources on the path between the MN and its communication partner, the availability of required resources should be checked during or even before the handoff. Moreover, the resources must be reserved after or even during the handoff as quickly as possible to minimize QoS degradations.

Mostly, mobility management and QoS problems are handled separately, although they relate to each other as we mentioned above. Therefore, it is a challenge to develop new solutions capable of supporting seamless mobility while simultaneously providing QoS guarantees during and after handoffs. Addressing this challenge is the main objective of this dissertation, which provides a detailed investigation of the mentioned problems, highlights the state of the art and proposes a new approach to simultaneously support mobility management and QoS in future all-IP mobile communication networks while satisfying the requirements of real-time applications.

1.3 Objectives and Contributions

1.3.1 Objectives

As mentioned in the previous section, this dissertation addresses how QoS can be guaranteed in all-IP mobile communication networks while considering mobility of users. More concrete, it focuses on the coupling between mobility management and QoS techniques in these networks, so that both mobility and QoS are simultaneously supported. Thus, the main goal of this work is to develop a new solution that couples well-known QoS mechanisms and mobility management solutions in a way capable of satisfying real-time requirements. Therefore, the main objectives of this work are defined as follows:

1. **Study the possibilities to couple QoS and mobility management techniques**.
2. **Development of a solution capable of providing the required QoS guarantees during and after handoffs**. The new solution should:
    a. provide mechanisms for an in-advance proof of resources availability in the new point of attachment,
    b. minimize or even eliminate the latency required to reserve resources on the new path and release the resources reserved previously on the old path,
    c. use resources efficiently,
    d. reduce signaling overhead,
    e. be scalable and flexible and
    f. assure security.

1.3.2 Contributions

Throughout the dissertation, the following contributions have been made:

1. **A brief review of mobility management solutions in IP-based networks.** More specifically, mobility management in different layers of the TCP/IP reference model is presented.
2. **A thorough overview and discussion of QoS provision mechanisms in IP-based networks**.
3. **An in-depth analysis of how QoS can be provided and, simultaneously, mobility managed** in future all-IP mobile communication networks.

4. **A comprehensive review of the approaches coupling between mobility management solutions and QoS mechanisms** along with a discussion of the pros and cons of each approach. Following the review, **a qualitative comparison of the described approaches** with respect to the tunneling problem, triangular routing problem, double reservation of resources during handoffs, passive reservation, dependency on layer 2 triggers, network topology, new nodes that should be introduced to the network, nodes that should be updated and security is provided. The comparison provides readers with a summary of the state of art that clearly highlights where the research has to focus on.
5. **Development of a robust and secure solution** capable of providing fast and seamless mobility management while guaranteeing QoS during and after handoffs. The new solution advances the state of art since it achieves performance improvements without

   a. wasting network resources,
   b. constraining the network topology or
   c. introducing new intermediate nodes.

   This solution is called **QoS-aware Mobile IP Fast Authentication protocol** (**QoMIFA**) and couples Mobile IP Fast Authentication protocol (MIFA) [Dia10] as a mobility management protocol with the resource ReSerVation Protocol (RSVP) [BZB97] as a protocol for reserving resources.
6. **Investigation of the advancement of the state of art provided by the developed solution is done** by means of mathematical analysis and simulation studies modeled in the Network Simulator 2 (NS2)[1] [NS2]. The studies **evaluate QoMIFA** compared to Simple QoS[2]. The simulative evaluation focuses on studying the impact of network load and MN speed on the performance deploying real-time traffic. The performance figures studied comprise the

   a. resource reservation latency,
   b. number of dropped packets per handoff,
   c. number of packets sent as best-effort per handoff until the reservation is accomplished and
   d. probability of dropping sessions.

   The mathematical analysis estimates the signaling cost of both studied protocols with respect to the location update and packet delivery cost.

1.4 Dissertation Outline

This dissertation is structured as follows: Chapter 2 provides a brief overview of mobility management in IP-based communication networks. The chapter briefly describes mobility management in different layers of the TCP/IP reference model and overviews the mobility management approaches necessary to follow the following chapters of the dissertation.

A review of well-known mechanisms used to provide QoS in IP-based networks is presented in Chapter 3. This chapter provides a qualitative comparison of QoS mechanisms at the end, as well. The qualitative comparison is done with respect to the type of resource reservation, scope of resource reservation, initiation of resource reservation, states stored in routers, transport protocols used, scalability, complexity, QoS guarantees that can be provided, support of mobility and security.

[1] NS2 is a widely-used discrete event simulator. It is targeted at the simulation of wired as well as wireless networks. This simulator will be briefly described in Chapter 6.

[2] Simple QoS couples Mobile IP version 4 (MIPv4) [Per02] and RSVP to support mobility and QoS. This protocol will be explained in detail in Chapter 4.

Chapter 4 provides a thorough review of the approaches for the coupling of mobility management solutions and QoS mechanisms. This chapter also provides a qualitative comparison of the described approaches with respect to the tunneling problem, triangular routing problem, double reservation of resources during handoffs, passive reservation, dependency on layer 2 triggers, network topology, new nodes that should be introduced to the network, nodes that should be updated and security.

Chapter 5 details the specification of the developed approach (i.e. QoMIFA), while Chapter 6 provides an evaluation of the new approach compared to Simple QoS by means of simulation studies modeled in NS2. The evaluation comprises the study of the impact of network load and MN speed. The performance figures studied in this evaluation include the resource reservation latency, number of dropped packets per handoff, number of packets sent as best-effort per handoff until the reservation is accomplished and probability of dropping sessions.

Chapter 7 investigates the signaling cost resulting from employing QoMIFA as compared to Simple QoS. The signaling cost investigated comprises the location update as well as packet delivery cost. The chapter also provides a study of the performance gain while taking the related cost into account.

Finally, the main obtained results and some proposals for future work are concluded in Chapter 8.

Chapter 2: Mobility Management in Mobile Communication Networks

As mentioned in Chapter 1, suitable IP-based mobility management solutions are necessary to realize one of the main goals of all-IP networks, i.e. always-on connectivity. Thus, a thorough understanding of IP-based mobility management is essential. For this purpose, an insight into the main principles and approaches of such mobility management is provided in this chapter.

This chapter is structured as follows: section 2.1 introduces basic principles of IP-based mobility management, mobility management requirements. Section 2.2 briefly reviews network layer mobility management approaches necessary to follow the next chapters of this dissertation. Finally, a conclusion of the main results is provided in section 2.3.

2.1 Overview

As mentioned in [Dia10], wireless IP-based mobile communication networks are divided into wired and wireless parts. The wireless part comprises APs that enable MNs to communicate with other fixed or mobile terminals. As a MN moves outside of the coverage area of its current AP and enters the coverage area of a new AP, the responsibility of serving the MN must be transferred to the new AP. This procedure is called a handoff and comprises three phases: handoff detection, initiation and execution. During the handoff detection phases, the MN detects that the current AP is no longer available. This triggers the initiation phase, in which a new AP should be detected. Following this, the execution phase should be performed, which includes an exchange of control messages, also termed signaling.

It is known that QoS figures (such as delay[1], jitter[2], reliability[3] and bandwidth[4], etc.) affected by users' mobility. Thus, according to [Dia10], solutions for mobility management in IP-based networks should consider the following requirements.

1. A fixed identifier of the MN should be kept regardless of MN's movements.
2. A proper interworking with IP routing and features should be guaranteed.
3. The MN should be able to be located by its communication partners after movements.
4. Applications' impairments should be minimized or even eliminated.
5. The cost for mobility support (e.g. location update cost, packet delivery cost, etc.) should be minimized.
6. No new security vulnerabilities are allowed to be introduced to the network.
7. Scalability, robustness and employability must be guaranteed by concept.

As mentioned above, when the MN moves beyond the coverage area of its current AP and enters the range of a new AP, it has to establish a new wireless link with the new detected AP. The procedures required to establish wireless links are referred to as link layer or layer 2 mobility and are implemented in the first layer of the TCP/IP reference model. Notice that although these procedures are implemented in the first layer of the TCP/IP reference model and

[1] Delay refers to the time required to forward a data packet from a certain source to a certain destination.

[2] Jitter defines the variation of data packets delay at the destination.

[3] Reliability expresses how applications can tolerate packet loss.

[4] Bandwidth represents the data transmission capability of a network.

the TCP/IP reference model does not have a layer named data link layer, they are referred to in the literature as link layer or layer 2 mobility procedures. The name comes from the ISO/OSI reference model, since both the physical and data link layer are combined to form the first layer of the TCP/IP reference model.

Mobility procedures are performed below the second layer of the TCP/IP reference model if the new and old APs are located in the same subnet. If this is not the case and, instead, a change in the subnet occurs, a higher layer mobility scheme is required to enable the MN to be reachable again from elsewhere in the Internet. This scheme can be implemented either in the Internet layer[1], transport layer or application layer. It is also possible that multiple layers cooperate to support mobility, i.e. hybrid layer mobility.

Supporting mobility functions in the network layer (i.e. Internet layer) aims at the transparency to higher layer protocols, e.g. Transmission Control Protocol (TCP) [Pos81a] and the User Datagram Protocol (UDP) [Pos80], etc. Transparency means that the protocols of the higher layer should not notice the mobility of MNs. Thus, no updates are required to these protocols. Changing the infrastructure to achieve such transparency is, however, permitted. Well-known examples include Mobile IP version 4 (MIPv4) [Per02] and version 6 (MIPv6) [JPA04], Mobile IP Fast Authentication protocol version 4 (MIFAv4) [Dia10], Mobile IP Regional Registration (MIPRR) [FJP07], Cellular IP (CIP) [Val99] and Handoff-aware Wireless Access Internet Infrastructure (HAWAII) [RLT00], etc. Network layer mobility management techniques play a major role in this dissertation. Therefore, these techniques will be described in greater detail in this chapter.

The major benefit obtained from implementing mobility management functions in the transport layer is the achievement of end-to-end mobility management without changing the infrastructure. The basic idea is to allow end-hosts to take care of mobility. More concrete, TCP and/or UDP should be updated to support mobility. Examples include the Radial Reception Control Protocol (R^2CP) [HKZ03], the Mobile Multimedia Streaming Protocol (MMSP) [MYO03], Freeze-TCP [Sch03], TCP Redirection (TCP-R) [FYT97], etc.

In the gateway-based mobility management protocols, the connection between the MN and the CN is split into two connections by means of a gateway. The first connection is between the CN and the gateway, while the second is between this gateway and the MN. In such a way, the updates required due to movements of MNs are restricted to the connection between the MN and the gateway. Examples include Indirect TCP (I-TCP) [BBa95], Mobile TCP (M-TCP) [HAg97], etc.

Application layer mobility follows different principles than those followed by the transport and network layer mobility. The main motivation is also to provide mobility support without changing the infrastructure. The basic idea is to utilize the IP telephony infrastructure and extend it to fulfill mobility requirements. Application layer mobility extends, in principle, the Session Initiation Protocol (SIP) [HSS99], [SRo99], which allows users to establish sessions containing multiple media streams between each other. The extension associates each permanent user identifier with a temporary IP address or host name, see [SWe00]. Well-known example is the Hierarchical Mobile Session Initiation Protocol (HMSIP) protocol [VPK03], [VPK03a].

Notice that all principles and approaches presented above are single layer-based. It is, however, known that each layer contributes either positively or negatively to mobility support. So, it makes sense to let more than one layer contribute to support mobility efficiently through the development of multi-layer mobility management approaches. The basic idea is to enable

[1] The solutions implemented in this layer are referred to in the literature mostly as network layer or layer 3 mobility management solutions, see [Dia10].

cross-layer signaling interaction, so that more than one layer can interact to support mobility. Examples of cross-layer architectures are presented in [WAb03] and [WMa03]. A more detailed description of the above presented approaches and principles can be found in [Dia10]:

2.2 Network Layer Mobility Management

As mentioned in section 2.1, network layer mobility management aims at the transparency to higher layer protocols and allows, for the benefit of this purpose, some updates to network topologies. The approaches supporting mobility in the network (i.e. Internet) layer are categorized into terminal- and network-based solutions based on whether MNs are involved in mobility activities or not, see [Dia10].

Terminal-based mobility management solutions say that MNs must contribute to allow for further provision of their services while moving. In other words, they must support mobility and, thus, must install specific software if mobility is desired.

On the contrary to terminal-based mobility management, the basic principle of network-based mobility management is the excluding of MNs from mobility management procedures. Access networks must execute all mobility procedures on behalf of MNs, which, therefore, do not require mobility support. As a result, all nodes with legacy IP stacks can be mobile.

The solutions belonging to both terminal- and network-based categories are broadly classified into macro and micro mobility management solutions based on whether handoffs are processed globally or locally. In the following only an insight into the approaches necessary to follow the next chapters of the dissertation will be provided. For more details regarding mobility management issues, the reader is referred to [Dia10].

2.2.1 *Mobile IP*

As stated in [DMA04], MIP is a well-known standard developed by the Internet Engineering Task Force (IETF) [IETF] to enable the MNs supporting the TCP/IP reference model to be mobile. It comes in two versions, namely MIPv4 and MIPv6.

Network topology: MIPv4 introduces two new entities to the network, namely a Home Agent (HA) and a Foreign Agent (FA). The HA is a router located in the MN's home network and is responsible for authenticating, authorizing and serving the MN inside the home network as well as forwarding the data packets to the MN when residing outside of the home network. The FA is a router controlling the subnet currently being visited by the MN. MIPv6 has a similar network topology. However, no FAs are required since normal IPv6 Access Routers (ARs) are sufficient. Figure 2.1 shows an example network topology of MIP. Notice that the communication partner of the MN is called a Corresponding Node (CN).

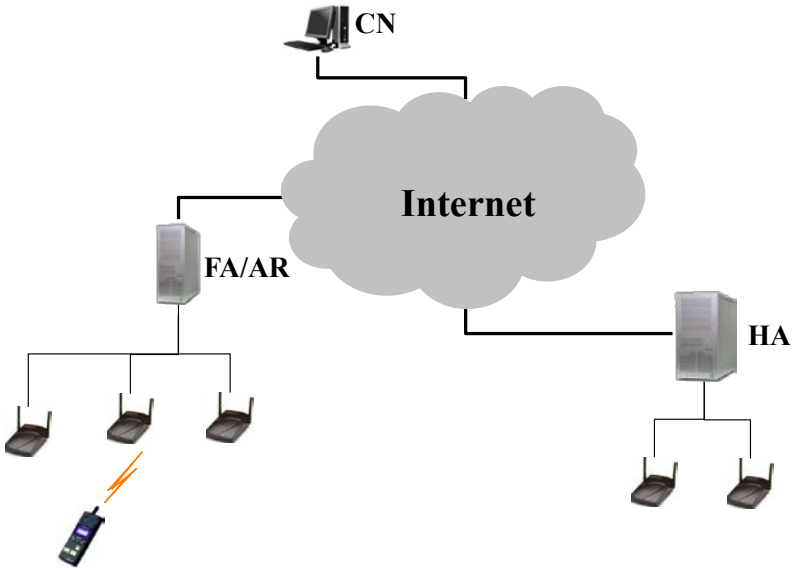


Figure 2.1: An example network operating MIP

Basic idea: the basic idea of MIP is to assign each MN two IP addresses. The first is permanent and called Home Address (HoA), while the second is temporal and referred to as Care of Address (CoA). The HoA is the address of the MN in its home network and acts as a unique identifier of the MN. On the contrary, the CoA is an address representing the current point of attachment. The CoA is registered with the HA, so that the HA always knows where to find the MN. When a CN wants to communicate with a MN, it transmits data packets to the MN's HoA. These data packets pass the HA based on the standard IP routing. The HA in turns intercepts, encapsulates and forwards the packets to the CoA. Changing the CoA always necessitates a re-registration with the HA.

Operation overview of MIPv4: each FA advertises its properties by means of a periodic Agent Advertisement (Agnt_Adv) messages. As the MN moves outside of the range of its current AP, it must first establish a new wireless link with a new, adequate AP. After or even during the establishment of this link, the MN must check whether it is still in the same subnet or not. This is normally done advertisement-based. The basic idea is pretty simple and says that the receipt of Agnt_Adv messages from a new FA means that the MN has moved to a new subnet.

Once the MN detects that it has changed its subnet and is assigned a new CoA, it notifies it's HA of the newly acquired CoA. For this purpose, the MN transmits a Registration Request (Reg_Rqst) message to the HA, which authenticates and authorizes the MN. Following this, the HA replies a Registration Reply (Reg_Rply) message to the MN. Following a successful registration with the HA, the MN will be available again. The handoff procedure of MIPv4 is shown in Figure 2.2.

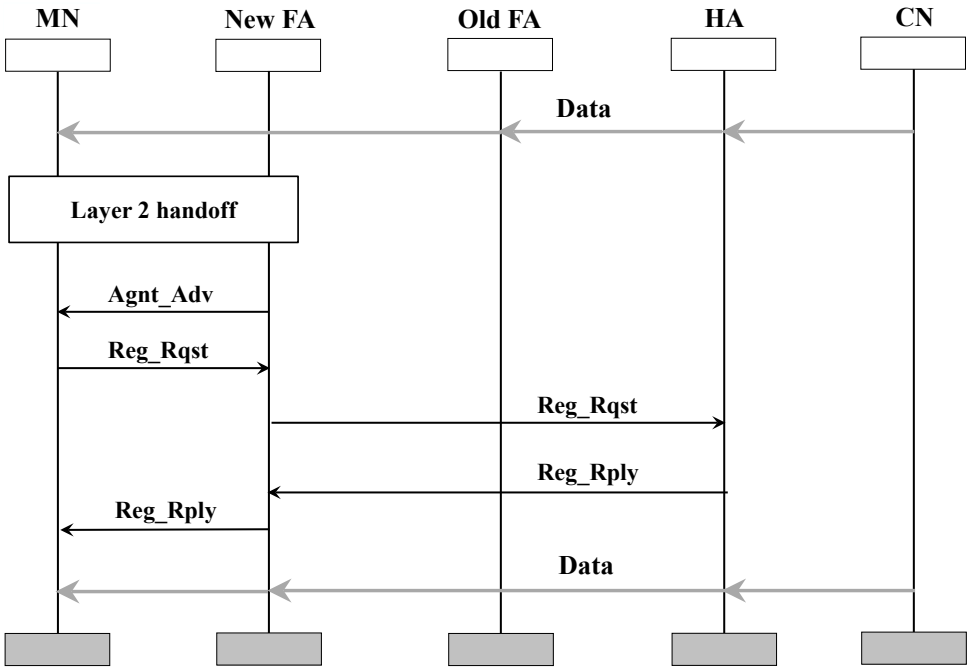


Figure 2.2: MIPv4 handoff procedure

Operation overview of MIPv6: in general, MIPv6 functions similar to MIPv4. After the MN receives a Router Advertisement (RA) message from a new AR, it configures a new CoA and notifies the HA and the CN of the newly acquired CoA. This is done by exchanging Binding Update (BU) and Binding Acknowledgement (BA) messages with both the HA and the CN, see Figure 2.3.

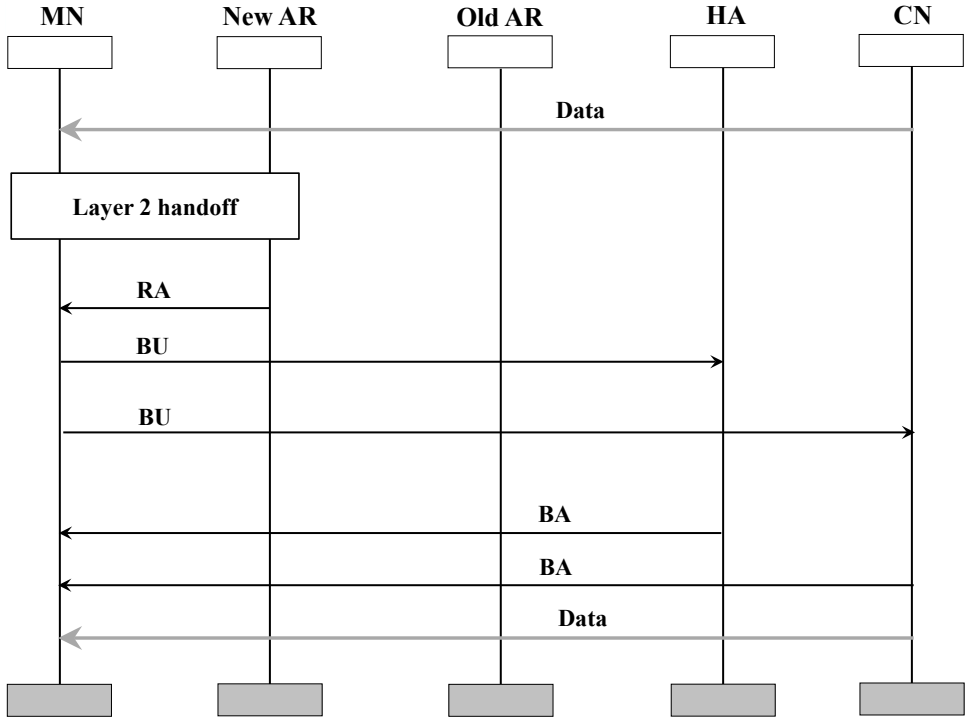


Figure 2.3: MIPv6 handoff procedure

As a CN wants to communicate with a MN, it first looks for a valid mobility binding for this MN. If no valid binding is found, the CN transmits data packets towards the MN's HoA. The

HA intercepts the packets and tunnels them to the MN's current CoA. On the contrary, if the CN has found a valid mobility binding, it transmits the packets directly to the CoA.

Pros and cons: the main advantage of MIP is the allowance of the nodes that implement the TCP/IP reference model to be mobile. This protocol is, however, not capable of providing seamless handoffs. Keep in mind that the HA, and possibly the CN as well, must be notified of each change in the point of attachment. Clearly, this produces a considerable latency and signaling, as well, especially if the HA is far away. Therefore, MIP is stated to be suitable for supporting global mobility. A further drawback of MIP is the triangular routing, since data packets forwarded to the MN are routed in most cases to the HA. Triangular routing can be seen clearly when MIPv4 is used. For MIPv6, this problem exists, as well. However, as the route optimization is part of the specification of MIPv6, triangular routing is less critical than by MIPv4. Notice that the triangular routing implies tunneling of data packets, which also causes more overhead. In addition to the long handoff latency, considerable signaling and triangular routing, MIPv4 suffers from the ingress filtering problem. Keep in mind that the MN uses its HoA for the uplink traffic. Clearly, this IP address will be detected as a topologically-incorrect IP address and will be dropped, if the FA operates an ingress filter. The ingress filtering problem does not appear when using MIPv6 due to the use of the configured CoA as a source address for uplink traffic.

2.2.2 Regional Registration for MIPv4 (MIPRR)

Network topology: MIPRR [FJP07] extends the principles of MIP and deploys a hierarchical network topology containing two or more hierarchy levels. A Gateway Foreign Agent (GFA) forms the top hierarchy level and controls the whole domain. The FAs providing IP connectivity form the undermost hierarchy level. Nodes supporting mobility and located in the access domain between FAs and the GFA form one or more hierarchy level(s) between the FAs and the GFA. These nodes are named Regional Foreign Agents (RFAs).

Basic idea: each MN is assigned two CoAs additional to the HoA. The first CoA is registered with the HA when the MN moves into a new domain. This CoA is called a Global CoA (GCoA) and is normally the address of the GFA. Notice that the GCoA does not change as long as the MN moves inside the same domain. The second CoA is termed Local CoA (LCoA) and determines the current point of attachment inside the domain. The LCoA is changed each time the MN changes the point of attachment and is updated at the GFA or possibly a RFA.

Operation overview: the operation of the protocol is relatively simple and can be summarized as follows: as the MN is switched on or moves into a new domain, it registers the GCoA with the HA employing MIPv4, as Figure 2.4 shows. The registration process is termed inter-domain mobility management.

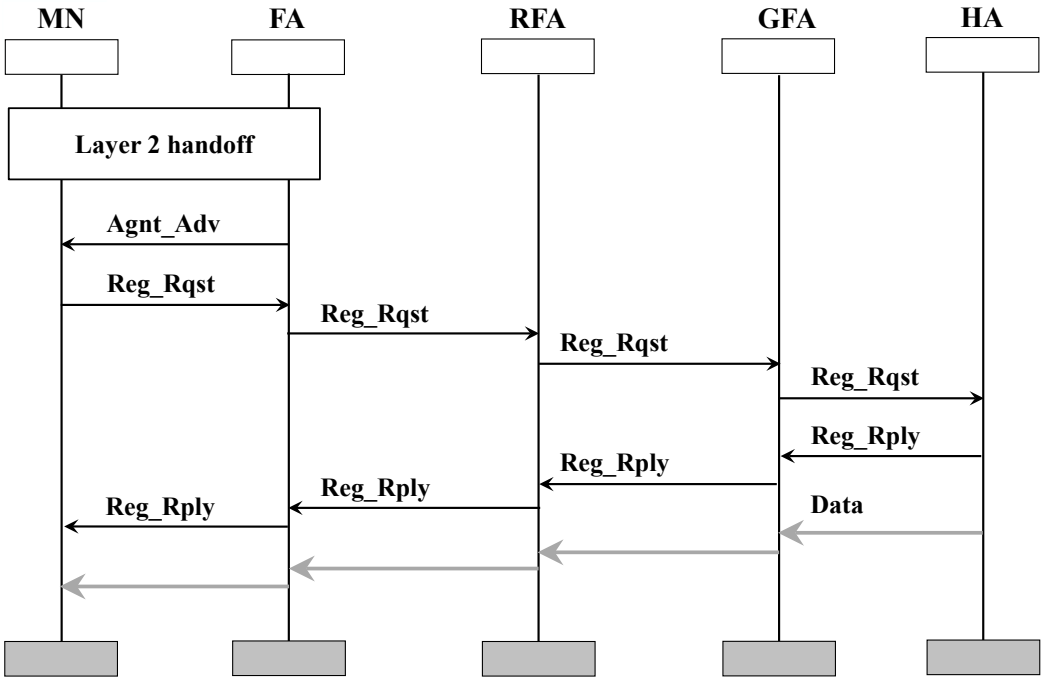


Figure 2.4: Inter-domain mobility management employing MIPRR

Tracking of movements inside the domain is handled locally without notifying the HA. This is done by exchanging a regional Reg_Rqst and a regional Reg_Rply message with the cross-over node, which can be either a RFA or the GFA. This kind of movement is called intra-domain mobility management, see Figure 2.5 (the crossover node assumed to be a RFA).

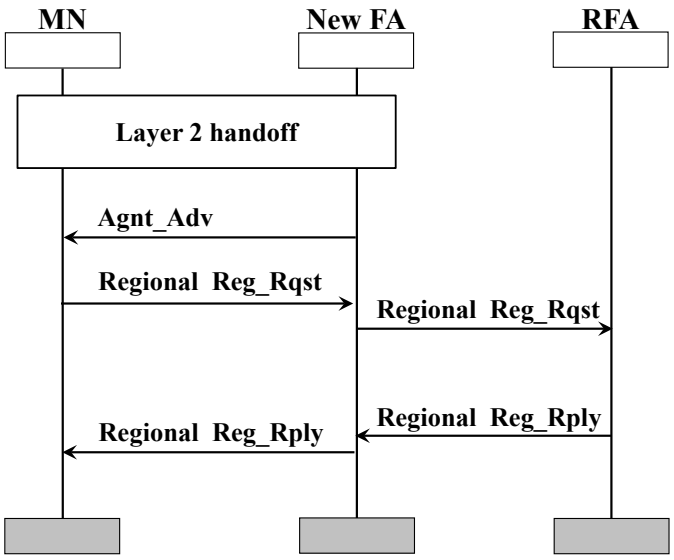


Figure 2.5: Intra-domain mobility management employing MIPRR

As mentioned while discussion of MIP, when a CN wants to communicate with a MN, it transmits data packets to the HoA of the MN. These packets bypass the HA based on the standard IP-routing. The HA in turn intercepts and tunnels the packets towards the CoA of the MN. Notice that the CoA is the address of the GFA when employing MIPRR. Assuming that there exist only one hierarchy level of RFAs between the GFA and FAs, tunneled packets will be de-tunneled and re-tunneled to the RFA serving the MN. The RFA further de-tunnels and

re-tunnels the packets towards the FA whose serves the MN currently resides inside its range.The FA de-tunnels the packets and forwards them to the MN. Notice that the MN receives the packets as they originate from the CN.

Pros and cons: MIPRR handles movements of MNs locally as long as MNs remain in the same domain. Of course, this reduces the handoff latency, number of lost packets, etc. as compared to MIP. That MIPRR requires a hierarchical network topology can be seen as a drawback since this prohibits employing MIPRR in existing topologies where such a hierarchy does not exist. Moreover, the construction of the required hierarchy implies the introduction of new intermediate nodes to handle the mobility locally. A main drawback of MIPRR is the single point of failure, since any crash or error in the GFA strongly degrades performance and may prohibit offering the mobility service. The security is not considered in the protocol. There should be mechanisms to secure the control messages exchanged between the HA and the GFA as well as between the MN and the GFA/RFAs.

2.2.3 Mobile IP Fast Authentication Protocol

So as to avoid the problems of MIP and achieve seamless handoffs, MIFA has been developed. This protocol enables fast resuming of communication without waiting for the registration with the HA, and possibly the CN, to be completed [DMA04].

Network topology: the network topology deployed when using MIFA is the same as used by MIP.

Basic idea: MIFA uses a simple idea saying that movements of MNs from any subnet are normally limited to a small set of neighboring subnets. In other words, each subnet can construct a group of candidate subnets, to which the MNs currently served by the subnet may move. These groups are termed Layer 3-Frequent Handoff Regions (L3-FHRs). Thus, providing neighbor subnets in advance with knowledge of the possible incoming MNs aids in accelerating the network layer handoff. More concrete, the authentication and authorization of MNs is delegated to these subnets, so that MNs must only contact the server controlling the new subnet to resume their communication following the handoff.

Operation in IPv4 overview: MIFA can be operated in two modes, namely reactive and predictive mode. The predictive mode realizes the make-before-break principle and is possible for MNs that are capable of receiving signals from more than one AP at the same time. The reactive mode is based on the break-before-make principle. It is possible for the MNs capable of listening to only one AP at the same time. The predictive mode is the default operation mode. If the predictive mode could not be operated, the reactive mode is fired.

The operation of MIFA can be briefly described as follows: when the MN is powered on, it establishes a wireless link first and then waits for an Agnt_Adv message. The Agnt_Adv message follows the specification of MIP. One bit from the reserved bits in the message (termed as MI flag) is utilized as a flag indicating the support of the MIFA protocol.

After the MN receives an Agnt_Adv message with the MI bit set, it first checks whether the message has been issued by the HA or a FA. Let us suppose that the MN is located in the range of a foreign subnet. The MN proceeds with the initial registration procedure, which is done using MIP. This procedure implies exchanging Reg_Rqst and Reg_Rply messages with the HA. During the initial registration procedure, two Security Associations (SAs) are generated. One is used to secure the control messages exchanged between the HA and the current FA (the key is $K1_{FA,HA}$), while the other is between the MN and the current FA (the key is $K1_{MN,FA}$). How the keys are generated and distributed is described in detail in [Dia10]. In addition to these both SAs, a third SA ($K2_{MN,FA}$), along with two random variables (R_1' and

R_2'), are generated and distributed to the MN. $K2_{MN,FA}$ will be used to secure the control messages that will be exchanged between the MN and the next new FA, with which the MN will associate after the next movement. It should be noticed that in most cases the new FA will be a member of the L3-FHR of the current FA. Figure 2.6 shows the initial registration procedure described above.

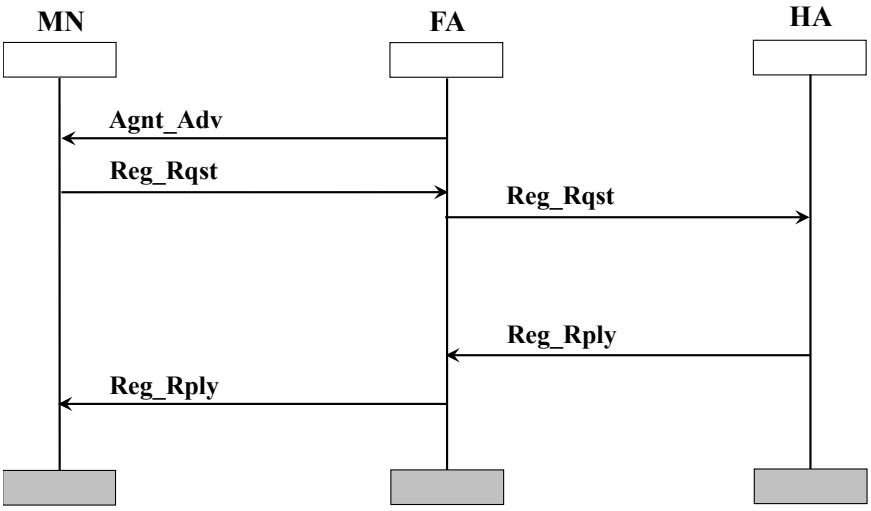


Figure 2.6: MIFA initial registration procedure

Following the initial registration procedure, the current FA executes the **initial authentication exchange** procedure to obtain the data required to authenticate and authorize the MN during the next registration with the subsequent new FA. The procedure is executed as follows: first a SA (the key is $K2_{FA,HA}$) is generated. This SA will be used to secure the control messages that will be exchanged between the HA and the next new FA, which will host the MN after the next movement. The newly generated key and random variables (R_1' and R_2') are transmitted to the HA by means of a Movement Probability Notification (M_P_Not) message.

As the HA receives the M_P_Not message, it generates two authentication values, referred to as $Auth_1$ and $Auth_2$. The authentication values are calculated by applying a hash algorithm on the random variables and some additional information related to the MN. For more details about how these values are generated, the reader is referred to [Dia10]. $Auth_1$ is what the MN has to send with the Reg_Rqst message during the next registration, while $Auth_2$ is the value the HA has to generate and send with the Reg_Rply message. Figure 2.7 shows the initial authentication exchange procedure illustrated above. Notice that the initial registration and initial authentication exchange procedures are executed once (only after the MN is switched on or wants to connect to the Internet for the first time).

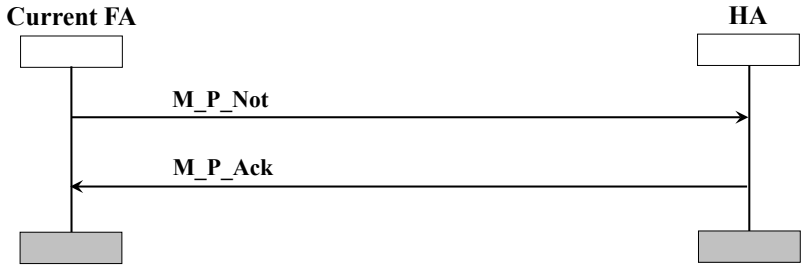


Figure 2.7: MIFAv4 initial authentication exchange procedure

The normal operation of MIFAv4 in all handoffs following the initial registration and initial authentication exchange procedures implies executing two main procedures, namely the **information distribution and handoff procedure**. The information distribution procedure aims at notifying the neighbor FAs of the possible incoming MN. This procedure is simple and implies sending a M_P_Not message to each member of the current L3-FHR. The M_P_Not message contains all information neighbor FAs require to authenticate the MN. This information is termed **MN-specific data.**

The handoff procedure in reactive mode is shown in Figure 2.8. After the MN moves out of the area of its serving FA and receives an Agnt_Adv message from a new FA, it transmits a Reg_Rqst message to this new FA. The Reg_Rqst message contains a MIFA authentication extension that includes the authentication value $Auth_1$. Once the new FA receives and successfully authenticates the Reg_Rqst, it compares the authentication value $Auth_1$ sent from the MN with the value of $Auth_1$ that the HA has previously calculated (notice that this value has been distributed as part of the MN-specific data). The MN is trusted by the HA, if the two authentication values match.

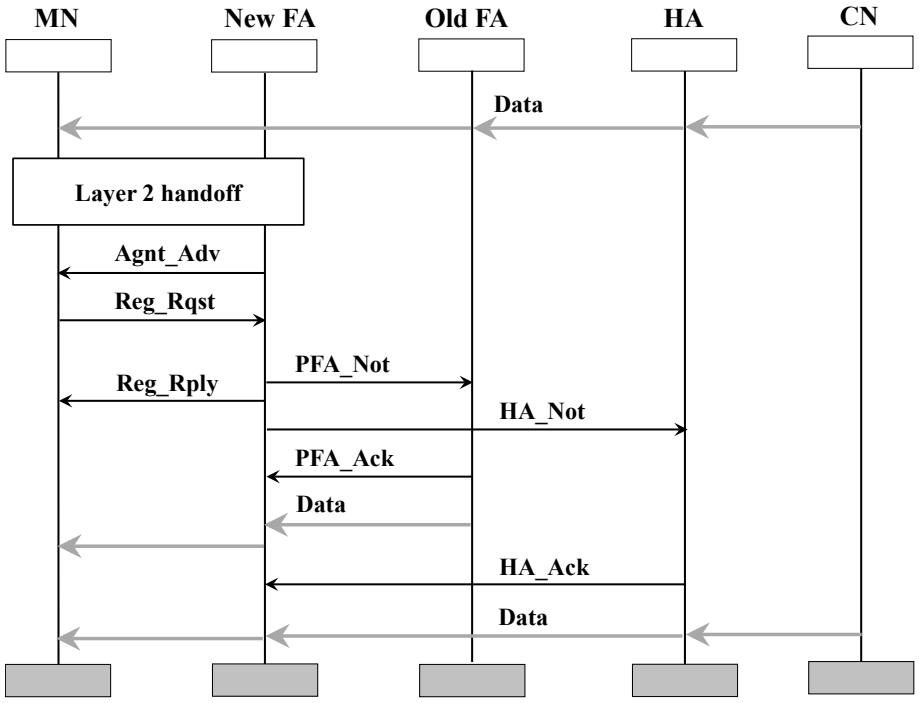


Figure 2.8: MIFAv4 operation in reactive mode

In order to inform the old FA, the new FA sends a Previous FA Notification (PFA_Not) message to the old FA, which responds with a Previous FA Acknowledgement (PFA_Ack) mes-

sage and then begins forwarding data packets to the new CoA. In addition to the notification of the old FA, a new SA (the key is $K3_{MN,FA}$) as well as two new random variables (R_1' and R_2') are generated. The new key will be used to secure the control messages that the MN will exchange with the subsequent new FA. The new variables (R_1' and R_2') will be used to calculate the authentication values during the next registration with the subsequent new FA. Following this, the new FA transmits a Reg_Rply message to the MN. This message contains $Auth_2$, the new random variables and the newly generated SA[1]. As the MN receives the Reg_Rply message and successfully authenticates it, the MN declares the handoff to be successfully completed.

After the new FA sends a Reg_Rply message to the MN, a new SA (the key is $K3_{FA,HA}$) is generated to be used for securing the control messages that the subsequent new FA will exchange with the HA. Following this, the new FA sends a HA Notification (HA_Not) message to the HA, see Figure 2.8. The message contains the newly generated SA and random variables. When the HA receives the HA_Not message and successfully authenticates it, the HA responds by transmitting a HA Acknowledgement (HA_Ack) message, containing the information required to construct the MN-specific data**,** to the new FA. Moreover, the HA redirects the tunnel from the old FA to the new one.

The last task the new FA must do is the distribution of MN-specific data to all members of its L3-FHR. For this purpose, the new FA executes the information distribution procedure discussed previously.

The handoff procedure in predictive mode is shown in Figure 2.9. Once the MN notices that a handoff to a new subnet will occur in the near future, it fires a Layer 2-trigger (L2-trigger)[2]. The trigger prompts the MN to begin the layer 3 handoff in advance. For this purpose, the MN exchanges a Proxy Router Solicitation (Pr_Rt_Sol) message and a Proxy Router Advertisement (Pr_Rt_Adv) message with the old FA. The Pr_Rt_Adv message provides information about the new FA. Of course, the MN needs this information to proceed with the handoff. Following this, the MN constructs and sends a Reg_Rqst message to the new FA via the old one. As the old FA receives this message, it issues an Initial Acknowledgement (Int_Ack) message to the MN and forwards the Reg_Rqst further to the new FA.

The new FA authenticates the Reg_Rqst and behaves similarly to when operating in the reactive mode. Following a successful authentication and authorization of the MN, the new FA notifies the old FA by means of a Reg_Rply message. Moreover, the new FA generates two new random variables and a new SA to be used to secure the control messages that should be exchanged between the HA and the subsequent new FA. The new SA and random variables are transmitted to the HA with a HA_Not message. The HA in turn calculates the new authentication values and sends them along with the information required to construct the MN-specific data to the new FA with a HA_Ack message. Clearly, at this time the HA begins tunneling data packets destined for the MN to the new CoA.

As the old FA detects a Layer 2-Link Down (L2-LD) trigger[3], it forwards the MN's data packets to the new FA, which buffers them until a Layer 2-Link Up (L2-LU) trigger[4] is raised. The appearance of the L2-LU trigger prompts the new FA to send a Reg_Rply message to the MN containing the new random variables, $Auth_2$, as well as a newly generated key ($K3_{MN,FA}$) to authenticate the messages that should be transmitted between the MN and

[1] Notice that all SAs are transmitted encrypted (not as clear text), see [Dia10].

[2] L2-trigger is used to predict handoffs prior to a break of the wireless link.

[3] L2-LD trigger provides notification that the old wireless link with the old FA has just been destroyed.

[4] L2-LU trigger indicates that the new wireless link with the new FA has just been constructed.

the subsequent new FA. Following the Reg_Rply, the MN resumes its communication on downlink as well as on uplink.

The last task is the distribution of MN-specific data to the members of the L3-FHR of the new FA. This is done by the execution of the information distribution procedure.

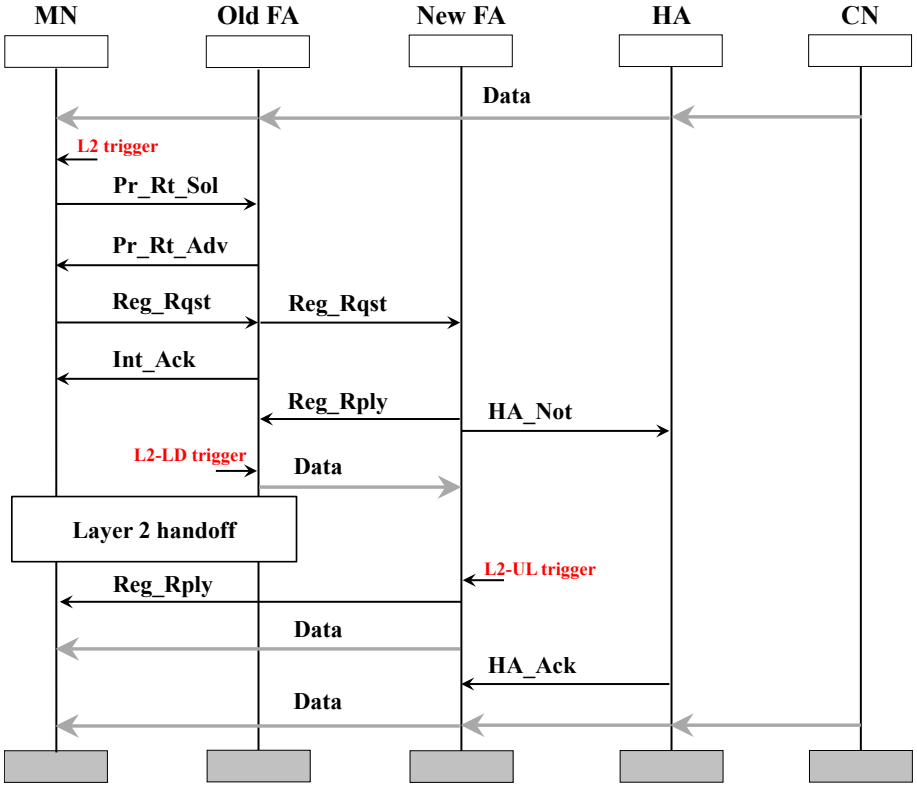


Figure 2.9: MIFAv4 operation in predictive mode

Pros and cons: MIFA eliminates the most problems MIP suffers from. It achieves seamless handoffs without restricting the network topology or introducing new intermediate nodes beyond those currently known from MIP. Sure, this simplifies the employment of this protocol in existing IP-based networks, which is a great advantage. Triangular routing and ingress filtering problems remain in MIFA unsolved. Considering security, this protocol does not add any new vulnerability. It is even more secure than other protocols, since authentication and authorization are required when exchanging any control message between any two nodes (i.e. a FA, the HA and the MN) in the network

2.3 Conclusion

This chapter has addressed the mobility management issues. The pros and cons of implementing mobility procedures in the different layers of the TCP/IP reference model have been discussed as well. Because network layer mobility management plays a major role in this dissertation, this chapter provided an insight into the main principles and approaches of network layer mobility. From the investigation of network layer mobility, one notices that substantial effort has been made to support fast and seamless handoffs. However, QoS was not a key issue in existing approaches for mobility management. Although supporting fast as well as seamless handoffs is essential to provide QoS guarantees, it is not enough, since issues such as the check of the availability of required resources in the new subnet, reservation of these resources, etc. should be considered as well, which is not the case in existing mobility management solutions.

Chapter 3: Quality of Service in Mobile Communication Networks

As mentioned in the Chapter 1, "all-IP" is a seminal future of mobile communication networks. These networks aim at offering ubiquitous access to information as well as use of services anytime, anywhere and anyhow. They follow the pattern "everything over IP, IP over everything" and aim at providing a wide range of services beyond voice communication, e.g. e-banking, video conferencing, AtHome Desktop, etc. We also mentioned in Chapter 1 that the provision of QoS in such networks is one of the essential challenges that we will address in this dissertation.

This chapter provides an insight into the challenge mentioned above and is structured as follows: section 3.1 addresses the question: What is QoS? This section provides a definition of QoS and describes the different types and attributes. We then review the basic mechanisms of QoS in section 3.2. Section 3.3 discusses QoS in IP-based networks and describes known mechanisms and protocols for QoS in detail.

3.1 *What Is QoS?*

3.1.1 Definition of QoS

Different organizations use different definitions of QoS, for example:

- ISO 8402 defines the word quality as "the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs" [ISO 8402].
- The word quality is defined by ISO 9000 as "the degree to which a set of inherent characteristics fulfills requirements" [ISO 9000].
- ITU-T (Recommendation E.800 [ITU-TE.800]) and ETSI [ETSI- ETR003] define QoS as: "the collective effect of service performance which determines the degree of satisfaction of a user of the service".
- IETF defines QoS as "the ability to segment traffic or differentiate between traffic types in order for the network to treat certain traffic flows differently from others. QoS encompasses the service categorization and the overall performance of the network for each category." [ETSI-TR102].

One may notice that all definitions reflect similar meanings of QoS. From the network point of view, QoS is defined as the ability of a network element (e.g. a network node, an application, etc.) to have level(s) of assurance that the traffic handled by this element as well as the service(s) requirements offered by the element can be satisfied.

The meanings of QoS change based on many factors, e.g. the application the user runs, the conditions under which the application is being used, etc. Furthermore, QoS meanings range from the perception of service users to a set of parameters necessary to assure particular quality [Mar07].

Three types of QoS are identified in [Har01], namely intrinsic, perceived and assessed QoS. Under the Intrinsic QoS (I-QoS) indicates the quality assurance that the network itself directly provides. I-QoS is typically described by means of objective parameters such as packet loss, throughput, delay, etc. Under Perceived QoS (P-QoS) one understands the quality perceived by users. P-QoS reflects how satisfied users are with the service. P-QoS is measured by the

"average opinion" of users. Assessed QoS (A-QoS) expresses the intention of a user to continue using a specific service, see [Mar07] and [Har01]. The three types of QoS mentioned heavily relate to each other. If very good objective performance is provided by the network (e.g. low delay, low packet loss, etc.), users will appreciate this and, thus, continue using the services of this network. In other words, if I-QoS is high, P-QoS will be high, as well. This may lead to an improvement in A-QoS. Keep in mind that users may not perceive an object performance increase or decrease as an I-QoS variation. The study of the relation between QoS types is certainly a topic of great importance and gains, therefore, great attention. Note that telecommunication companies invest great economic efforts in designing and implementing newly developed schemes, protocols, etc. They expect, therefore, great acceptance by users. If, for example, a specific algorithm produces performance improvement of 30 % concerning a certain performance metric (e.g. 30 % lower delay, lower packet loss, etc.), this will appear as a positive result from the I-QoS point of view. However, it is very important to investigate whether users will appreciate this improvement or not.

A-QoS relates to the P-QoS and, of course, implicitly to the I-QoS. Again, this relation is not straightforward since A-QoS is affected by factors other than users' perception, e.g. marketing and commercial aspects. For instance, users may accept degradation in service quality if the service is offered for free, see [Mar07].

At the moment, QoS is offered in terms of I-QoS Service Level Specification (SLS), which is "a set of parameters and their values which together define the service offered to traffic" [Gro02]. The SLS is a part of the Service Level Agreement (SLA) between a service provider and a service user. The SLA is a contract signed from both the service provider and the user. This contract describes the policies applied to provide certain SLS for that user, see [Mar07].

3.1.2 QoS Metrics

Most important metrics concerning IP-based networks include IP packet transfer delay, IP packet delay variation (also known as jitter), IP packet loss ratio, IP packet error ratio and bandwidth [ITU-T-Y.1541], [ITU-T-Y.1540]. The following provides a more detailed description of these metrics.

IP packet transfer delay: the IP packet transfer delay represents the delay required for a packet to be sent from a sender via a network to a receiver. This delay includes the processing delays the packet experiences within the routers on the path from the sender to the receiver as well as the transmission delays across the links on this path. Applications typically set an upper bound for this delay (termed as play-out time) and consider each packet experiencing a transfer delay exceeding this upper bound as lost. The IP packet transfer delay is termed IP packet end-to-end delay, as well.

IP packet delay variation (jitter): the IP packet delay variation is the variation in the end-to-end delays that IP packets experience at a receiver. Jitter is introduced by queue delays within routers that process various traffic passing through shared resources. Jitter heavily affects real-time applications such as voice, video, etc. High levels of jitter are not acceptable by such applications.

IP packet loss ratio: the IP packet loss ratio is the ratio of packets that get lost during an active session to those that are sent from a sender. This metric occurs in:

- wireless connections due to interferences, weather changes, handoffs between cells, etc. and
- routers (congestion points) due to the limited buffering capacities.

The IP packet loss ratio is critical since it highly affects applications that use TCP.

IP packet error ratio: the IP packet error ratio is the ratio of packets experiencing errors to those sent on the path from a sender to a receiver during an active session. This metric is also critical since it results in possible data loss, retransmissions, etc.

Bandwidth: the bandwidth of a network link is the maximum transmission rate that can be maintained between the two end-points of the link. In this way, the bandwidth maintained for traffic on a path from a sender to a receiver is restricted based on many factors such as the physical infrastructure specifications of the path, the type of traffic, number of other flows sharing the resources on this path, etc. The higher the bandwidth the application can obtain, the higher the throughput of the application.

Of course some applications may require other QoS metrics, such as the skew[1]. Moreover, each application requires a different combination of some/all metrics discussed above and possibly additional metrics, as well, in order to maintain the required QoS. For instance, video streaming requires low IP packet end-to-end delay, low jitter and high bandwidth. Such an application, however, tolerates some loss of IP packets and is not heavily interrupted when it receives some packets containing errors.

3.1.3 QoS Guarantees

Based on [Hoa03] and [Fer90], QoS guarantees can be either deterministic or statistical guarantees. Deterministic QoS guarantees are provided as hard bounds on performance metrics, e.g. a specific value for the jitter, end-to-end delay, loss ratio, etc. These guarantees have the form $var \leq bound$ where *var* is a performance metric that has an upper bound and bound is user-specified.

On the contrary to deterministic QoS, statistical guarantees place no hard bounds on performance metrics. Instead, they provide a probability that this bound will be satisfied. These guarantees have the form $prob(var \leq bound) \geq P$ where *P* is the lower bound of the probability that the performance metric *var* will be satisfied.

3.1.4 Concatenation of QoS Guarantees

Meeting QoS guarantees is fundamentally an end-to-end issue – that is, from application to application. In the context of interconnected communication networks, end-to-end QoS is not easy to realize and will certainly be composed of a combination of edge-to-edge QoS guarantees from individual networks. It is, however, essential that edge-to-edge QoS guarantees are hidden from applications and thus provide an end-to-end QoS guarantees from the applications' point of view, see Figure 3.1 and [Hoa03]. Note that each edge-to-edge QoS guarantee should be able to provide the required end-to-end QoS parameters (e.g. delay, jitter, bandwidth, etc.) regardless of the mechanisms applied to realize the edge-to-edge QoS, more details will be provided in section 3.3.2.1.

[1] Skew expresses the difference between the delays packets belonging to different media encounter, e.g. voice and video within a video conference service. The larger the skew, the less the voice and the video within the videoconference are synchronized (users experience poor dubbing) [Mar07].

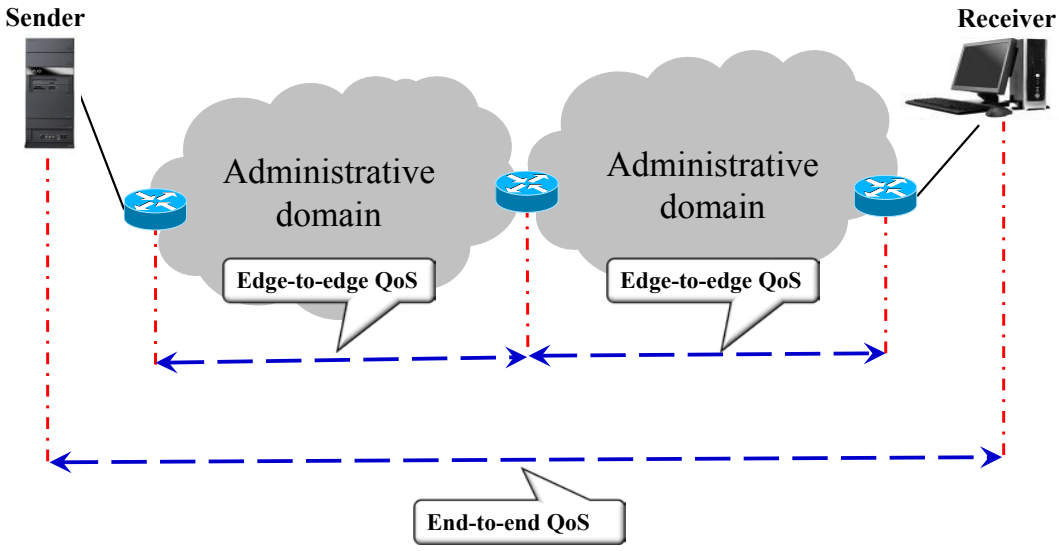


Figure 3.1: Concatenation of QoS guarantees

3.1.5 Relative and Absolute QoS

The services that a network offers are normally classified into several service categories, which are treated individually [FHu98]. Relative QoS relates to this differential handling. The network provides a kind of guarantee that the traffic of a higher class will receive in most cases better handling than all lower classes. At any network situation, not worse handling than all lower classes is guaranteed [Hoa03]. The definition of absolute QoS is done by means of quantitative performance metrics. Examples are statements like “no packet loss”, “end-to-end delay of lower than 150 msec”, etc., see [Hoa03].

3.1.6 QoS Specification

Based on the layered model known from communication systems, a QoS layered model was developed, as well [CCH94], [Hoa03], see Figure 3.2. As the figure shows, one distinguishes four layers, namely: user level, application level, system level and network level QoS. User level QoS relates to the QoS users notice and perceive. Application level QoS is determined by parameters of the application level, e.g. resolution, frame rate, etc. System level QoS relates to operating system parameters. Finally, network level QoS is captured by network parameters, e.g. throughput, packet loss, error rate, etc.

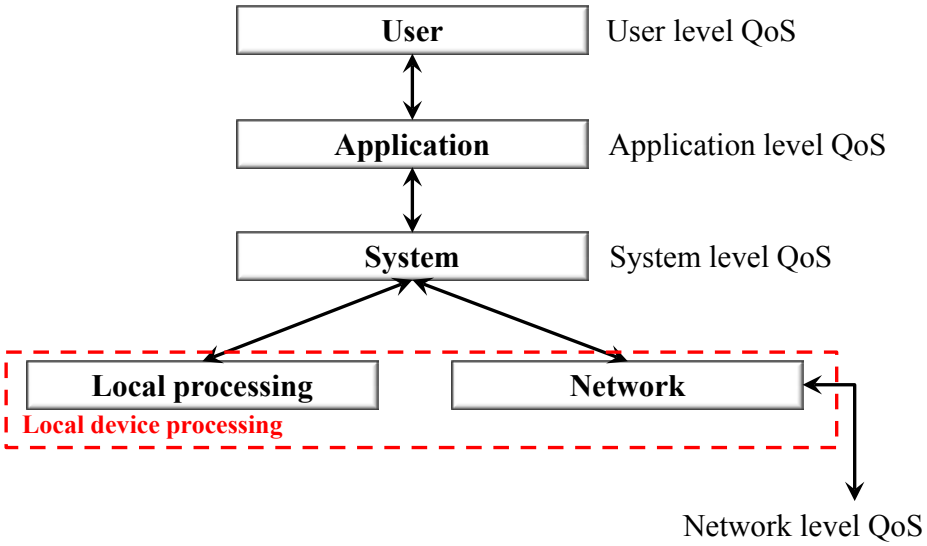


Figure 3.2: QoS layered model

As described in [Hoa03] and [ACH96], QoS specification is primarily user- rather than system-oriented. Therefore, QoS specification focuses on capturing user or application level QoS requirements and management policies, and mapping them, after that, into different layers.

A wider view of QoS specification that focuses not only on the user and/or application level QoS is presented in [HLa97]. The authors define this QoS specification as a triple of a service descriptor, a traffic descriptor and a QoS profile. The QoS profile is used to determine the attributes of the application, the regulation of traffic and the guaranteed QoS. [ACH96] provides an even more detailed view and illustrates that QoS specification encompasses a flow synchronization specification, a flow performance specification, a level of service, QoS management policies and a cost of service. The flow synchronization specification determines the degree of synchronization between multiple related flows [LGh90], e.g. audio and video flows in a video conference. The flow performance specification specifies performance metrics of the user's flow [Par92] in a quantitative way, e.g. jitter, loss rates, etc. On the contrary to the flow performance specification, the level of service determines performance metrics in a qualitative way. This allows for a distinction between hard[1], firm[2] and soft[3] performance guarantees [Ami07]. In other words, the level of service captures the degree of certainty that the QoS requirements requested will actually be honored. The QoS management policies specify the QoS adaptation degree as well as scaling actions to be carried out when the contracted QoS is violated [CCH95]. For instance, the play-out time of a specific video stream may be adapted in response to variations in the end-to-end delay so that the distortion of the video presented remains perceptually minimal. The service cost captures the price the user is willing to incur for the level of service he required. For sure, the cost of service plays major role. The user may accept QoS deterioration if the service he uses is offered for free. On the contrary, the user will not accept any QoS distortion if his service is financially expensive.

[1] Hard real-time system: defines a hard limit (deadline) that cannot be missed - otherwise a negative effect is imparted on the system.

[2] In contrast to hard real-time system, the deadline from a firm real-time system may be missed. If this occurs, a zero value is imparted on the system.

[3] Similar to the firm real-time system. However, if this deadline is missed, the value imparted on the system after the deadline decreases monotonically to zero.

3.2 QoS Mechanisms

QoS mechanisms are selected based on the defined QoS specification, the resources available and the resource management policies applied. One distinguishes between two categories of QoS mechanisms, namely static and dynamic QoS mechanisms [ACH96]. Static QoS mechanisms are termed QoS provision mechanisms, they deal with end-to-end QoS negotiation phases. Dynamic QoS mechanisms are known also as QoS control and management mechanisms, they relate to media-transfer phases. The reason behind the distinction between QoS control and management mechanisms is the different timescale they operate on. QoS control mechanisms operate on a faster timescale than QoS management mechanisms. Concerning QoS provision mechanisms, we distinguish between five components, namely service specification, QoS mapping, admission testing, QoS negotiation and resource reservation protocols. The service specification component captures the QoS requirements related to call setup or participation time [Hoa03]. The QoS mapping component describes how QoS requirements have to be translated between various levels (i.e. application level, system level, etc.). The admission testing component, also known as admission control, describes the functions that decide whether a new request can be accommodated or not. The QoS negotiation component determines the way in which all communication parties reach an agreed specification of QoS [ACH96]. Resource reservation protocols are responsible for the allocation of resources according to the QoS specification the user requested.

With respect to QoS control mechanisms, one also distinguishes five fundamental components, which are flow shaping, flow scheduling, flow policing, flow control and flow synchronization [ACH96]. The flow shaping component regulates flows depending on either user or network supplied performance specifications. The regulation is achieved, for instance, by regulating a fixed throughput or statistical performance metrics (e.g. sustainable rate, burstiness, etc.). The flow scheduling component controls the forwarding of flows within schedulers. The flow policing component observes whether the QoS contracted between a user and a provider is being maintained. Flow policing is essential in cases where administrative and/or charging boundaries are being crossed. The flow control component determines how to deal with the resources allocated. It is normally implemented either using open- or closed-loop schemes, [ACH96], [Jac93] and [Jai95]. Finally, the flow synchronization component controls the ordering of events and the precise timing of multimedia interactions.

It is commonly agreed upon that just committing resources and providing real-time control are often not sufficient to maintain an agreed QoS level. There is a need for additional **QoS management mechanisms** to ensure that the contracted QoS is sustained. QoS management is, in principle, similar to QoS control from a functionally point of view. However, it operates on a slower timescale [PSt95]. QoS management mechanisms include QoS monitoring, maintenance, adaptation and degradation. QoS monitoring techniques are applied to trace ongoing QoS levels. Of course, the more often the QoS monitoring techniques are executed, the more overhead is produced and the more accurate the monitoring is. The time, over which a certain parameter is monitored, is selected based on QoS management policies that may be specified in cooperation with the user, e.g. using certain QoS signaling mechanisms. The QoS maintenance has the function of sustaining QoS. This is achieved by comparing the monitored to the expected QoS and making decisions based on the results of the comparison, e.g. tuning operations to maintain the QoS provided. The QoS adaptation determines how applications adapt to fluctuations in the end-to-end QoS, see [PSt95] and [FSV97]. The QoS degradation signals the user that the QoS level he requested can no longer be sustained. In response, the user decides how to proceed, e.g. accept the available QoS level, adapt to the QoS level offered, etc. see [ACH96].

3.3 QoS in IP-based Networks

As often mentioned in this work, the Internet is continually expanding and providing many types of services with various resource demands. It is well-known that the current Internet architecture was designed to provide best-effort services. The necessity to satisfy the ever increasing demands for QoS guarantees has motivated companies, engineers and researches to develop new approaches capable of providing different levels of QoS. Therefore, it is helpful to provide an insight into the approaches designed to enhance IP-based networks, especially the Internet, with QoS capabilities.

3.3.1 Integrated Services Architecture

The Integrated Services architecture (IntServ) [BCS94] aims at providing real-time applications with QoS guarantees. It is well known that the Internet, as originally designed, offers no QoS guarantees. Of course, before real-time applications are widely in use, the Internet architecture should be extended to provide real-time QoS. Moreover, the extension should support unicast and multicast applications. Extending the Internet to realize these goals is the purpose of the IntServ architecture.

IntServ is a per-flow QoS model. Each flow is a packet stream that requires a specific QoS level and is identified by the 5-tuple “IP source address, IP destination address, protocol, TCP/UDP source port and TCP/UDP destination port”. The routers implementing IntServ should support appropriate QoS for each flow. The functions that IntServ-enabled routers support to create such different QoS include traffic control and resource reservation. Traffic control, in turn, is implemented by three components, namely a packet scheduler, packet classifier and an admission control.

Resource reservation is carried out by a specific protocol. This protocol is necessary to create and maintain flow-specific states in end-hosts as well as in each router along the path between the end-hosts. Note that invoking states in routers is a fundamental change to the Internet architecture, since the Internet was designed based on the concept that says: routers should be kept simple and all flow-related states should be placed in end systems [Cla88]. This concept is one of the main points that led to the robustness and tremendous success of the Internet. Thus, to create states in routers while preserving the robustness of the Internet, the states are created “soft” and deleted if not explicitly refreshed. The standard resource reservation protocol that IntServ architecture employs is the Resource reSerVation Protocol (RSVP), see [BCS94] and [BZB97].

Figure 3.3 displays the structure of an IP router that implements the IntServ architecture. As the figure shows, the router implements two main functional parts, namely a forwarding and a backbone part.

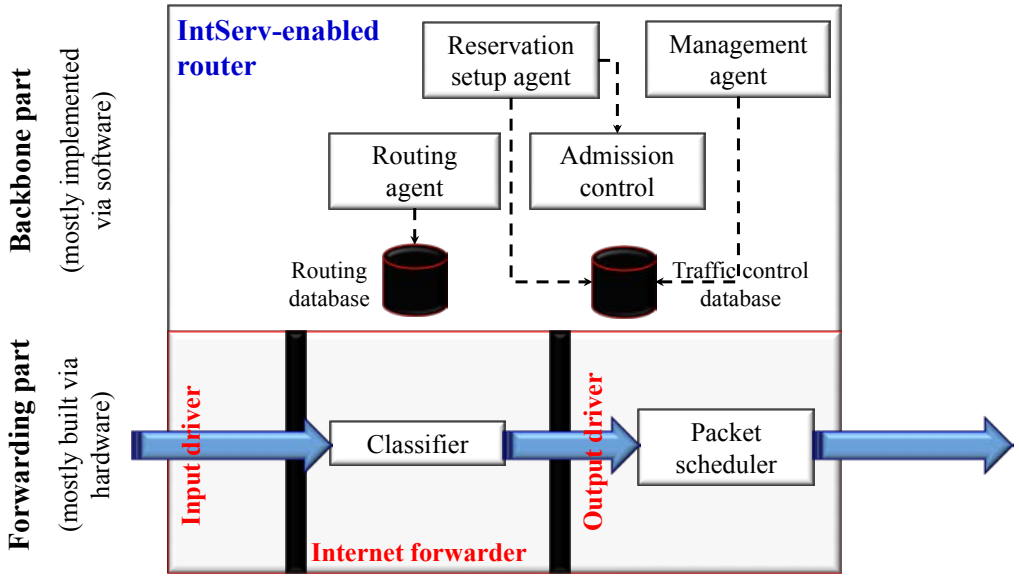


Figure 3.3: Structure of an IP router with IntServ support [BCS94]

The forwarding part handles each packet passing through the system. Therefore, it involves a hardware assist containing three main components, namely an input driver, an Internet forwarder and an output driver. The Internet forwarder forwards the packets based on their IP headers. This component implements a classifier to classify incoming packets and forward them accordingly to the adequate output driver which, in practice, implements a packet scheduler.

The backbone part is recommended to be implemented as software. It creates data structures required to control the forwarding part. The routing agent implements a specific routing protocol (e.g. IP) and constructs a database to assist in making routing decisions. The reservation setup agent implements a protocol to reserve resources. If the admission control mechanism decides to accommodate a new request, the required changes to the classifier as well as packet scheduler database are done. The management agent is used to manage the work of the router as a whole. It should be able to modify the entries of the classifier as well as packet scheduler databases and to create new admission control policies.

The structure of a host that supports the IntServ architecture is similar with the addition of applications. Rather than forwarding packets, the packets originate from applications and terminate in applications.

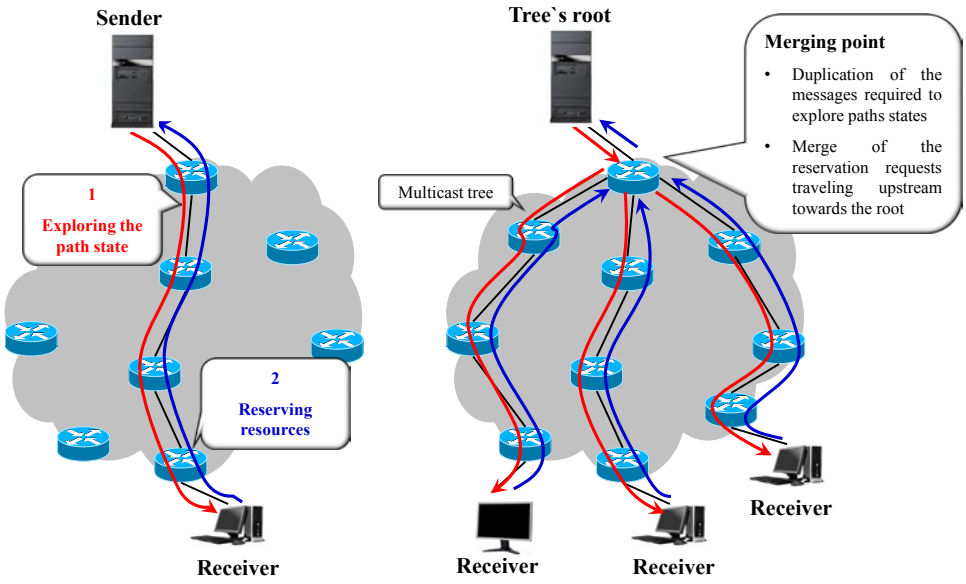
Based on the IntServ model, there are two service types, namely guaranteed [SPG97] and controlled-load services [Wro97a]. The guaranteed service guarantees that packets will arrive at their destinations within a certain delay bound and will not be discarded within queues. It is assumed, however, that the traffic of the user remains within its specified parameters. This service is intended for applications requiring a firm end-to-end delay bound, e.g. video and voice applications. The controlled-load service is intended for a wide range of applications such as file download, adaptive real-time applications, etc. These applications tolerate a certain amount of packet loss and delay and have been shown to work well in non- or low-loaded networks. However, they quickly degrade with increasing loads. The basic idea of controlled-load service is to make the end-to-end behavior, applications operating between certain sources and destinations experience, tightly approximates the behavior visible to applications when exchanging best-effort traffic in low-loaded networks. Note that the controlled-load service does not use specific parameters preferred and provided by end-users, e.g. the required delay, packet loss, etc. Instead, this service makes a commitment to provide the user with a

service closely equivalent to what the user sees in case it exchanges best-effort traffic under low load situations.

3.3.1.1 Resource ReSerVation Protocol (RSVP)

The Resource reSerVation Protocol (RSVP) [BZB97], [BEB95] is a setup protocol designed to be used by the IntServ architecture for resource reservation within the Internet. The purpose of RSVP is the creation of flow-specific states in routers and hosts and thus the allocation of resources on these elements. RSVP was designed to deliver a robust, scalable, flexible and heterogeneous resource reservation setup for both unicast and multicast applications. This led to a number of basic features as [BEB95] mentioned, namely:

- Support of point-to-point, point-to-multipoint as well as multipoint-to-multipoint communication models.
- Employment of receiver-initiated reservation procedures to enable scaling well for a large number of receivers. This strategy works especially well in multicast scenarios, in which reservations for separate receivers are merged in crossover nodes residing in the multicast tree root, see Figure 3.4.



a) Unicat scenario **b) Multicast scenario**

Figure 3.4: Receiver-initiated reservation for unicast and multicast scenarios

- Management of soft states to achieve robustness and simplicity as well as guarantee that reserved resources will be released under all circumstances. Soft states time out if not periodically refreshed.
- Separation of reservations from routing since RSVP is a pure reservation setup protocol. The routing protocol decides how control messages should be routed towards the destination. RSVP solely creates the states on this path.

Operation of RSVP for unicast applications: the operation of RSVP for such applications is relatively simple and can be described as follows: when a sender would like to reserve resources for an application, it transmits a PATH message towards the receiver. The PATH message contains objects describing the QoS parameters required. These objects are summarized as follows:

- A sender template (SENDER_TEMPLATE) object: defines the format of data packets that the sender will generate and has the form of a filter specification (FILTER_SPEC), which could be used to select the packets of the sender from others in the same session or link.
- A sender traffic specification (SENDER_TSPEC) object: specifies the characteristics of the data flow that the sender will generate.
- A previous hop (PHOP) object: contains the previous hop address, which is the address of the interface in the previous hop, through which the PATH message was sent.
- The session object: is contained in every RSVP control message. It contains the IP destination address, IP protocol ID and generalized destination port.
- An optional advertisement specification (ADSPEC**)** object: carries the advertising One Pass With Advertising (OPWA[1]) information for the flow.
- TIME_VALUES object: contains the value of the refresh period applied to refresh the resources reserved.

Each router residing on the path between communicating peers that is capable of satisfying the request creates a state, termed path state, for the sender and the specific flow. As mentioned earlier, the states are stored soft and refreshed periodically as long as the resources are required. Each state is primarily defined by the SENDER_TEMPLATE and the SESSION object. Of course, all other objects present in the PATH message are stored in the state, as well.

Once the receiver receives a PATH message, it responds by sending a reservation request (RESV message) upstream to the sender. The RESV message follows the exact path that the concerned PATH message followed. Each RESV message contains mainly the following objects:

- The session object: contains, as mentioned earlier, the IP destination address, the IP protocol ID and a generalized destination port.
- The TIME_VALUES object: contains, as cited above, the value of the refresh period applied to refresh the reserved resources.
- The STYLE object: determines the reservation style and any style-specific information that does not exist in the FLOWSPEC or FILTER_SPEC objects.
- The FLOW SPECification (FLOWSPEC) object: defines the desired QoS.
- The FILTER SPECification (FILTER_SPEC) object: defines the subset of session data packets that should be handled according to the QoS specification specified in the FLOWSPEC object.

Each router or end-host reserves the required resources once it receives a RESV message for the flow, for which a PATH message has been previously received. Moreover, the RESV message creates and maintains states, termed reservation states, in each node along the path from the receiver to the sender including the sender itself. The operation of RSVP for unicast applications is illustrated in Figure 3.5.

[1] As stated in [BZB97], the basic reservation model for RSVP is termed “one pass”. This model implies that the receiver transmits a reservation requests to the sender. Each router on this path makes an independent decision regarding the requests, i.e. accept or reject. Using this model, the receiver cannot determine what the end-to-end service will look like. Therefore, an enhancement to this model (the OPWA model) was proposed in which RSVP control messages (PATHs) are sent downstream to the receiver in order to gather information to help predict end-to-end behavior.

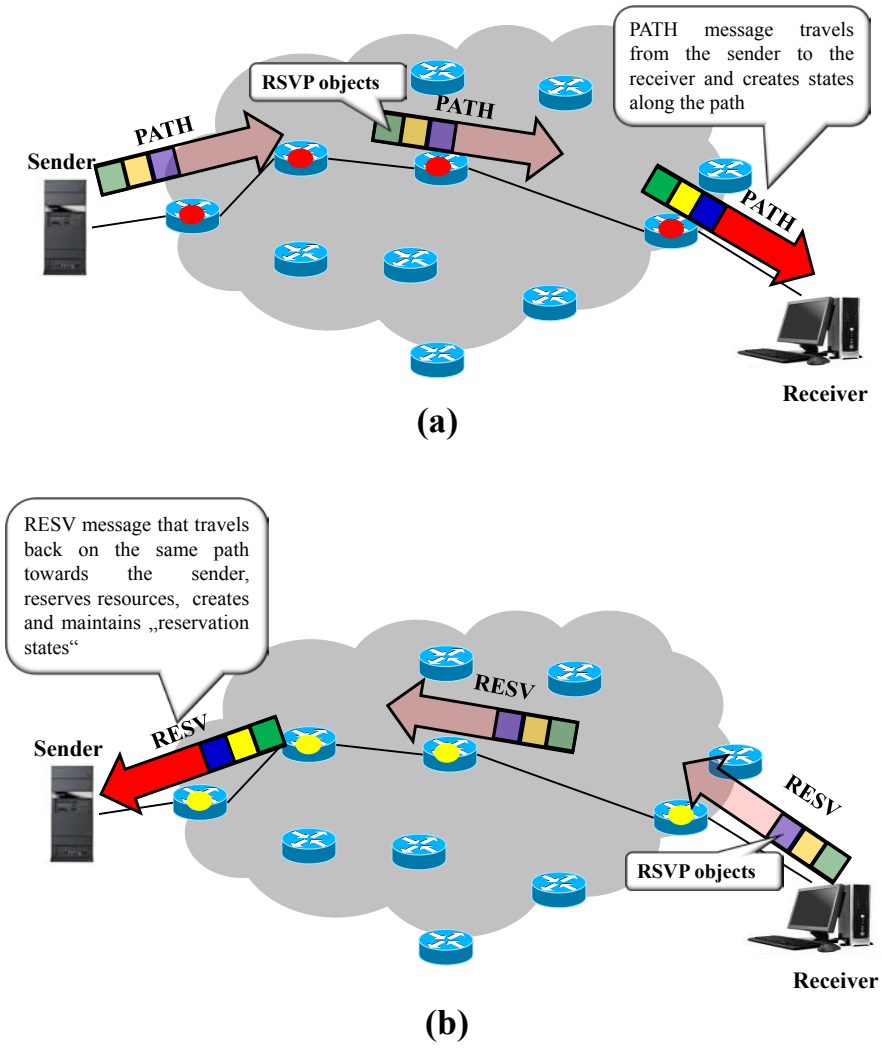


Figure 3.5: Operation of RSVP for unicast applications

Operation of RSVP for muticast applications: the operation of RSVP for such applications is, in principle, similar to the operation for unicast applications. It is well known that a multicast tree is constructed from a sender to a group of receivers. The sender sends first a PATH message to the receivers. Each node belonging to the multicast tree and having more than one node, participating in the multicast tree beneath it, duplicates the PATH message and forwards them further. As mentioned above, each PATH message creates and maintains a path state. Once a receiver receives a PATH message, it responds by sending a RESV message towards the sender. The RESV message causes resources to be reserved and reservation states to be created as well as maintained. It is important to note that RESV messages originating from different branches of a multicast tree for the same sender must be merged in crossover nodes as these messages travel upstream. Other operation aspects are similar to those discussed for unicast applications.

3.3.1.2 Pros and Cons

As is apparent from the discussion above, the IntServ architecture works on the basis of a per-flow or per-customer QoS model and therefore delivers high QoS guarantees. The wide variety of applications and services that the proposed architecture can serve is a main advantage. The soft state principle mentioned is a main advantage, as well, since it ensures that resources

will always be released if not in use. The soft state principle can, however, be seen as a disadvantage, namely in terms of signaling overhead. That is, renewing these states requires a periodical exchange of PATH and RESV messages.

When a network provider would like to support the IntServ architecture, he must update all routers in his domain. Of course, this complicates the employment of this architecture. Moreover, as the network grows, the IntServ architecture begins to suffer since all nodes on the paths between senders and receivers should support the IntServ model.

3.3.2 Differentiated Services Architecture

Differentiated Services (DiffServ) Architecture [BBC98] is proposed to provide a scalable and flexible QoS model for the Internet. The DiffServ architecture depends on the following two basic principles, see Figure 3.6.

- The computational complexity should be shifted to the routers residing on the domain's boundaries, called edge or boundary routers. Other routers located inside the domain remain simple and are termed core or interior routers.
- QoS is provided based on a per-traffic class model rather than per-traffic flow or per-customer model. This implies that traffic of various users is handled in the same manner if belonging to the same class. Of course, this reduces the overhead[1] and consequently improves the scalability as compared to the IntServ architecture.

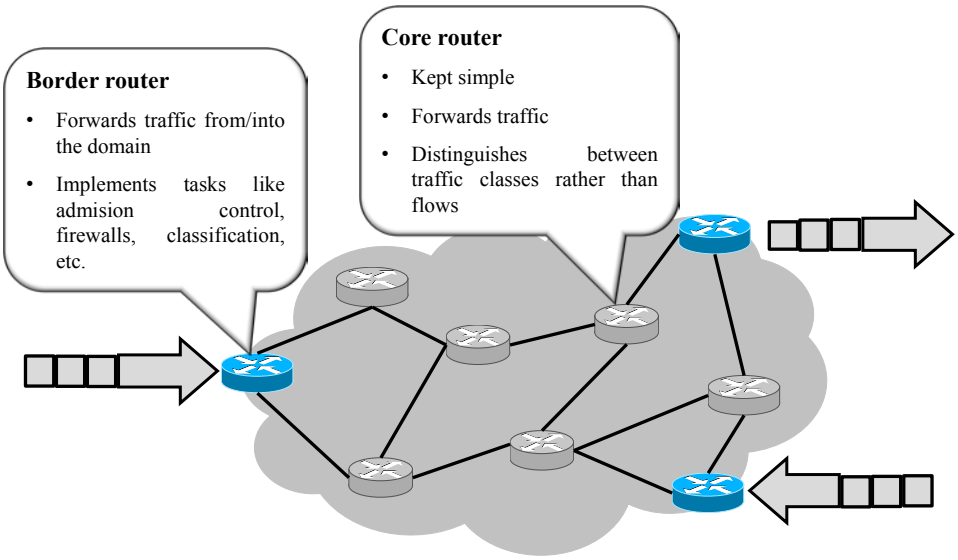


Figure 3.6: Basic principles of DiffServ architecture

As shown in Figure 3.6, edge routers interconnect the DiffServ domain with other DiffServ or non-DiffServ domains. All traffic is routed into/from the domain through these routers, which should be able to classify incoming traffic into a small number of classes, as Figure 3.7 shows. Moreover, tasks such as admission control and traffic conditioning are also implemented in edge routers. The main task of edge routers, however, is the prediction of the Per-Hob Behavior (PHB), which is defined in [BBC98] as "a description of the externally observable forwarding behavior of a DiffServ node applied to a particular DiffServ behavior aggre-

[1] Note that smaller number of states should be stored as compared to the IntServ architecture. Thus, the overhead resulting from the storage, maintenance and release of the states is also reduced.

gate". This simply refers to the prediction of how a traffic flow mapped to a certain traffic class will be handled by other DiffServ nodes residing inside the DiffServ domain.

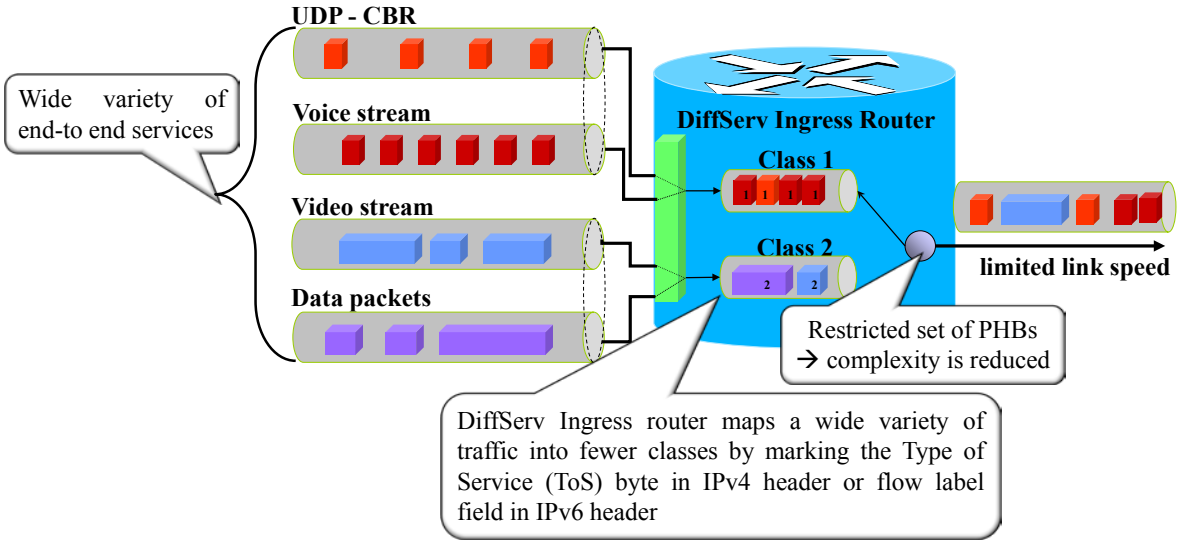


Figure 3.7: Operation of an edge router

The classification of traffic is done using two primary types of classifiers, namely Behavior Aggregate (BehAgg) and Multi-Field (MF) classifiers. The BehAgg classifier classifies incoming packets based on the Differentiated Services Codepoint (DSCP) values of the IP header, which is used to choose the PHB. However, the MF classifier classifies packets depending on other fields such as source address, destination address, source port, destination port, protocol ID, etc. Sure, the MF classifier is more complex than the BehAgg classifier.

As cited above, the prediction of PHBs is a major task within the DiffServ domain. Failed predictions result primarily in undesirable effects, e.g. QoS deterioration, inaccurate resource allocation, etc. The DiffServ working group [IETFDiff] divides the PHBs into three types, namely Expedited Forwarding (EF) [DCB02], Assured Forwarding (AF) [HBW99] and best-effort. The EF PHB is designed to support low delay, low loss and low jitter services by guaranteeing that EF packets are served at or above a certain configured rate independent of the intensity of other non-EF traffic attempting to transit the node. The AF PHB is intended to provide different levels of forwarding assurance for IP packets. It defines four AF classes - each is allotted a certain amount of resources. Each IP packet within an AF class is assigned one of three possible drop precedence probabilities. The lower the drop precedence probability that the packet has, the higher the importance of the packet is. Best-effort PHB is the default behavior of the Internet traffic. This PHB does not support any performance guarantee nor does it define a QoS level.

3.3.2.1 Enhancements to the DiffServ Architecture

Resource Management in DiffServ (RMD): as mentioned above, the DiffServ architecture was mainly developed to avoid the scalability and complexity problems the IntServ architecture suffers from. This is done by providing services on an aggregated rather than per-flow basis and pushing as many per-flow states as possible to network edges. The proposed DiffServ architecture, however, does not provide any means to enable nodes outside the domain to dynamically reserve resources or receive indications of resources availability inside the domain. In practice, services are provided based on SLAs between services providers. The

SLAs [Gro02] statically define what the accepted traffic parameters a certain user generates are.

The concept of Resource Management in DiffServ (RMD) was introduced to enable the dynamic reservation of resources inside DiffServ domains [WSK02], [MPS01] and [CTS02]. RMD mainly describes the following:

- Mapping of individual resource reservation requests into PHBs at ingress nodes.
- Hop-per-hop admission control based on the PHBs that the domain has. Interior nodes admit resources employing one of two possible operation modes: a measurement-based mode (also termed stateless mode) and a reservation-based mode (also referred to as reduced state mode[1]). Sure, the hop-per-hop admission control negatively affects the scalability of RMD since interior nodes have to do other tasks than just forwarding.
- A method to forward original reservation requests across the domain up to border nodes and even beyond.
- A Congestion control algorithm to notify border nodes of congestions inside the domain due to sudden failures (e.g. link error and rerouting of data as a result). This algorithm should be capable of terminating an appropriate number of flows to dynamically handle congestions.

The two main tasks of RMD are admission and congestion control. As mentioned above, the admission control is done using either measurement- or reservation-based mode. The measurement-based algorithm continuously measures traffic levels and available resources in interior nodes. The algorithm can, therefore, predict whether a new request can be accepted or not. The reservation-based algorithm reserves resources for PHBs that are currently handled by interior nodes. In this context, ingress nodes aggregate individual flows into PHBs and dynamically signal the change in the resources required for each PHB being handled. Of course, the reservation-based algorithm is used in cases where hard bounds on the resources are necessary, while the measurement-based algorithm is used for services that are capable of tolerating changes in the number of resources available.

The congestion control algorithm is based on probing. Interior nodes set thresholds for the traffic of PHBs. Border nodes transmit special packets for probing across the domain. Interior nodes re-mark the DSCP field of probing packets when pre-defined thresholds of PHBs are exceeded. Moreover, exceeding a pre-defined threshold of a certain PHB results in re-marking all packets of the PHB, as well. In this way, border nodes can admit new flows requesting resources.

IntServ over DiffServ: to benefit the advantages of both IntServ and DiffServ architectures, the authors in [BFY98] propose combining and interoperating between IntServ and DiffServ. The basic scenario is to deploy the IntServ end-to-end architecture across a network containing one or more DiffServ regions, see Figure 3.8. IntServ-aware nodes classify data packets on per-flow basis based on their IP addresses and port numbers. DiffServ-capable nodes support admission control and classify data packets into a limited number of aggregated classes of flows depending on the DSCP values present in IP headers.

[1] Reduced state mode differs from the stateful mode in terms of the kind of states stored. In other words, the reduced state mode stores states that do not contain as much information as the states maintained by the stateful mode.

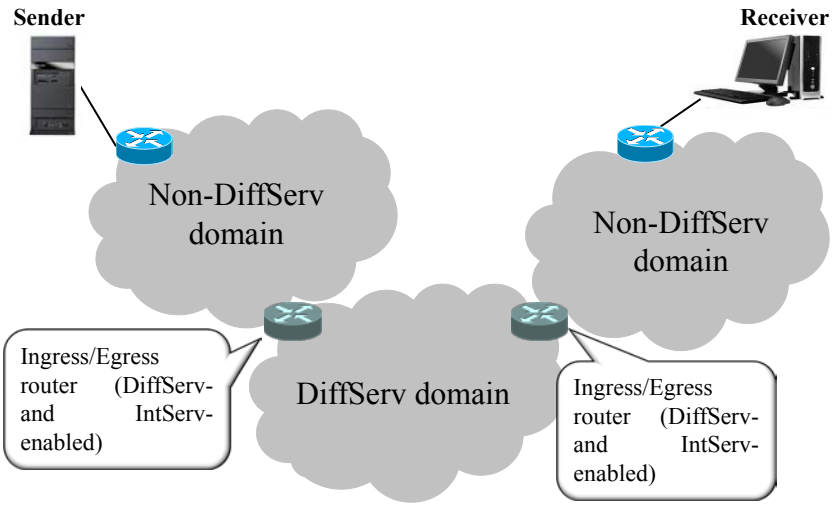


Figure 3.8: Basic scenario of IntServ over DiffServ architecture

As the sender would like to allocate resources, it transmits a PATH message containing adequate RSVP objects. The PATH message travels through the network towards the receiver and is processed in each RSVP-aware node that the message passes through. As the PATH message crosses an ingress node of the DiffServ domain, the message is forwarded transparently1 across the DiffServ domain. It is worth mentioning that both ingress and egress nodes are IntServ-aware nodes, as well. In our example, the ingress node will maintain a state for the PATH message. When the message leaves the DiffServ domain, it will be handled again by RSVP-aware nodes until the message reaches the receiver, which responds by sending a RESV message.

The RESV message returns back on the same path towards the sender and transits the DiffServ domain transparently, as well. The RESV message results in reserving resources only in all RSVP-aware nodes within non-DiffServ domains. The sender begins sending data after receiving the RESV message. Data packets use the resources reserved in non-DiffServ domains. However, as they reach the ingress node, they are assigned an adequate DSCP value. Data packets are forwarded according to the assigned PHB until they leave the DiffServ domain. The benefits of the proposal can be summarized as follows:

- The scalability problem of IntServ is solved, because IntServ parameters can be translated into DiffServ codepoints and vice versa. In this way, the network scales better. However, this involves more intensive processing in edge nodes (i.e. more complex ingress and egress nodes) and slight extra processing in the core network (mainly to forward PATH and RESV messages transparently).
- The performance of the DiffServ architecture is enhanced due to the use of RSVP as a mechanism to provide end-to-end reservations across networks.

3.3.2.2 Pros and Cons

As often mentioned, the DiffServ architecture offers QoS on the basis of per-class traffic rather than per-flow or per-customer QoS model. This principle reduces the complexity[2]. However, it does not offer as strong QoS guarantees as that the IntServ architecture provides.

[1] This means that the PATH message will not be handled as a control message nor assigned to a certain class.

[2] In terms of the implementation of the DiffServ architecture and also the management of DiffServ domains.

The key issue in the DiffServ architecture is the prediction of the PHB(s) since PHB(s) determine(s) whether new requests will be accepted and specifie(s) how incoming traffic will be handled in interior nodes. An accurate prediction is the key to a successful employment of this architecture. This implies that border routers must have an accurate view of the domain and the resources currently available. The later also includes traffic that border routers are currently handling as well as the amount of traffic they possibly expect in the near future. Means of providing such capabilities are not discussed in detail in the DiffServ architecture and are a complex issue.

Let us now discuss the deployment of DiffServ domains. When a network provider would like to employ the DiffServ architecture, he must update all routers in his domain. Sure, this complicates the employment of this architecture. Note that network providers may update edge routers only. Sure, the prediction of PHBs in this case is more difficult and less accurate than when updating all routers in the network. However, QoS guarantees remains manageable. Note also that the support of new traffic classes and the management of the domain are simpler than in other architectures.

3.3.3 Multiprotocol Label Switching (MPLS) Architecture

MPLS [RVC01] is an architecture developed to provide a fast switching and routing of traffic flows throughout the network. MPLS is a 2.5 layer protocol when considering the TCP/IP reference model. Note that the MPLS architecture is independent of the protocols of both layer 2 and 3.

MPLS does not specify a new QoS architecture. Rather, it utilizes the principles the DiffServ architecture defined, see section 3.3.2. A main feature of MPLS is the Traffic Engineering (TE) capabilities [AMA99] it provides. TE aims at the optimization of the use of network resources in an intelligent way, thus, improving network performance as a whole. The following overviews the basics of MPLS and provides an insight into its work with the DiffServ as well as its TE capabilities.

3.3.3.1 Basics of MPLS

As known, as a packet of a connectionless network layer protocol (e.g. IP) traverses network routers, routing decisions are done in each router independently. Of course, the information included in routing headers is considerably more than that required to select the next hop. Choosing the next hop can, thus, be considered as a composition of two tasks, the first assigns each packet into a Forwarding Equivalence Class (FEC) and the second maps the selected FEC to a next hop. MPLS addresses that in existing network layer protocols, the assignment of packets to FECs is done each time a packet enters a router. Therefore, it proposes that this assignment is done just once as the packet enters the network. The specific FEC is encoded as a short fixed length value (20 bits long) termed “label” used to select the next hop. In subsequent hops, no assignment to a FEC is done anymore. Rather, the packet is sent to next hop after replacing the old label with the new one. Note that each router may have its own labels, which necessitates the change of the label while forwarding the packet, see Figure 3.11, which shows an example scenario with MPLS deployment. The figure also illustrates the basic principles of MPLS. As the figure shows, the path packets belonging to a specific FEC take is referred to as a Label Switched Path (LSP).

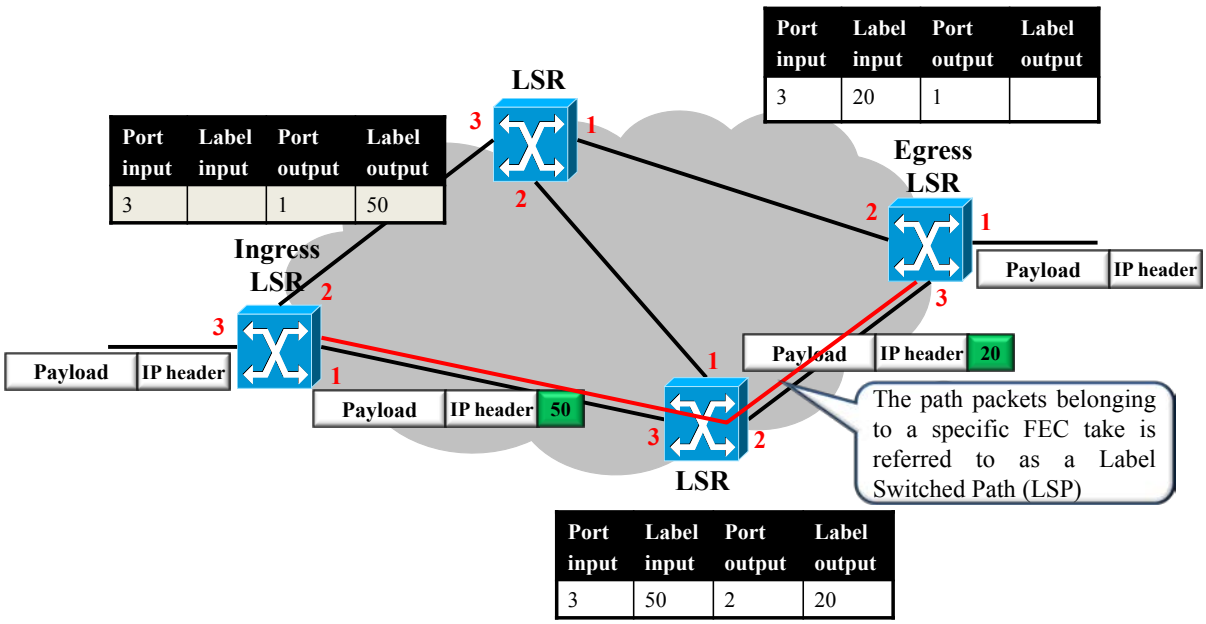


Figure 3.9: An example scenario with MPLS architecture

The operation of MPLS is pretty simple and depicted in the figure. The network layer routing protocol assumed in the example scenario is IP. Once an IP packet enters an MPLS domain, the specific Label Switching Router (LSR)[1] assigns it a label (label 50 in the example) based on the label-FEC bindings[2]. This LSR is called an ingress LSR. Following that, the packet is forwarded to the next hop. Note that the next hop is determined based on the label. The next LSR swaps the old label with a new one (label 20 in the example) and forwards the packet to the next LSR and so on. Once the packet traverses outside of the MPLS domain, the label is deleted and the IP packet is forwarded based on the standard IP protocol.

As we mentioned above, the paths packets take are termed LSPs. The paths are controlled either independently or in an ordered manner, see [RVC01]. Independent LSP control simply mimics the behavior of IP. This means: each LSR makes an independent decision to bind a certain label to a specific FEC. In ordered LSP control, a label is bound to a particular FEC in the ingress LSR[3]. This ensures that traffic of a particular FEC follows a specific path with a particular set of properties.

So as to enable forwarding of packets based on labels, LSRs have to exchange information about how they assign labels to FECs (label-FEC bindings). Such information is exchanged by means of a specific label distribution protocol. In fact, some existing protocols have been extended to piggyback label-FEC bindings, see [RVC01]. New label distribution protocols are also developed, see [RVC01].

3.3.3.2 MPLS and the DiffServ Architecture

As we noted, MPLS uses the principles of the DiffServ architecture. So, the functions implemented in DiffServ ingress and egress nodes are implemented in ingress and egress LSRs, see Figure 3.11. Note that the same PHBs specified in the standard DiffServ architecture are further used.

[1] A router that supports the MPLS architecture.

[2] The rules that determine which labels are assigned to which FECs.

[3] A label is also bound to a specific FEC if the LSR has already received a label binding for that FEC from its LSRs neighbors.

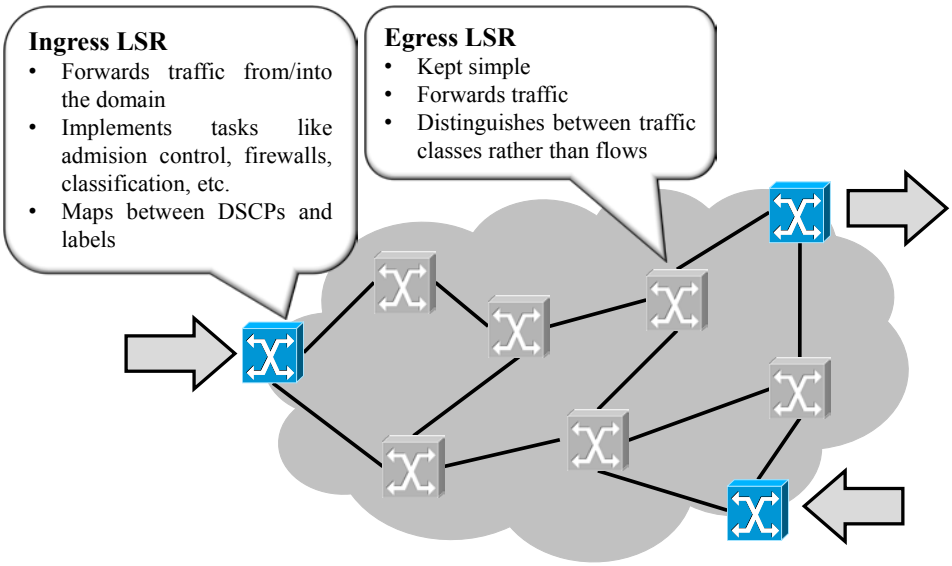


Figure 3.10: The DiffServ architecture over MPLS

So as to support the DiffServ architecture over MPLS, LSRs must be capable of coping with the DSCP values of the DiffServ architecture, i.e. an adequate mapping between DSCPs and labels should be carried out. The problem here is that MPLS specifies only 8 different PHBs (3-bits for Experimental-bits field (EXP) of the MPLS header, see [RRT01]. This came originally from the 3-bits precedence in Type of Service field of the IP header, see [Pos81]). The DiffServ architecture can specify considerable more PHBs (64 possible DSCPs). To achieve such mapping, MPLS uses either an EXP-Inferred-PSC [1] LSP (E-LSP) or a Label-only-Inferred-PSC LSP (L-LSP). E-LSP is used in situations where networks support up to 8 PHBs. In such cases, the EXP field of the MPLS header is sufficient to express the PHB inside MPLS domains. Thus, the label tells the LSR where to forward the packet and the EXP determines the PHB. On the contrary to E-LSP, L-LSP is applied when the number of PHBs the specific domain offers exceeds 8. This implies that it is not possible to express the PHB by means of the EXP field. In such cases, the labels themselves are used to convey PHBs. Sure, some labels will be reserved for PHBs, while the others will be further used to determine the next hop.

3.3.3.3 MPLS Traffic Engineering

Traffic Engineering (TE) refers to the process responsible of selecting a particular path for a specific traffic, so that given constraints (bandwidth, delay, etc.) are not violated [Cisco92]. The goal is to optimize the utilization of network resources and achieve a reliable network operation.

TE involves the following components [Cisco92]:

- Information distribution: responsible of the distribution of information related to the network topology, constraints pertained to links (e.g. available resources), errors, etc.
- Path selection algorithm: responsible of selecting the paths that obey the given constraints.
- Route setup protocol: applied to signal the setup of LSPs.

[1] PHB Scheduling Class (PSC) is a set of one or more PHB(s) applied to a behavior aggregate that belongs to a specific Ordered Aggregate (OA). For instance EF PHB is a PSC with a single PHB, namely the EF itself.

- Link admission control: used to decide which tunnel may have resources and ob a new request can be accommodated or not.
- TE control: responsible of the establishment and maintenance of trunks[1].

The wide-employed route setup protocol is RSVP-TE [ABG01]. The protocol operates in a similar manner as the standard RSVP. The sender sends an RSVP-TE PATH message towards the receiver[2]. The PATH message creates and maintains soft sates. The message also contains, as known, a set of objects. Important are the Explicit Route Object (ERO) and the Record Route Object (RRO). The ERO identifies the route from the sender to receiver. The PATH message follows this route[3]. The RRO keeps track of the LSRs traversed by the PATH message.

The receiver responds by sending an RSVP-TE message towards the sender. The RSVP-TE message follows the path the PATH message followed (stored in the RRO object). It is worth mentioning that the RSVP-TE message results in the distribution of labels for the LSPs between the sender and receiver. These labels are carried with the RESV message in an object named label object. The result is an LSP tunnel between the sender and receiver. All packets belonging to this tunnel are assigned the same label and handled the same within the LSP tunnel.

3.3.3.4 Pros and Cons

Because MPLS utilizes the principles of the DiffServ architecture, it inherits the pros and cons of this architecture, see section 3.3.2.2. Additional advantages of MPLS are obtained from the TE capabilities, which enable the establishment of LSP tunnels with given parameters. This improves the QoS guarantees than can be provided. Sure, the main pros of MPLS is the fast forwarding of data packets due to the dependency on short labels rather than long headers (as by IP). Keep in mind, however, that this implies considerable signaling to distribute labels and operate TE capabilities.

3.3.4 Next Steps In Signaling (NSIS) Framework

The Next Steps In Signaling (NSIS) [HKL05] framework was developed by the IETF NSIS working group [IETFNSIS] and aims at the development of an extensible and generic signaling framework to signal information concerning data flows along their paths in the network. NSIS assumes that the paths data flows take are determined independent of the signaling itself. The signaling problem in this context is similar to that encountered by RSVP. NSIS aims, however, at the generalization of the signaling problem with two intentions. The first states that NSIS framework components should be usable in different parts of the Internet and for different needs without requiring a complete end-to-end deployment. The second highlights that signaling, in principle, is intended for more purposes than just resource reservation, e.g. security, mobility, etc.

Figure 3.11 shows an example scenario with NSIS deployment. The scenario shows a single data flow running between a sender and a receiver via three routers. The example shows that it is not necessary to implement NSIS Entities (NEs) in all network nodes.

---

[1] Trunking is a concept applied to enable communications system to provide clients with network access by sharing a set of links, channels, etc. instead of providing them to each client individually [Tru12].

[2] We assume that the sender and receiver are parts of the MPLS network and operate RSVP-TE.

[3] Note that this object is constructed based on the information distributed inside the network. This object maybe also changed by any LSR on the path. The change is done, of course, using certain administrative policies.

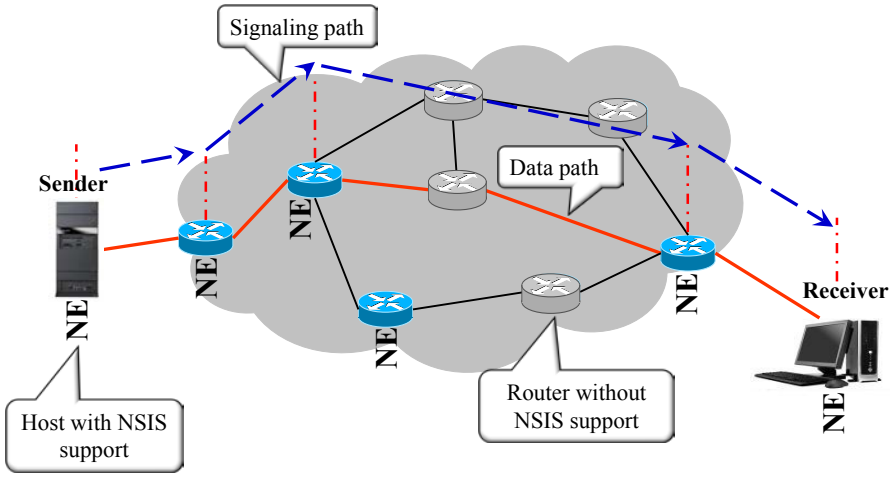


Figure 3.11: An example scenario with NSIS deployment

NSIS considers two basic signaling paradigms, namely a path-coupled and a path-decoupled paradigm. Signaling messages are routed in the path-coupled paradigm only via the NEs existing on the data path, see the figure above. Between NEs, the paths taken by signaling messages may differ from that taken by data. The path-decoupled paradigm does not require that signaling messages be routed only via NEs residing on the data path. NEs not residing on the data path, however, must be aware of the data path.

One notices from the above description that the NSIS framework aims at flexibility. Thus, in order to achieve a modular and flexible solution for the NSIS requirements, the NSIS signaling framework is divided into two layers:

- a lower generic layer responsible for transporting signaling messages in the network independent of any signaling application and
- an upper layer specific for signaling applications.

The lower generic layer is referred to as the NSIS Transport Layer Protocol (NTLP), while the layer responsible for hosting signaling applications uses the term NSIS Signaling Layer Protocol (NSLP). Both layers are displayed in Figure 3.12.

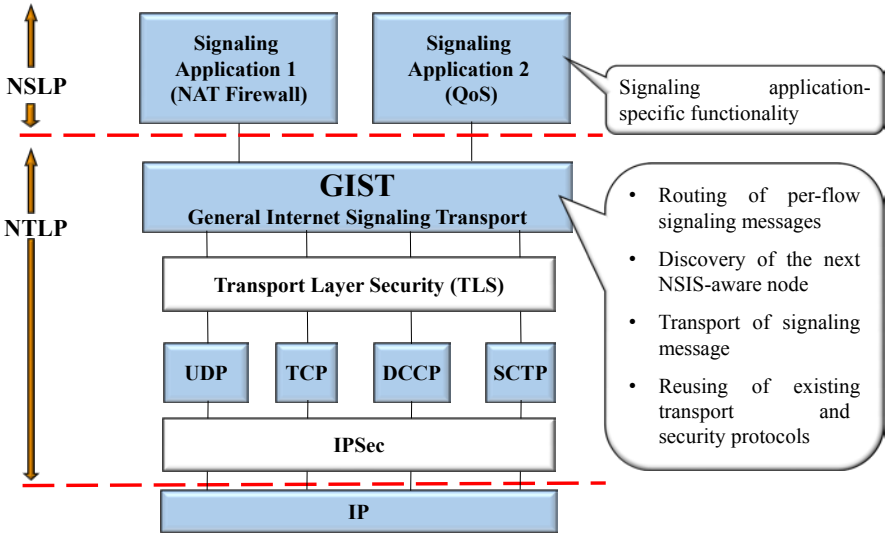


Figure 3.12: Layers of the NSIS framework

Taking a deeper look at Figure 3.12, one notices that the NTLP layer is designed as a structured layer, as well. The design employs pre-existing security and transport protocols, e.g. TLS, UDP, TCP, IPSec, etc., under a common messaging sub-layer named General Internet Signaling Transport (GIST), see [SHa10]. GIST provides a common service for diverse signaling applications to enable the transfer of signaling applications' messages in both directions along the path of their flows. It is worth mentioning that the GIST sub-layer does not handle signaling applications' states. Rather, it only manages its own states and configures the underlying security and transport protocols to ensure the transfer of signaling messages on behalf of signaling applications. More concrete, two main tasks should be solved by the GIST, as [SHa10] highlights, namely:

- Routing: handles the transportation of signaling messages between adjacent NSIS-aware nodes and, if necessary, establishes addressing and identity information about neighbor NSIS-aware nodes.
- Transport: delivers signaling messages to the adjacent NSIS-aware node.

The NSLP layer hosts diverse signaling applications, e.g. IntServ, QoS NSLP [MKM10]. The following provides an insight into well-known applications in this context.

3.3.4.1 QoS NSLP

QoS NSLP is a protocol developed for reserving resources along paths from senders to receivers. The functional design of QoS NSLP is basically similar to that of RSVP. The protocol creates, manages and maintains soft states in NSIS-aware nodes along data paths. QoS NSLP uses four control messages rather than two, as RSVP does. The messages are listed below:

- QUERY message: used to discover available resources along a certain data path.
- RESERVE message: creates, modifies, maintains or deletes reservation states stored in NSIS-aware nodes along a path of data.
- RESPONSE message: an acknowledgement that indicates the receipt of either a RESERVE or QUERY message.
- NOTIFY message: the message notifies in case of errors.

QoS NSLP is flexible since both sender-initiated and receiver-initiated scenarios are possible. Figure 3.13 shows an example network topology with QoS NSLP support. As the figure shows, QoS NSLP Entity (QNE) is an NSIS entity with QoS NSLP support. QoS NSLP Initiator (QNI) stands for the first node in the sequence of QNEs that initiates and issues a reservation request for a session (i.e. the first QNE that issues a RESERVE message). QoS NSLP Receiver (QNR) represents the last node in the sequence of QNEs that receives and handles a reservation request (i.e. the last QNE that consumes the RESERVE message) for a session.

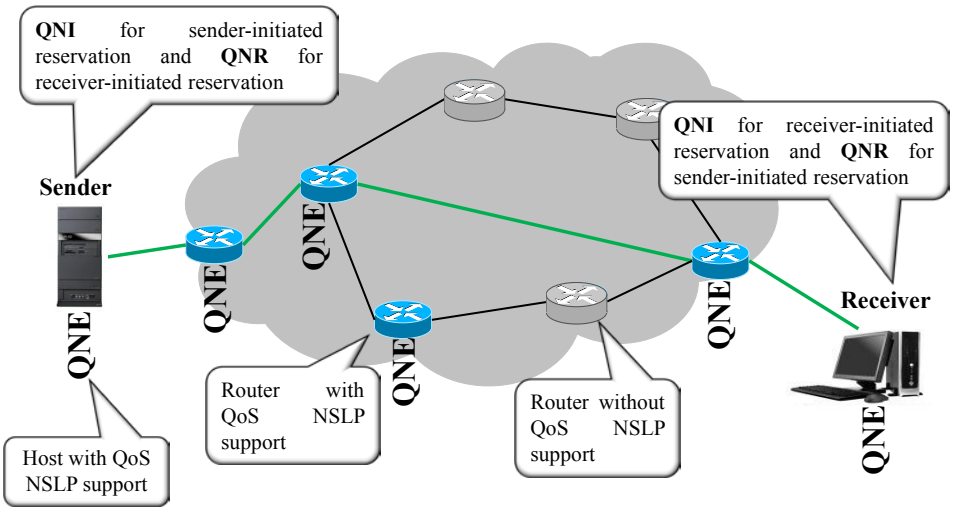


Figure 3.13: An example network topology with QoS NSLP support

Sender-initiated reservation is shown in Figure 3.14. The QNI first constructs a RESERVE message with a QoS SPECification (QSPEC) object carried with it. This object characterizes the QoS required. The message is passed to the GIST, which transfers it to the next QNE. There, the message is delivered by the GIST to the QoS NSLP protocol, which handles it based on the QSPEC object included. Following that, the QoS NSLP protocol generates a new RESERVE message based on the one received and passes it to the GIST that transfers it to the next QNE. The same processing is done in further QNEs along the path up to the QNR, which consumes the incoming RESERVE message and does not forward it further.

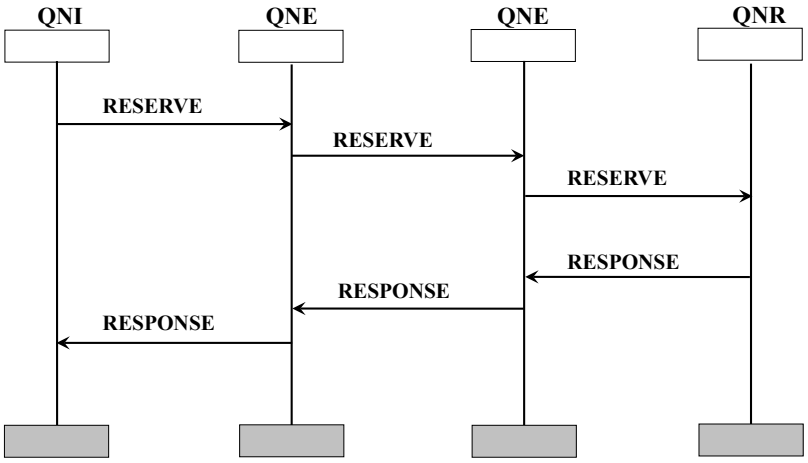


Figure 3.14: Sender-initiated reservation employing QoS NSLP

In the case the RESERVE message includes a Request Identification Information (RII) object, which indicates that a RESPONSE message should be returned to the QNI, the QNR constructs a RESPONSE message and passes it to the GIST, which transfers it back to the QNE that has sent the related RESERVE message. Notice that the resources are reserved along the path by means of the RESERVE message. In this way, the time required to build, refresh or update a session is RTT/2, where Round Trip Time (RTT) denotes the time to exchange a RESERVE and a RESPONSE message between the QNI and the QNR.

The receiver-initiated reservation is displayed in Figure 3.15. The key issue here is that the sender of data (QNR) must trigger the reservation process. For that purpose, the QNR issues a

QUERY message to be transferred by the GIST to the receiver (the QNI). The message gathers information about resources available along the path as well as the ability of concerned QNEs to satisfy the request. This information is carried within a QSPEC object. Notice that no RESPONSE message is sent by any QNE on the path to indicate the receipt of the QUERY message, as Figure 3.15 depicts.

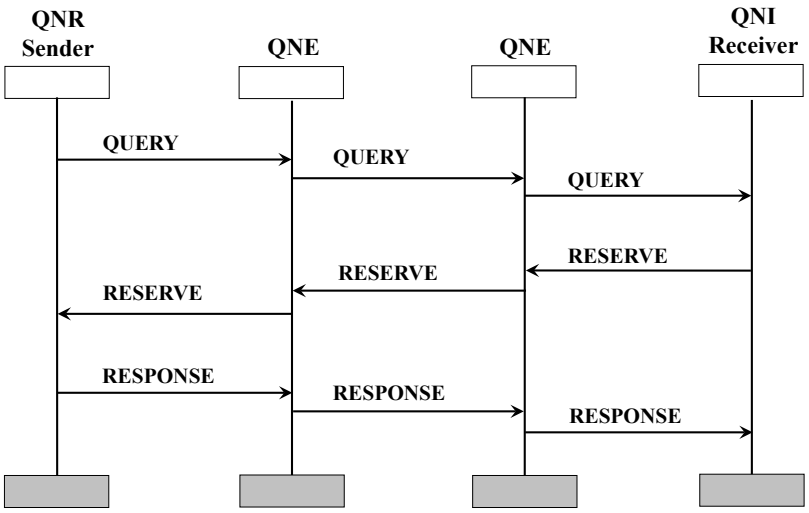


Figure 3.15: Receiver-initiated reservation employing QoS NSLP

As the QNI processes the QUERY message, it constructs a RESERVE message, based on the QUERY message received and passes it to the GIST. The GIST then transfers the message back to the QNR using the same path taken by the QUERY message. The procedure is similar to that discussed for the sender-initiated reservation, as Figure 3.14 illustrates.

The fact that receiver-initiated and sender-initiated reservations are possible is not the only feature that makes QoS NSLP eminently advanced. Additional features contribute to this as well - mentionable are the following:

- Bidirectional reservations can be made either by binding two sender-initiated reservations or sender-initiated and receiver-initiated reservations.
- Various QoS approaches are supported, e.g. IntServ, DiffServ, etc. The key idea is to construct and process the QSPEC object based on the QoS approach being applied.
- QoS NSLP supports layered reservations, also termed layering. Layering simply means that a specific QNE, residing on the path signaling messages traverse, may construct a new QSPEC object that encapsulates the original one if necessary to allow special handling for QoS inside parts of the network, as it will be illustrated in the example below. Layered reservations are helpful in various scenarios. For instance, when certain parts of the network support one or more local QoS models, local transport characteristics (e.g. use of GIST unreliable transfer mode instead of the reliable one), local combination of several per-flow reservations into an aggregate reservation, etc. An example is shown in Figure 3.16. The example shows a scenario that contains a local domain with a local QoS model that is different from that supported by other parts of the network. As the figure illustrates, when one of the QNEs residing on the local domain borders receives a RESERVE message from outside this local domain, the QNE (Ingress QNE) in question constructs a new local QSPEC object that encapsulates the end-to-end QSPEC object. Of course, the local QSPEC object is built based on the end-to-end QSPEC object. QNEs of the local domain handle then the local QSPEC object when they receive the RESERVE message. When the message is

transferred outside the local domain, the local QSPEC object is replaced by the end-to-end QSPEC object.

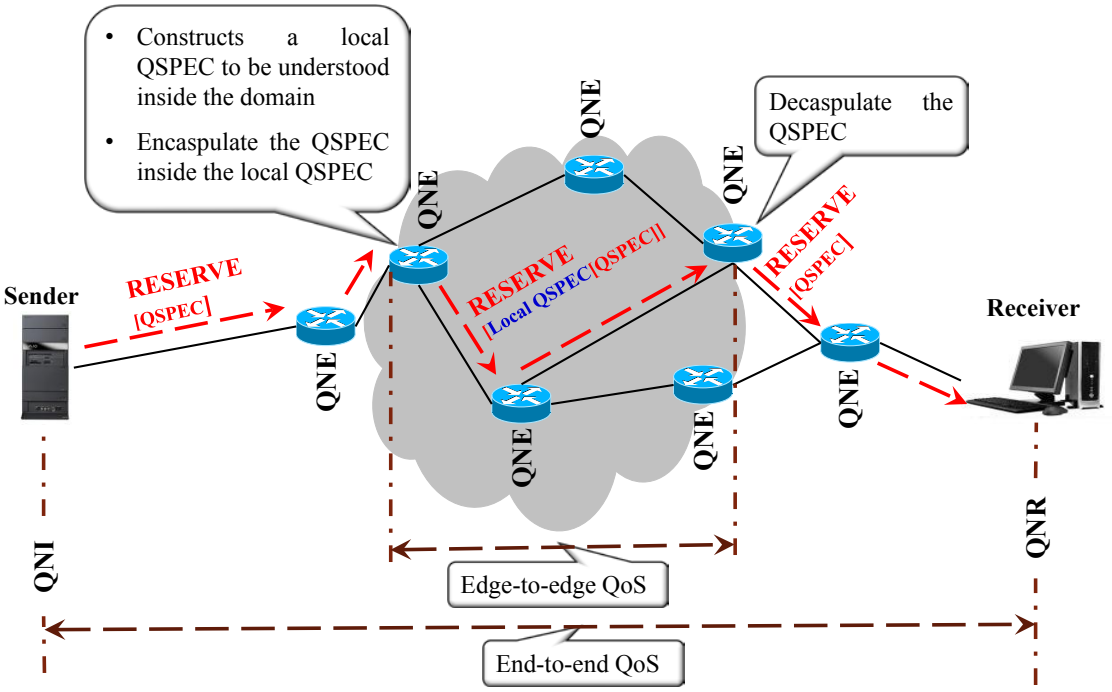


Figure 3.16: An example for layered reservations

- To enhance scalability, QoS NSLP supports reduced refreshes, which simply means that the RESERVE messages used to refresh existing reservations are abbreviated forms of those applied to reserve the resources. In other words, all unnecessary fields are excluded from the mentioned RESERVE messages.
- To reduce signaling traffic traversing the network, QoS NSLP supports summary refreshes and summary tear messages. This simply means that a single control message is sent to refresh/tear down a group of sessions.
- QoS NSLP supports session binding to express relations between sessions. Session binding reflects a unidirectional relation between various sessions. This is helpful in many applications such as video conferencing since a session for voice transmission and another for video are normally established. There is, of course, a dependency between both sessions. Mostly, when one is terminated, the other should be terminated, as well.
- As highlighted above, there may be dependency between sessions. This implies that there could also be relation between control messages. QoS NSLP addresses this issue, as well. So as to reduce reservation states stored in QNEs and, as a result, reduce the processing load for signaling messages, diverse per flow reservations can be combined into one aggregate reservation.

3.3.4.2 Enhancements to the NSIS Framework

NSIS Resource Management in DiffServ (NSIS RMD) QoS model: the NSIS RMD [BWK10] was developed to provide a scalable and dynamic QoS Model (QoSM) within NSIS networks that contain DiffServ domains. The protocol model of NSIS RMD QoSM is shown in Figure 3.17. The figure shows an RMD-enabled DiffServ domain with QNE ingress and egress nodes. Internal nodes can be either QNE or NSIS unaware nodes, as the figure illus-

trates. The QNI and QNR are not part of the RMD domain. Rather, they represent the initiator and receiver of QoS reservation requests.

QNI, QNR and edge nodes of the RMD domain maintain QoS NSLP as well as NTLP states and are, therefore, stateful nodes. Interior NSIS-aware nodes are NTLP stateless and either QoS NSLP stateless or reduced state nodes[1].

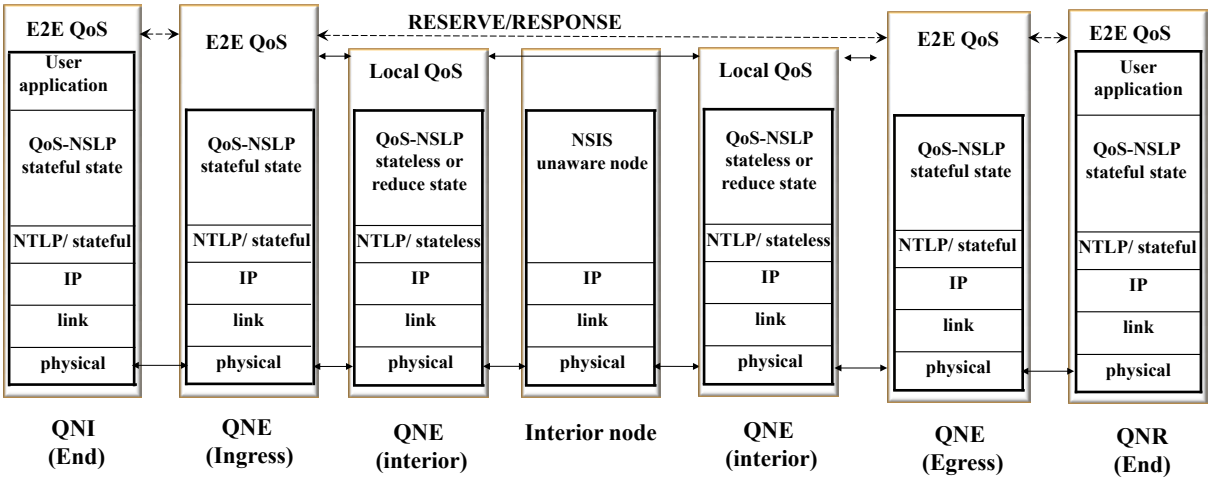


Figure 3.17: Protocol model of NSIS RMD QoSM

Figure 3.18 illustrates the way in which control messages are exchanged within NSIS RMD domains. The figure shows a sender-initiated scenario. Notice that due to the fact that the QoS models used within the RMD domain can be different from those applied outside the domain, the reduced states that interior nodes maintain can be updated independent of the states stored for end-to-end per-flow reservations. This is why one sees different RESERVE messages within the RMD domain (RESERVE and RESERVE` messages) in Figure 3.18. One type of RESERVE messages (RESERVE in the figure) is associated with end-to-end per-flow reservation states, while the other type (RESERVE` in the figure) is associated with the reduced states stored within the RMD domain.

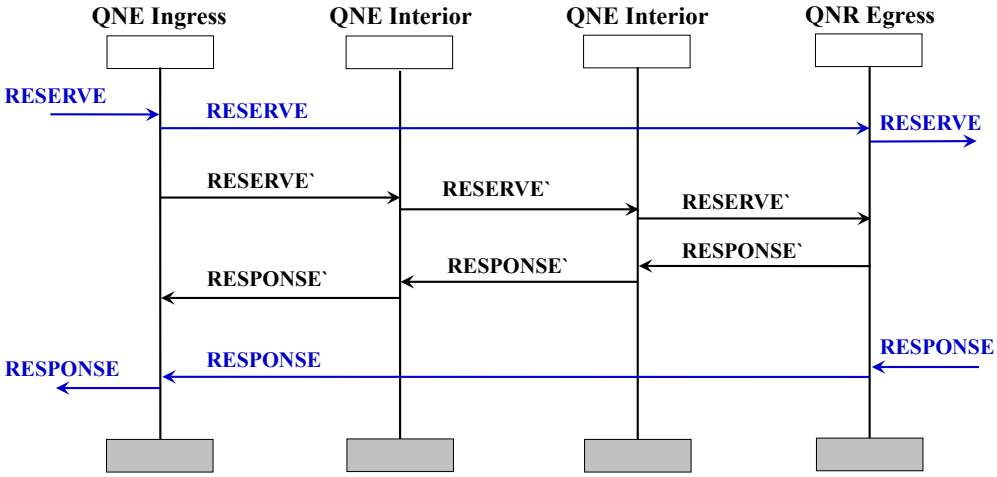


Figure 3.18: Sender-initiated reservation with reduced state interior nodes

A sender-initiated reservation is done as follows: first, a RESERVE message is created by the QNI and transmitted with an adequate QSPEC to the QNR. When the RESERVE message

[1] Interior NSIS-aware nodes will be QoS NSLP stateless for the measurement-based admission control mechanism and QoS NSLP reduced state for the reservation-based admission control mechanism since interior nodes store reduced states for PHBs, see section 3.3.2.1 for more details.

arrives to an RMD ingress node, a local RMD QSPEC is constructed by the ingress node based on the original QSPEC carried by the original RESERVE message. The RMD QSPEC is transmitted with an intra-domain RESERVE message (RESERVE` in the figure) to the QNR. Of course, the intra-domain RESERVE message is also sent applying the GIST datagram signaling mechanism and results in the reservation of resources inside the RMD domain. Meanwhile, the original RESERVE message is also sent to the QNR but does not result is reserving resources. When the RESERVE` message reaches an egress node, a local RESPONSE message (RESPONSE` in the figure) is transmitted back to the ingress node. Moreover, when the egress node receives the original RESERVE message, it forwards it towards the QNR. Notice that when the egress node receives a RESPONSE message, it forwards the message to the QNI, see Figure 3.18.

Signaling IntServ controlled-load service with NSIS: as mentioned in section 3.3.1, the controlled-load service aims at realizing the concept stating: let the end-to-end behavior that is visible to applications running between two end-points tightly approximates the behavior that the applications observe when they exchange best-effort traffic in low-loaded networks. This type of service was mainly developed for the IntServ architecture, see section 3.3.1 for details. This section describes how the controlled-load service can also be signaled within the NSIS framework. That is, how RSVP control messages are mapped to QoS NSLP messages, in order to support the controlled-load service in networks supporting both protocols.

QoS NSLP carries QoS-specific information in objects termed QSPEC, as mentioned in section 3.3.4.1. There are many types of QSPEC objects, namely QoS-Desired, QoS-Available, QoS-Reserved and QoS-Minimum objects. Of course, only a subset of these objects is normally carried by a single control message. The QoS-Desired object contains the QoS parameters that the application prefers. The QoS-Available object is used to describe the QoS parameters that can be offered to the application. The QoS-Reserved object describes the resources actually reserved. The QoS-Minimum object is normally included with the QoS-Desired object, in order to signal that the resources specified in the QoS-Desired object can be degraded to the level specified in the QoS-Minimum object in case not enough resources are available. In order to support controlled-load service in NSIS networks, it is essential to translate the RSVP objects into the NSIS QSPEC objects presented above. It is also essential to map between RSVP control messages and QoS NSLP control messages. Based on [KFS12], the RSVP PATH message is mapped onto the QoS NSLP QUERY message, the QoS NSLP RESERVE message is sent instead of the RSVP RESV message, while the RSVP RESVConf message is replaced by the QoS NSLP RESPONSE message. The objects included in the mentioned control messages are presented in Table 3-1.

	Message	Objects
RSVP	PATH	SENDER_TSPEC ADSPEC
QoS NSLP	QUERY	QoS-Desired QoS-Available QoS-Minimum
RSVP	RESV	FlOWSPEC

QoS NSLP	RESERVE	QoS-Desired QoS-Available
RSVP	RESVConf	
QoS NSLP	RESPONSE	QoS-Reserved

Table 3-1: The objects included in RSVP and QoS NSLP control messages

An RSVP PATH message normally carries a SENDER_TSPEC and an optional ADSPEC object. The SENDER_TSPEC specifies the traffic that the sender will send, while the AD-SPEC object carries the advertising OPWA information for the flow, as described in section 3.3.1.1. The RESV message carries, in typical operation scenarios, the FLOWSPEC object to the sender. This message results in reserving resources, as mentioned earlier. If a confirmation is required, the sender transmits a RESVConf message to the receiver to indicate that the end-to-end reservation has been established successfully.

QoS NSLP carries objects with similar tasks to those carried by RSVP control messages. The QNI sends a QUERY message with a QoS-Desired, an optional QoS-Available and an optional QoS-Minimum object. The QoS-Desired object specifies the resources desired by the sender, while the QoS-Available object reflects what is actually available. The QoS-Minimum indicates to which grade QoS is allowed to be degraded. The RESERVE message that the QNR issues to the QNI normally conveys the same objects that the QUERY message carried and results in reserving resources. As a confirmation, a RESPONSE message with a QoS-Reserved object is sent back to the QNR.

In brief, routers that support both RSVP and QoS NSLP and are border routers to either RSVP or NSIS sub-domains should translate between the control messages that we presented above so that controlled-load service is provided, see Figure 3.19. Note that the sender as well as receiver in the figure has support for RSVP only. The translation between RSVP and QoS NSLP control messages is done in the border routers of the QoS NSLP domain, as the figure shows.

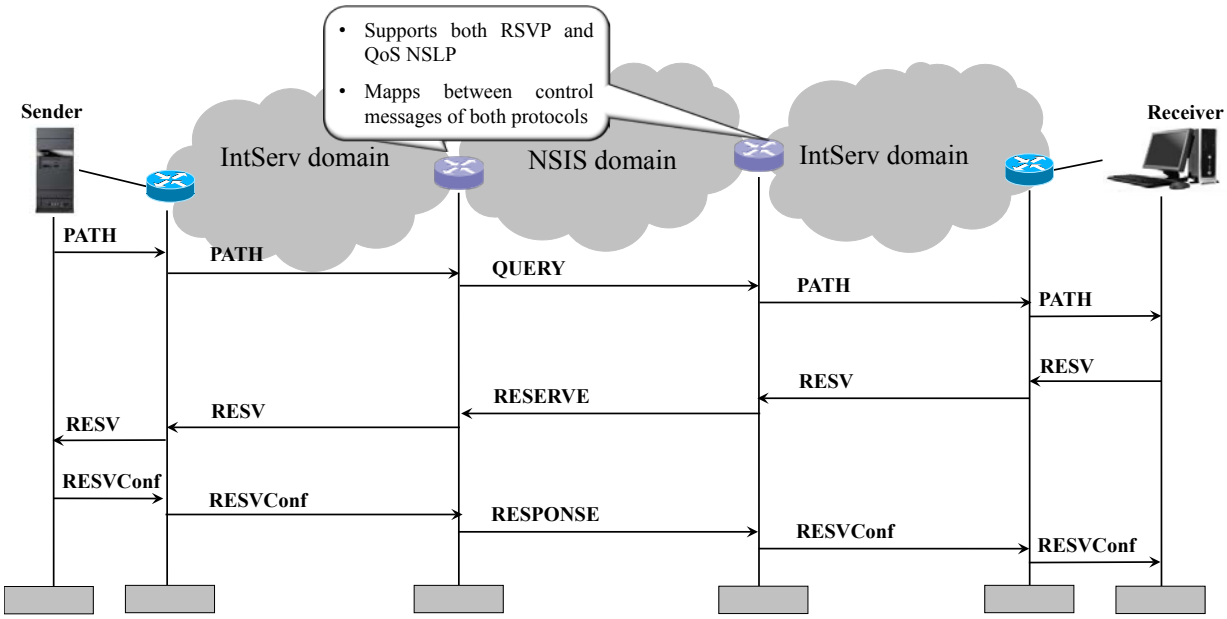


Figure 3.19: Provision of controlled-load service in networks containing IntServ/RSVP and QoS NSLP domains

3.3.4.3 Pros and Cons

After providing an overview of the NSIS framework, let us now discuss the related pros and cons. As mentioned earlier, NSIS presents a generic signaling framework to signal information concerning data flows along their paths in the network. In other words, the NSIS framework supports a wide range of signaling applications that may have goals far beyond the reservation of resources, for security aspects for instance. Even for providing QoS guarantees, various QoS signaling applications with different QoS models including the IntServ and DiffServ architectures can be applied. This enables simple application of NSIS framework to new- and already-deployed networks. Furthermore, the architecture of NSIS is flexible and also extensible in terms of integration of new signaling applications and new NSIS-aware nodes. Moreover, the architecture of NSIS enables existing protocols such as TCP, UDP, TLS, etc. to be used.

In contrast to IntServ and DiffServ architectures, the NSIS framework considers mobility by concept. The basic concept concerning mobility depends on assigning each new session a unique session identifier upon establishing the session. Note that the session may include various flows. After the MN moves to a new point of attachment, it begins establishing a new session. The new session is established using the same session identifier the MN had before the movement. Session flows identifiers, however, are assigned new. Sure, the establishment of the session implies exchanging signaling messages. As the signaling passes through an NSIS-aware node that has the same session identifier, this node maps the new flow identifiers to the old ones and replies to the MN directly. In this way, mobility is supported in a simple fashion. Of course, the focus in this context is on the avoidance of a full re-establishment of new sessions after movements rather than the support of fast and seamless mobility.

Since QoS NSLP is developed for the NSIS framework, the advantages of this protocol are advantages of the NSIS framework as well, e.g. the support of sender-initiated, receiver-initiated and bidirectional reservations, the support of layered reservations, support of reduced and summary refreshes, etc.

In addition to the wide range of advantages we mentioned, there are also many disadvantages of the NSIS framework. First, this framework is still under development. Therefore, there are many open issues under discussion, especially concerning its implementation. Although the architecture of NSIS offers scalability and extensibility, it suffers from complexity. As one can see, NSIS-aware nodes are more complex than IntServ- or DiffServ-enabled nodes. Furthermore, there is no support for multicast yet.

3.3.5 Qualitative Analysis

This section summarizes the chapter by a qualitative comparison of the main four QoS architectures presented, namely IntServ, DiffServ, MPLS (with the DiffServ architecture and the TE capabilities) and NSIS, with respect to the

- type of resource reservation,
- scope of resource reservation,
- initiation of resource reservation,
- states stored in nodes,
- transport protocols used,
- scalability,
- complexity,
- QoS guarantees that can be provided,
- support of mobility and
- security

is presented in Table 3-2. The first 7 metrics are important from the service provider/network operator point of view. These metrics relate to questions like, how resources are reserved, which network nodes are allowed to initiate a new reservation, how reservation states are maintained, how complex is the architecture, does it scale, etc. the last three metrics are of interest from the user side. Notice that users care of the QoS guarantees they get. Furthermore, they are interested in knowing whether they can be mobile, whether they are secure, etc.

Architecture	IntServ	DiffServ	MPLS	NSIS
Type of resource reservation	Per-flow	Per-class (aggregated)	Per-class (aggregated)	Both
Scope of resource reservation	End-to-end (assuming that all nodes between the sender and receiver are IntServ-enabled)	Edge-to-edge (assuming that the sender and receiver locate outside the domain)	Edge-to-edge (assuming that the sender and receiver locate outside the domain)	• End-to-end • Edge-to-edge • Host-to-edge
Initiation of resource reservation	Receiver-initiated	Controlled by ingress routers	• Controlled by ingress LSRs (when RSVP-TE not used) • Receiver-initiated (when RSVP-TE is applied)	• Receiver-initiated • Sender-initiated
States stored in nodes	Soft states (per flow)	No states (The standard architecture does not necessitate storing states. However, most implementations and some extensions to the standard architecture require creating and maintaining states for classes handled)	• No states (when RSVP-TE not used) • Soft states (when RSVP-TE is applied)	• Soft states, NSIS can store full states (termed as stateful) • Reduced states • No state (termed as stateless)
Transport of control messages	UDP	UDP	UDP	UDP/ TCP
Scalability[1]	P	G:VG	G:VG	VG
Complexity	Low	Middle	Middle	High
QoS guarantees that can be provided	• Highest QoS guarantees (guaranteed services) • Low-middle QoS guarantees (controlled-load services)	• Relative high QoS guarantees (EF PHB) • Middle QoS guarantees (AF PHB)	• Relative high QoS guarantees (EF PHB) • Middle QoS guarantees (AF PHB)	Depends on the QoS signaling application implemented
Support of Mobility	No	No	No	Yes
Security	Not addressed	Not addressed	Not addressed	Supported by reusing pre-existing security protocols

Table 3-2: Qualitative comparison of IntServ, DiffServ, MPLS and NSIS

In terms of the type of resource reservation, the IntServ architecture reserves resources for each flow, while DiffServ and MPLS aggregate flows into classes. The NSIS architecture

[1] **P**: Poor, **M**: Middle, **G**: Good, **VG**: Very good. When this field contains, for example, **M:G**, this indicates a variation between middle and good.

may contain per-flow and/or per-class QoS models. This depends on the QoS application being implemented.

With respect to the scope of resource reservation, the IntServ architecture aims at supporting end-to-end reservations. Notice that the sender and receiver should be IntServ-enabled nodes and all nodes in-between are aware of IntServ, as well. DiffServ and MPLS architecture support edge-to-edge reservations since resources are negotiated between ingress and egress routers/LSRs of the domain. NSIS, as mentioned above, is capable of supporting various scenarios. Thus, it can provide end-to-end, edge-to-edge and host-to-edge reservations. This depends on the QoS signaling application being operated.

Let us now discuss how resource reservation procedures are initiated as well as which kinds of states are stored inside nodes when employing the three studied architectures. As described earlier, the IntServ architecture employs a receiver-initiated resource reservation protocol, namely RSVP, which stores soft states inside routers in addition to end-hosts. The NSIS architecture is capable of supporting sender- and receiver-initiated resource reservation scenarios. The main protocol developed is QoS NSLP, which also maintains soft states. However, as NSIS is a framework capable of supporting diverse QoS signaling applications, various types of states can be maintained accordingly. Thus, NSIS supports stateful as well as reduced states. Moreover, NSIS-aware nodes may be stateless when no storing of states is necessary, as we mentioned while discussing the NSIS RMD. Reservation within DiffServ domains is done in a different manner since the reservation is controlled by ingress routers and achieved based on domain-wide polices. In the basic architecture, no states are maintained. However, in most implementations and enhancements, per-class states are necessary. Considering MPLS, reservations of LSPs are controlled by ingress LSRs when MPLS is applied within a DiffServ architecture. However, when RSVP-TE is used, LSPs are reserved based on a receiver-initiated basis. No states are stored when RSVP-TE not used, while soft sates are maintained in case RSVP-TE is applied.

Considering the transport protocols used to transfer control messages, one notices that IntServ, DiffServ and MPLS use UDP, while NSIS is capable of utilizing both UDP and TCP. Regarding scalability and complexity, Table 3-2 shows that the IntServ architecture scales poorly. However, it is the least complex architecture among the three. DiffServ and MPLS scale better than IntServ but are also more complex. The best architecture in terms of scalability is NSIS. It is, however, the most complex one compared to IntServ, DiffServ and MPLS.

Let us now consider the QoS guarantees that the studied architectures are capable of providing. The IntServ architecture supports either guaranteed or controlled-load services. Guaranteed services provide the highest QoS guarantee, while controlled-load services attempt to enforce networks to behave as they do when carrying low loads, even when they are heavily loaded. In other words, these services provide low-high QoS guarantees. The DiffServ architecture provides, in general, medium QoS guarantees. It operates two PHBs, namely EF and AF. Both PHBs do not provide as hard guarantee as the guaranteed service, the IntServ architecture supports. They are, however, in most scenarios better than controlled-load services. MPLS is similar to the DiffServ architecture in terms of QoS guarantees. NSIS provides various QoS guarantees varying from hard to loose ones based on the QoS signaling application supported.

With respect to the support of mobility, the chapter showed that mobility is only supported by the NSIS architecture. Security is also considered when developing the NSIS architecture in the context of the possibility to utilize pre-existing security protocols, e.g. TLS. IntServ, DiffServ and MPLS do not consider security as a part of their architecture.

Chapter 4: Coupling between QoS and Mobility Management Solutions

As mentioned in previous chapters, there should be solutions capable of simultaneously handling mobility management and QoS to meet the requirements of future all-IP networks. The basic principle to do this is to couple the solutions for mobility management and QoS, so that handoffs are accomplished and in-parallel required resources reserved. To provide an insight into how such coupling can be achieved, this chapter reviews well-known techniques capable of coupling between mobility and QoS solutions.

The rest of this chapter is organized as follows: section 4.1 discusses how does mobility of users affect QoS, while section 4.2 presents how mobility management techniques can be coupled with QoS mechanisms. Following that, well-known approaches accomplishing such a coupling are presented in section 4.3, 4.4 and 4.5. Section 4.6 provides a qualitative comparison between the reviewed approaches with respect to the tunneling problem, triangular routing problem, double reservation of resources during handoffs, passive reservation, dependency on layer 2 triggers, network topology, new nodes that should be introduced to the network, nodes that should be updated and security. Finally, section 4.7 concludes this chapter with the main results.

4.1 How Does Mobility of Users Affect QoS?

As we mentioned in section 3.1.1, the QoS focuses on users' satisfaction. Thus, mobility will negatively impact QoS if services disruptions result due to the mobility of users. In the following, we will discuss how an example basic QoS mechanism (IntServ) will interwork with mobility protocols. The example applies RSVP as a resource reservation protocol, while the mobility protocol used for clarification is the basic protocol, MIP. Following that, some issues concerning the other studied QoS mechanisms will be mentioned.

Let us first assume that the MN resides in the range of a FA and resources are reserved on the path from the CN to the FA serving the MN. Let us accept at this moment that resources are reserved on the whole path, although this is not really true. We will explain the reason later on in this section. When the MN moves to a new subnet served by a new FA, the MN loses the connection and stops receiving data packets. Note that data packets are further sent since the network is not notified of the new location yet. The MN will first establish a new wireless link with the new detected AP. Thereafter, it has to detect the change in the subnet, which in case of its occurrence prompts a configuration of a new CoA and then a new registration with the HA. The registration is done, of course, via MIP, see Figure 4.1 (a).

After the HA gets notified of the new CoA, downlink data packets are sent to the new location of the MN. Transmitting uplink data packets, however, starts after the MN completes the MIP procedure. Note that neither downlink nor uplink data packets obtain QoS guarantees at this moment since no resources are reserved yet, this implies that they are sent as best-effort, see the figure. Of course, other impacts appear as well. For instance, MIP uses triangular routing paths which contribute without intention to increase the end-to-end delay and, as known, packets delivered with delays exceeding a play-out time are considered lost. Sure, best-effort packets are a subject to such delays, while packets handled with QoS guarantees rarely see such situations.

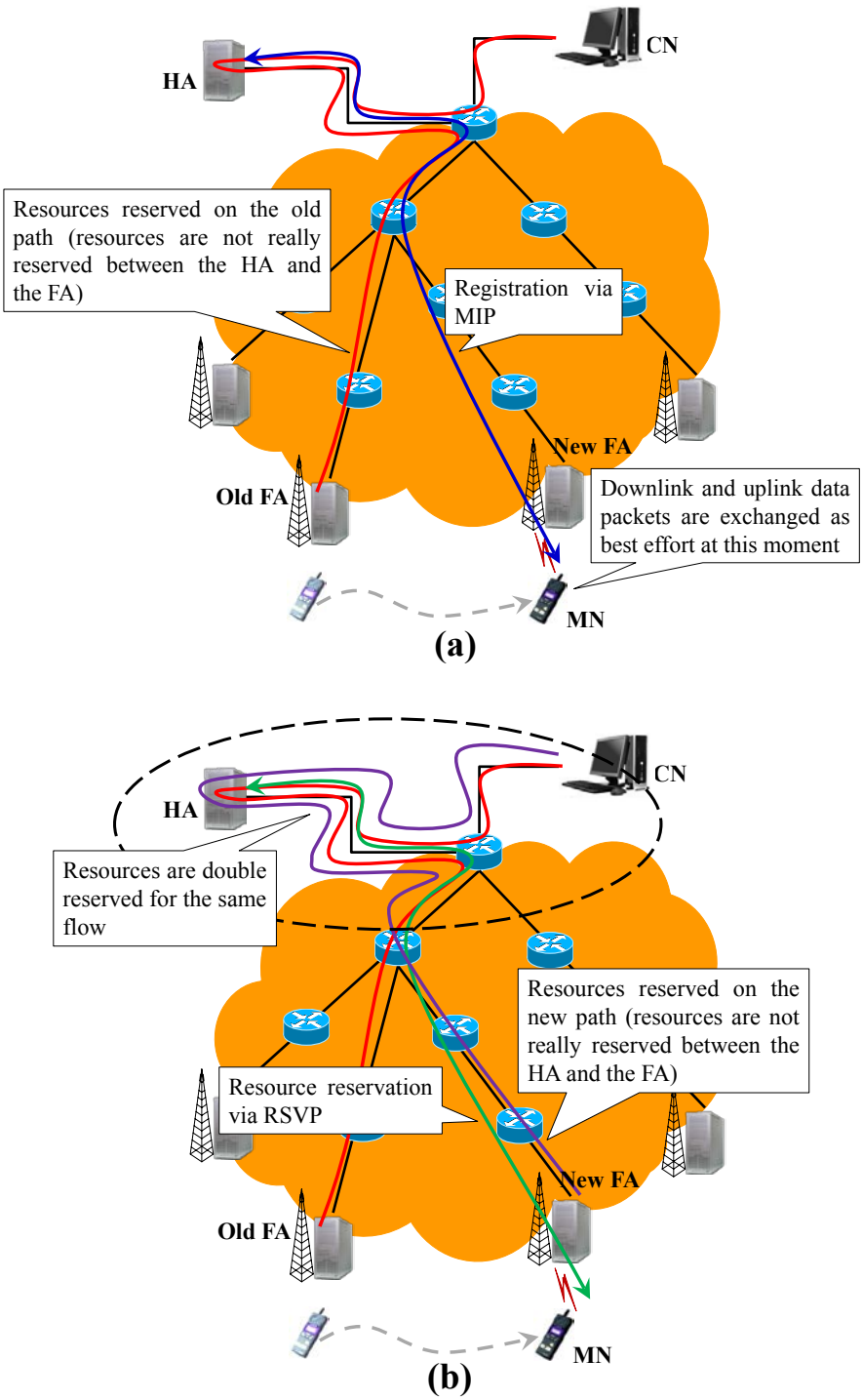


Figure 4.1: Interwork between RSVP and MIP

The exchange of data as best-effort remains until the resource reservation lifetime is about to expire. After that, RSVP comes into play. Note that RSVP does not note that the path has been changed, it only refreshes the reservations. Assuming the uplink session first, the refresh procedure is done, as already known, by sending a PATH message to the CN, which responds by a RESV message. Note that the PATH message is routed via the standard IP, while RESV message traverses the HA. This presents a serious problem since the RESV message has to follow the same path the concerning PATH message followed. The solution for such problem is the use of a reverse tunnel, which forces both PATH and RESV messages to be exchanged via the HA. Exchanging RSVP control messages on a triangular route means, however, that

the RESV message will, in practice, be tunneled to the new location of the MN, where it is de-tunneled and forwarded to its destination. So all routers on the path between the HA and the FA serving the MN will not recognize the RESV message and, as a result, will not reserve resources. Similar behavior is seen by the downlink RSVP session. The problem mentioned is known as tunneling problem. Keep in mind that we have assumed above that the resources are reserved on the whole path and noted that this is not true. The reason is the tunneling problem we already discussed. After the resource reservation procedure (or more accurate the refresh procedure) is accomplished, data packets obtain the contracted QoS guarantees. Note, however, that data packets are sent via the tunnel from the HA to the current subnet without QoS guarantees.

The figure also shows that resources are double reserved on parts of the path between the CN and the MN. More concrete, they are double reserved between the HA and the CN, see Figure 4.1 (b). Note that the figure shows doubled resources on other parts of the path. This is only because the semantic of RSVP is end-to-end. In practice there is no resources reserved between the HA and the FA that serves the MN. Although the resources reserved on the old path will be released after timeout, they are blocked for other users before they are released. This is, for sure, a serious problem.

Let us now summarize the issues discussed in this section:

1. Due to the movement, the MN loses the connection with the old FA. This implies that the MN loses its data.
2. Receipt of data (as best-effort) starts after the MN completes the handoff procedure.
3. Receipt of data as best-effort remains until RSVP notes that it has to refresh the reservation.
4. After the resources are reserved on the new path, QoS guarantees are retained (only on parts of the path).
5. Main problems noticed include the tunneling problem and double resource reservation.

Note that to minimize the number of dropped packets, handoff latency must be minimized. To accelerate the retaining of QoS guarantees, RSVP must be notified directly after the handoff is declared to be completed or even before. Furthermore, other work around is necessary to handle the tunneling and double resource reservation problems.

Similar problems arise when using DiffServ or NSIS frameworks instead of the IntServ architecture. The differences are in the applied QoS provision mechanisms that may differ. In addition, the double resource reservation problem disappears when employing NSIS because it notices that the reservation is for the same flow.

All in all, mobility must be managed while keeping QoS in mind. So, to enable proper interaction between mobility management protocols and QoS mechanisms, they should be coupled with each other. How such coupling can be achieved is handled in the next section.

4.2 How Can Mobility and QoS Techniques be Coupled?

There are three basic strategies to couple mobility management solutions and QoS mechanisms, see [GRu05] and [MLM02]. The first strategy attempts to integrate the solutions for both QoS and mobility in a single protocol. Mostly, new extensions to solutions of mobility management are implemented so that QoS can be handled or vice versa. The approaches that follow this strategy are referred to as hard-coupled solutions[1]. These approaches are stated to perform very well and be efficient as both tasks, mobility and QoS, are considered in their

[1] Hard-coupled solutions are referred to sometimes in the literature as closely-coupled solutions [MLM02], tight-coupling solutions and integrated solutions [MIND02].

design, see [GRu05]. This, however, makes them complex[1] and less applicable to current as well as future networks, since they require many changes in network nodes and even topologies, see [Man03] and [LVM01]. Well-known examples are the Wireless Lightweight Reservation Protocol (WLRP) [Par03] and mobile extensions to RSVP [AAg97].

Other researchers argue that solutions for mobility management and QoS should be kept separate. However, the operation of one affects the operation of the other. Thus, these solutions are termed loose-coupled solutions and do not perform as well and nor are they as efficient as compared to hard-coupled solutions. Keep in mind that because both QoS and mobility are handled separately, the signaling resulting is more than when handling them in one protocol as hard-coupling solutions do, see [MLM02]. They are, however, less complex and more applicable to current and future networks. Most existing approaches for the coupling between mobility management and QoS follow this strategy. Well-known examples are Mobile RSVP (MRSVP) [TBB01], Hierarchical Mobile RSVP (HMRSVP) [TLL03], Localized RSVP (LRSVP) [MRa03], NSIS with Advanced reservations [LKL08] and NSIS-based semi-proactive resource reservation [TMT08].

The third strategy represents a compromise between the both strategies mentioned above. The basic idea is to keep the solutions for mobility management and QoS separate from the implementation point of view (same as loose-coupled solutions). Both solutions, however, should work together, so that they look like one protocol, also similar to hard-coupled solutions. In this way, such hybrid solutions inherit the properties of both hard- and loose-coupled ones. Most existing hybrid techniques, however, perform worse and are less efficient than hard-coupled solutions. They do, however perform better and are more efficient than loose-coupled ones. Furthermore, hybrid techniques are less complex and more applicable to current and future networks than hard-coupled solutions. Compared to loose-coupled techniques, they are more complex and less applicable, see [AMD07] and [AMD10]. Well-known examples are RSVP and MIPv6 interoperation framework [SSL01], QoS extension for NSIS in MIPv6 environments [LPN07], etc.

The following provides an insight into the three strategies and investigates well-known solutions present. The investigation focuses on presenting the basic idea of each approach, its operation overview and pros and cons focusing on the

- triangular routing problem,
- tunneling problem, which is practically a consequence of the triangular routing,
- double reservation of resources,
- reservation of resources in adjacent subnets, also known as passive reservation,
- dependency on layer 2 triggers,
- deployed network topology,
- new nodes that should be introduced to the network,
- nodes that should be updated and
- security.

4.3 Hard-Coupled Approaches

As mentioned in section 4.2, hard-coupled solutions attempt to build new extensions for existing protocols or propose new approaches that support mobility simultaneously with QoS. To provide more detailed view, this section reviews well-known hard-coupled solutions.

[1] In terms of implementation.

4.3.1 Wireless Lightweight Reservation Protocol (WLRP)

WLRP employs RSVP to allocate resources in wireless environments [Par03]. It aims at overcoming the scalability problem of RSVP through lower per connection state storage[1] and fewer control messages.

Network topology and basic principles: Figure 4.2 illustrates the network topology and basic operation principles of WLRP. Reservations in the wired part are handled via the standard RSVP. The basic idea here is to reserve resources passively in neighbor BSs, to which the MN probably may move. Passive resources will be activated after handoffs, thus, QoS guarantee quickly retrieved.

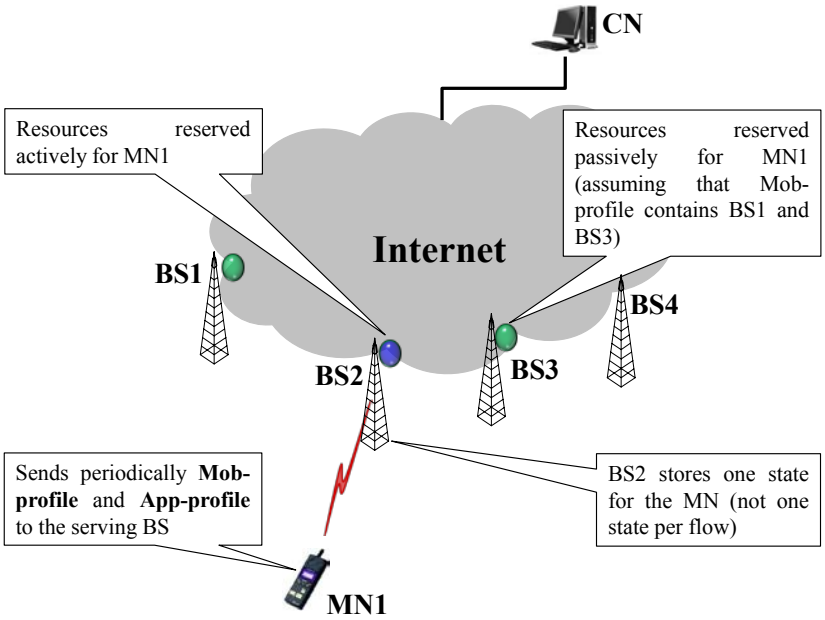


Figure 4.2: Network topology and basic principles employing WLRP

Operation overview: as the figure shows, the MN sends periodic reports to its serving BS. This periodic reports contain a mobility as well as an application profile of the MN (these profiles are referred to as Mob-Profile and App-Profile, respectively). The Mob-Profile specifies the BSs that the MN may move to in the future, while the App-Profile comprises the QoS parameters (i.e. loss negotiability[2], loss profile[3] and handover quality[4]) that the MN aims at having. The serving BS requests the BSs existing in the Mob-Profile to passively reserve the resources satisfying the QoS parameters included in the App-Profile. This is achieved by sending a Passive PATH message to each of these BSs [MSi00], see Figure 4.3.

[1] In contrast to RSVP, which stores one state for each flow, WLRP stores one state per connection. In other words, it stores a state for each MN that currently resides in the range of the BS or has reserved resources passively in this BS. This state conveys all flows of the MN.

[2] Loss negotiability quantifies the QoS degradation that the currently running application is capable of tolerating. This parameter is necessary to avoid rejecting or dropping the reservation in the case of overload.

[3] Loss profile can be either a distributed or bursty loss. Distributed loss means that the loss of separate packets can be tolerated, e.g. when operating video applications. In contrast, bursty loss indicates that any loss in a part of a data frame means the loss of the whole frame; this is the case for FTP applications.

[4] Handover quality indicates the level of service the application expects during handoffs.

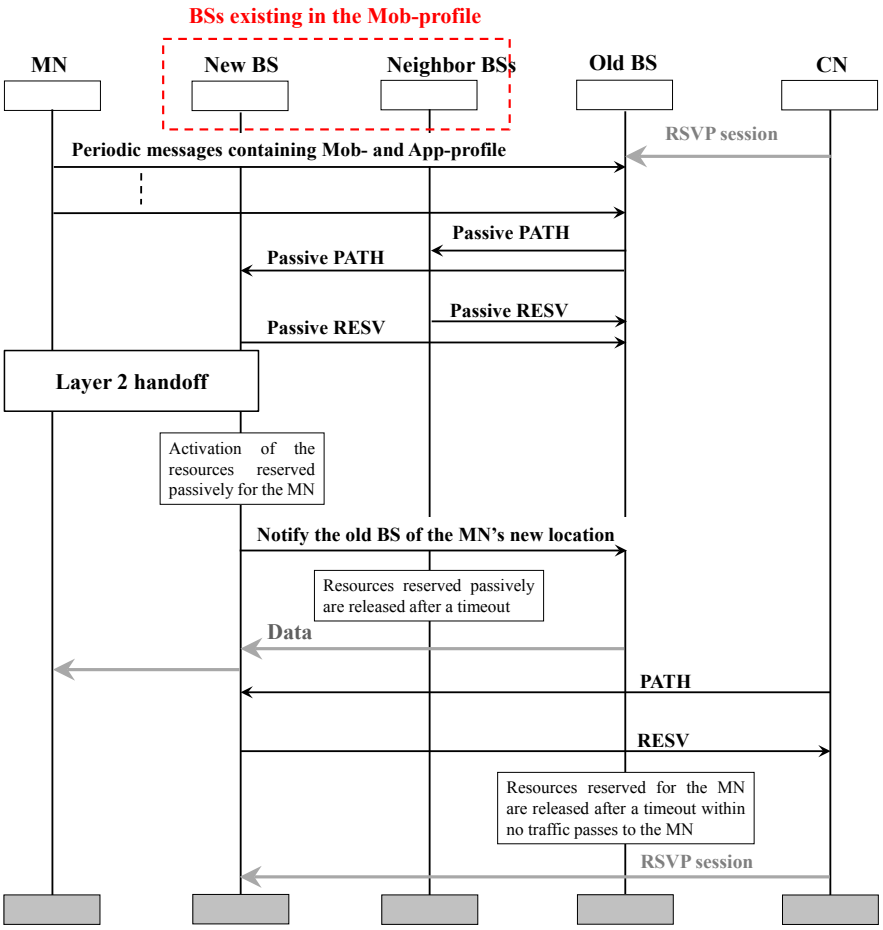


Figure 4.3: Operation overview of WLRP

Once a BS receives a Passive PATH message, it responds by sending a Passive RESV message and reserves resources for the MN passively. The BS uses the reserved resources, however, for best-effort traffic as long as the MN is located outside the range of this BS. Once the MN moves into the range of a new BS, the resources passively reserved for the MN will be activated. To achieve a smooth handoff, the new BS notifies the old one, which forwards the data destined for the MN to the new location. In this way, the MN guarantees its QoS. Other resources reserved passively for the MN on other BSs are released after a timeout. Active reservations employing WLRP are released either explicitly using TEARDOWN messages or after a timeout, during which no traffic passes the network to the MN.

Performance evaluation: simulation results presented in [Par03] proof that employing WLRP significantly reduces the blocking and dropping probability as compared to best-effort service. WLRP is capable of providing seamless handoffs and simultaneously reserving resources on the new BS. As compared to RSVP, WLRP does not require sending refresh messages to refresh the reservations since active reservations are hard and should be released by means of TEARDOWN messages or after the resources are not used for a certain time.

Pros and cons: as WLRP considers only the wireless part of the network and leaves the standard RSVP operating in the wired part, it does not address the tunneling problem[1]. For the same reason, the triangular routing problem as well as the double reservation of resources

[1] The tunneling problem is faced when the MN, residing in a range of a foreign subnet and reserving an end-to-end session to the CN, communicates with the CN via the HA (triangular route). The HA must tunnel data packets to the MN's CoA. These tunneled packets are not recognized by the reserved end-to-end session.

during handoffs is not addressed as well. WLRP relies on passive reservation of resources to achieve the performance improvements we mentioned above. Although passive reserved resources are utilized for best-effort traffic as long as the specific MN does use them and also are reserved on a soft state basic, this is seen as a drawback since these resources cannot be used to serve other users operating services different than best-effort. This problem clearly arises in dense networks, especially when serving large amount of MNs. In such situations, significant amount of bandwidth will be reserved for best-effort.

WLRP requires that MNs periodically sends Mob- and App-Profiles. This implies that MNs depend on layer 2 triggers to generate these profiles, this makes WLRP technology-specific. The periodic transmission of Mob- and App-Profiles implies considerable signaling over wireless links as well.

Although the hard state of active reservations can be seen as an advantage, such a state may result in some problems. For instance, in case the MN is crashed or the radio link with the MN has been broken for any reason, the network will not be able to release the resources until the resources are not being used for a certain time. Clearly, this results in a waste of resources.

For the employment of WLRP, one notices that WLRP does not introduce new nodes to the network and requires updating only MNs and BSs. Considering security issues, WLRP does not address them in its design.

4.3.2 Mobile Extensions to RSVP

The proposal presented in [AAg97] intends to extend RSVP signaling to support mobility. For the benefits of this purpose, three classes of reservations are defined, namely committed, quiescent and transient reservations. Committed reservations refer to the traditional form of reservation and allocation of resources. This form is used by the standard RSVP. Resources are reserved in quiescent reservations, however, not allocated (the MN is not within the subnet yet). These resources can be allocated temporarily for other MNs. However, they should be preempted when they become activated. Transient reservations are those quiescent ones currently allocated temporarily for other MNs.

Network topology and basic principles: the basic idea of the proposal is the construction of a dynamic multicast tree that is centered on the current subnet and additionally includes neighbor subnets, see Figure 4.4, which shows the network topology and basic principles of the proposal. The root of the multicast tree is called a fulcrum node. There exist a committed reservation between the fulcrum node and the subnet hosting the MN. Other branches of the multicast tree accommodate either quiescent or transient reservations. When the MN moves to a neighbor subnet, the quiescent reservations are activated, thus, QoS guarantee quickly retrieved. The multicast tree is also adapted, and so on.

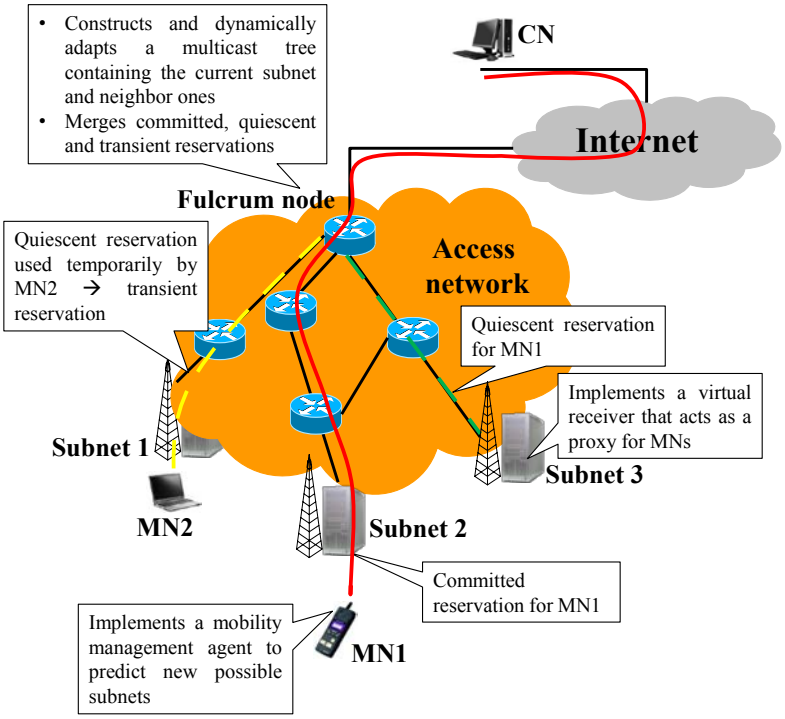


Figure 4.4: Network topology and basic principles employing the mobile extensions to RSVP proposal

Operation overview: the operation of the proposal requires many extensions to the standard RSVP. First, the RSVP engine in the MN should be extended to include a mobility management agent. The task of this agent is to predict future locations of the MN, thus, predicting new possible subnets. This agent also signals to the predicted subnets to trigger joining the multicast tree and establishing quiescent reservations. Second, the RSVP engine in the subnets is extended in two intentions. The first is a virtual receiver, while the second are extra control messages. Virtual receivers act as proxies for MNs. The extra control messages include PATHCom, RESVCom, PATHQui, RESVQui, PATHTra and RESVTra messages. These messages are used to establish and maintain the three types of reservations mentioned.

Let us now briefly discuss how this proposal operates. When the MN is powered on, it establishes a committed reservation with the CN. This is done by exchanging PATHCom and RESVCom messages with the CN. Following that, the mobility management agent implemented in the MN predicts the possible new candidate subnets and periodically signals them to prompt joining the multicast tree and the establishment as well as maintenance of quiescent reservations. The quiescent reservations are achieved by exchanging PATHQui and RESVQui messages with the fulcrum node. Note that no quiescent reservations achieved beyond the fulcrum node. When the MN moves into the range of a neighbor subnet, the quiescent reservation achieved previously on its behalf is activated. Note that the resources not required anymore, will not be refreshed, thus, released after timeout.

Performance evaluation: it is obvious that the proposal enables MNs to move freely inside the network while maintaining QoS guarantees. This reduces the number of packets getting lost due to handoffs and those encountering end-to-end delays more than the play-out time when comparing to the standard RSVP.

Pros and cons: because the proposal focuses on extending RSVP to improve the performance for mobiles, neither the tunneling nor the triangular routing problems are addressed. Concerning the double resource reservation problem, one notes that the introduction of a fulcrum node

to merge committed and quiescent reservations provides a non optimal solution for the problem. This is because, resources may be reserved doubled on the paths between the fulcrum node and the old as well as new subnet.

A main advantage of the proposed technique is that the performance improvements are obtained without reserving resources passively. However, MNs should be capable of tracking their movements. In other words, the technique is technology-dependent (it relies on layer 2 triggers).

Lets us now discuss the employability of the proposal. The discussion introduced above shows that the proposed technique does not restrict network topology. However, a hierarchical topology would be better. A fulcrum node must be introduced to the network. Moreover, considerable updates are required since MNs as well as all domain subnets and RSVP-enabled nodes must be extended to implement the extensions discussed in this section. Sure, this negatively affects the employability of the proposal. In terms of security, one notes that the proposal does not address issues related to security.

4.4 Loose-Coupled Approaches

As mentioned in section 4.2, loose-coupled solutions separate between the protocols managing mobility and those handling QoS from the implementation point of view. However, the operation of one of them influences the operation of the other. So as to provide an insight into loose-coupled solutions, well-known approaches will be reviewed in this section.

4.4.1 Mobile RSVP (MRSVP)

MRSVP [TBB01] extends RSVP to support mobility. It distinguishes between two types of resource reservations, namely active and passive reservations. The resources reserved actively are those the MN currently uses, while the resources reserved passively are the resources reserved for MNs expected to come into the subnet in the near future.

Network topology and basic principles: Figure 4.5 shows the network topology as well as basic principles of MRSVP. The agent serving the MN is called a local proxy agent, while the agents to which the MN may move from the local proxy agent are referred to as remote proxy agents. The figure also shows that the MN is assigned to an anchor point in the network (typically a crossover node). The anchor point manages the active and passive reservations for the MN. Note that as active and passive reservations share the same link inside the access domain, they will be merged with each other to avoid doubled reservations. The required FLOWSPECs in the resulting combined reservation is determined according to the IntServ model, see [Wro97]. Note also that MRSVP does not constrain the topology deployed. However, a hierarchical topology would be better.

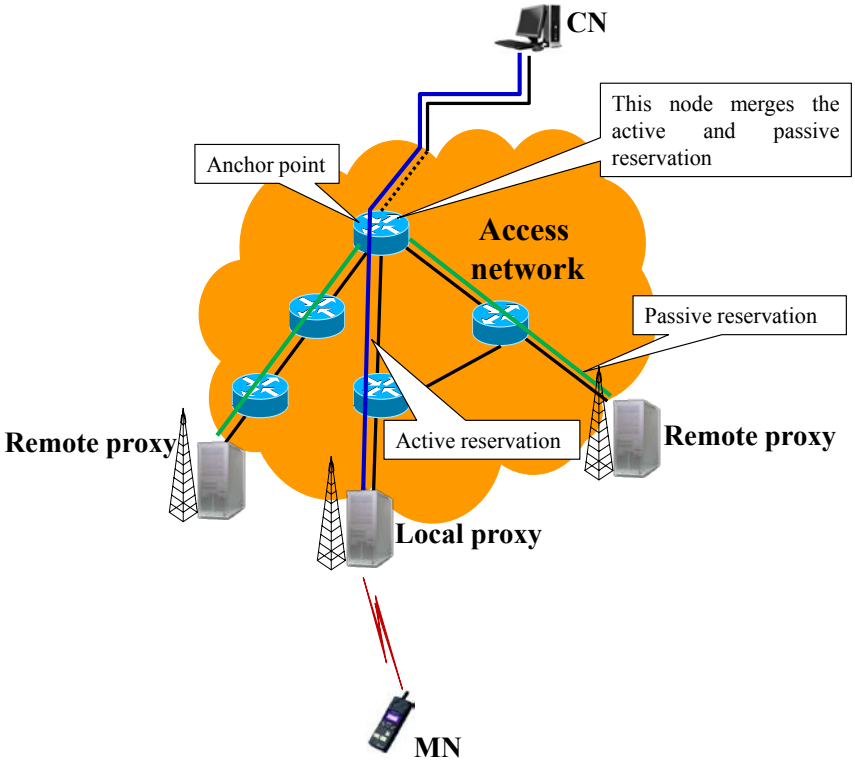


Figure 4.5: Network topology and basic principles employing MRSVP

Operation overview: first, the MN determines, by means of a proxy discovery protocol [TBB01], the IP addresses of the remote proxies to which the MN will probably move. The operation of this protocol implies the exchange of a remote Agent Solicitation (remote Agnt_Sol) and a remote Agent Advertisement (remote Agnt_Adv) messages between the MN and each remote proxy agent via the local one, see Figure 4.6. Following that, the MN registers itself with the HA and eventually the CN[1] according to MIP. Note that the example provided in the figure below considers MIPv4.

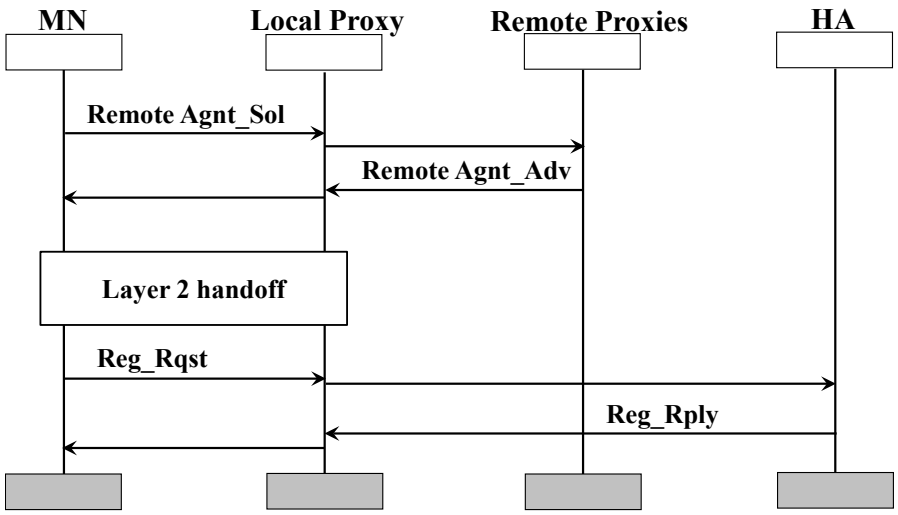


Figure 4.6: Determination of the remote proxies' IP addresses and registration with HA

[1] In case the MN operates MIPv4, it registers with the HA only. However, when MIPv6 is employed, the MN must register with the CN, as well.

The resource reservation procedure for downlink sessions employing MRSVP is illustrated in Figure 4.7. As shown, the MN sends periodic Mobility SPECification (MSPEC) messages containing the addresses of remote proxy agents to its anchor node. In addition, the MN sends periodic SPECification (SPEC) messages to each remote proxy to notify it of the required QoS metrics (FLOWSPEC, ADSPEC and flow identification). It is worth noting that the anchor node may be the CN or the HA. Notice that if the MN uses route optimization, it makes sense to select the CN as an anchor point, while the HA should function as an anchor point for the traffic sent on the triangular route.

Let us assume that the HA is the anchor point as the figure shows. The CN sends a PATH message to the MN. This PATH message passes the HA, which in turn sends a PATH message to the local proxy as well as each remote proxy. The local proxy responds with an active RESV message, while each remote proxy replies with a passive RESV message. Notice that passive RESV messages have, in principle, the same structure as active RESV messages. However, they are used to reserve resources in advance for the MN. These resources are termed passive resources. Once the HA receives the active RESV message from the local proxy, it forwards it to the CN. Passive RESV messages will not be forwarded beyond the HA. As the MN moves into the range of one of the remote proxy agents, resources reserved passively for the MN will be switched to active. This switching is assumed to be carried out as the MN attempts to register itself employing MIP.

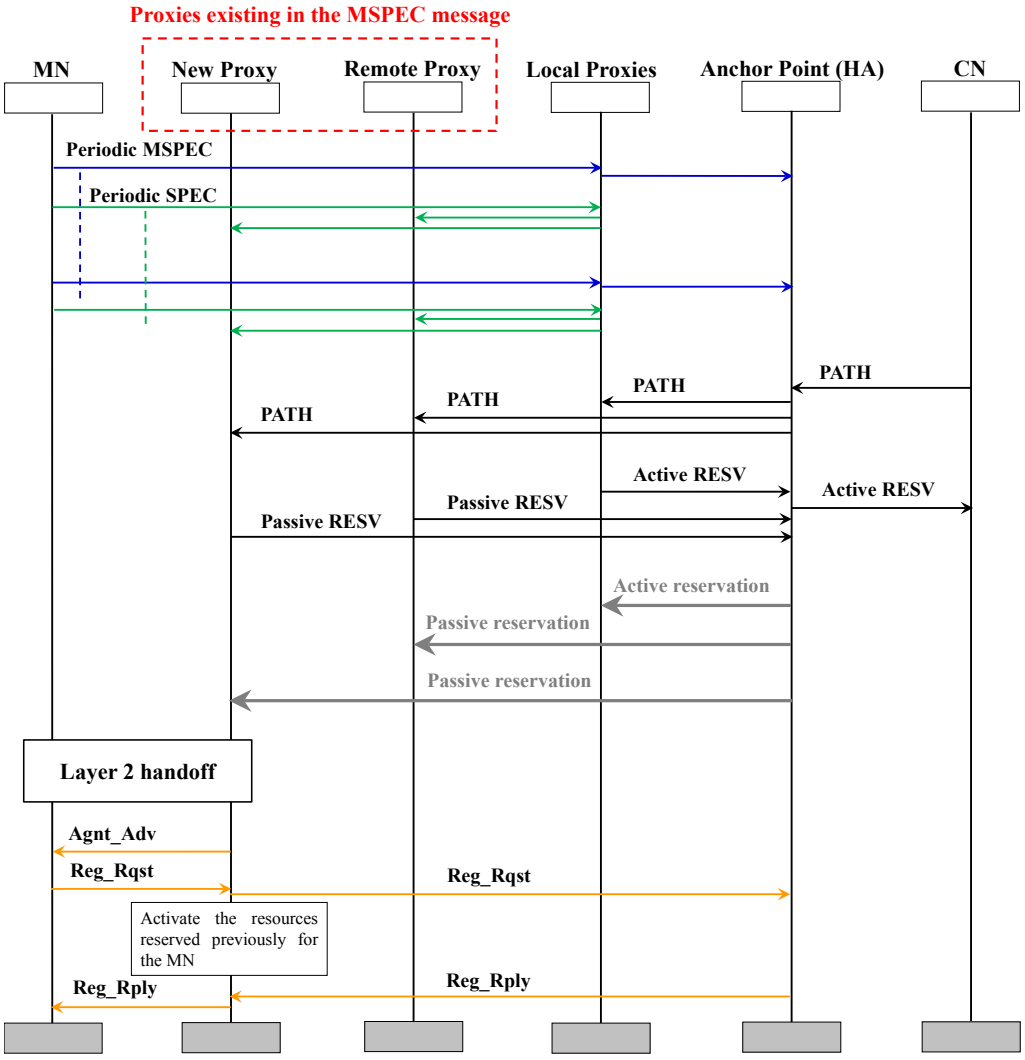


Figure 4.7: Operation overview of MRSVP (resource reservation for downlink sessions)

Resource reservation for uplink sessions is achieved in a similar way, see Figure 4.8. Notice that PATH messages are issued from the local as well as remote proxies towards the anchor point (the HA), which forwards only a PATH message to the CN. Once the HA receives the active RESV message from the CN, it transmits an active RESV message to the local proxy as well as a passive RESV to each remote proxy. It should be noticed that MRSVP in the case of triangular routing uses RSVP tunnels to enable resource reservation at the routers within the tunnel. Furthermore, to release the unwanted passive reservations, the MN sends **Terminate** messages to its proxy agents. The proxy agents send a RESV TEARDOWN or PATH TEARDOWN message to tear down the passive reservations unless some other users share these reservations.

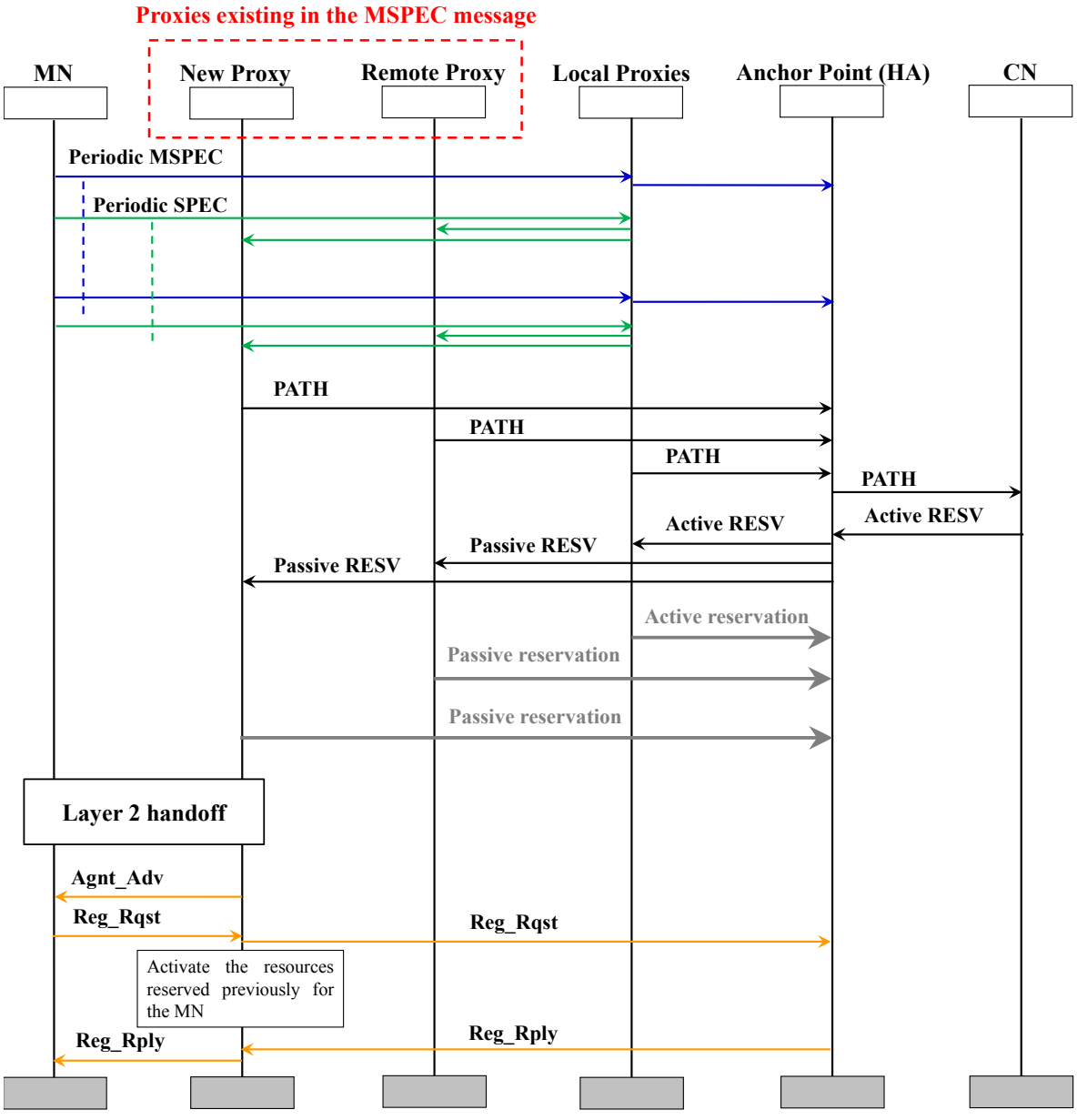


Figure 4.8: Operation overview of MRSVP (resource reservation for uplink sessions)

In the case of multicasting to a MN that operates as a receiver, the MN joins the multicast group and notifies all its remote proxies to join the multicast group, as well. In addition, the MN also sends a SPEC to each of its remote proxy agents. The sender sends a PATH message

to the multicast destination address. Once the MN receives an active PATH message, the MN responds with an active RESV message towards the sender. If any of the remote proxy agents receives a passive PATH message, it replies a passive RESV message. Notice that the multicast routing protocol employed determines the routes of active and passive reservations.

The authors define in [TBA99] three service classes to be provided by networks that employ MRSVP, namely mobility-independent guaranteed service class, mobility-independent predictive service class and mobility-dependent predictive service class. The mobility-independent guaranteed service class is appropriate for delay-intolerant applications. This service class guarantees that MNs obtain the QoS required as long as their movements are limited to the remote proxies present in the *MSPEC* messages and their traffic characterizations do not change during the currently active sessions. MNs admitted to mobility-independent predictive service class obtain the predictive QoS rather than the absolute guaranteed level as long as movements of MNs are restricted to the proxies existing in the *MSPEC* messages and MNs' traffic characterizations do not change during the currently active sessions. This service class is appropriate for delay-tolerant applications. The mobility-dependent predictive service class offers services similar to those of mobility-independent predictive service class. However, the MN occasionally fails to obtain the predictive QoS level and, thus, suffers from service degradation. The sessions of such MNs are even allowed to be dropped for the benefit of the flows of both other service classes. The mobility-dependent predictive service class is adequate, therefore, for applications that can tolerate data loss, sessions dropping, etc.

Performance evaluation: simulation results presented in [TBA99] show that employing MRSVP negatively affects network utilization since many resources will be passively reserved, thus, network utilization reduced. The decrease in network utilization will not be significant, however, if suitable multiplexing of mobility-independent and mobility-dependent flows is permitted. In addition, the flow dropping rate is reduced in such cases as compared to the case in which all flows in the network are mobility-dependent.

Pros and cons: one can conclude from the above discussion that MRSVP provides QoS guarantees while moving in the network. In this way, MNs, in general, do not suffer from service degradation due to movements. No statements are made, however, on how MNs' velocities affect these guarantees. Moreover, the accurate definition of remote proxies is challenging, especially at high speeds. This requires MNs to track their movements, which consumes their power. Furthermore, this implies dependency on layer 2 triggers, which makes MRSVP technology-specific and implies considerable signaling over wireless links.

MRSVP employs the standard RSVP to reserve resources. Thus, both tunneling and double resource reservation problem appear. Note that due to the usage of anchor points to merge passive and active reservations, the double resource reservation problem is minimized. The triangular routing problem arises, as well, since MIP is used as a mobility management protocol. With respect to passive reservation of resources, MRSVP reserves resources passively to improve the performance. Sure, this reduces the network utilization as mentioned above. Furthermore, this problem will be more critical in dense networks, especially when large amount of MNs are served. In addition, passive reservation of resources results in wasting resources.

Another drawback of MRSVP is its complexity in terms of implementation. Keep in mind also that MRSVP does not constrict network topology. However, it introduces anchor points to the network. In addition, updates should be accomplished on MNs, CNs, HA and all proxies in the domain. Considering security issues, MRSVP does not address them in its design.

4.4.2 Hierarchical Mobile RSVP (HMRSVP)

HMRSVP attempts to utilize the principles of micro mobility management to localize the resource reservation inside the access domain. For this purpose, HMRSVP integrates between RSVP and the MIPRR protocol.

Network topology and basic principles: Figure 4.9 shows the network topology and basic principles for intra- and inter-domain handoffs. Note that HMRSVP uses the same topology used by MIPRR protocol, see section 2.2.2. The RSVP session between the MN and the CN is split into two sessions by the GFA. Movements of the MN inside the domain affect only the reservations to the GFA, while the resources reserved between the GFA and the CN remain unaffected. Clearly, this results in a significant reduction of the resource reservation latency after handoffs. To guarantee that the reservation of resources will not take a long time when moving between different access domains, HMRSVP reserves resources passively for the MNs that are located on the domain boundaries and may move into another domain.

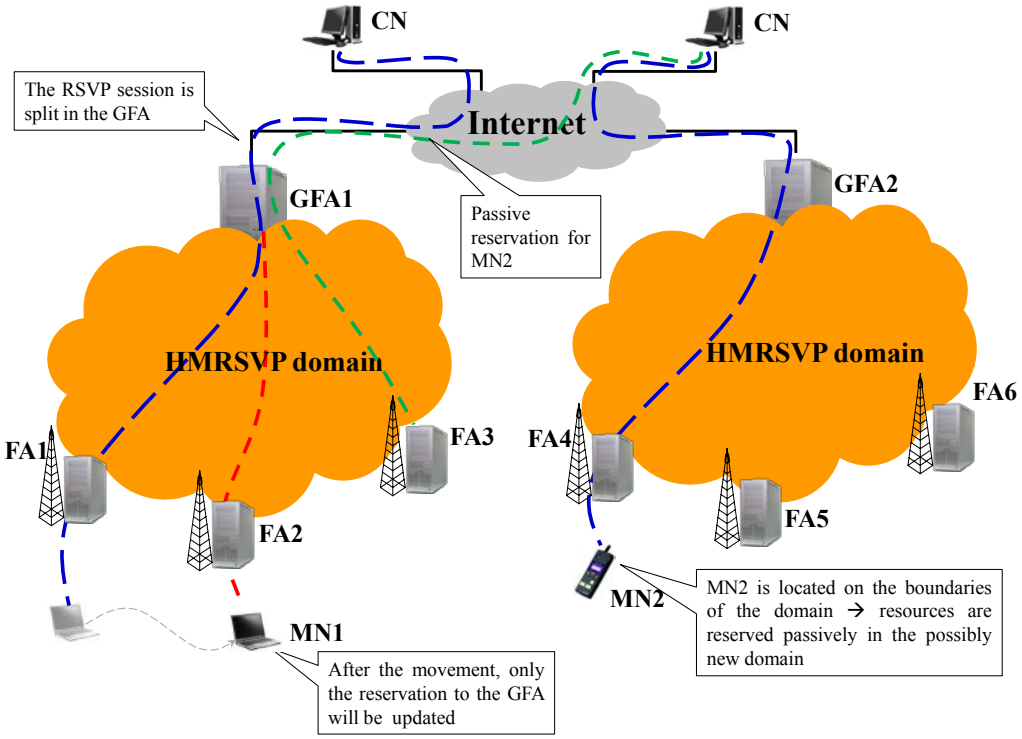


Figure 4.9: Network topology and basic principles employing HMRSVP

Operation overview: once the MN is switched on, it registers the address of the GFA with the HA as it's CoA (Global CoA (GCoA)) and registers the address of the local proxy[1] with the GFA as the current point of attachment (Local CoA (LCoA)). Following this, the MN transmits a Receiver Mobility SPECification (Receiver MSPEC) message to the CN to notify it of the current location of the MN[2], see Figure 4.10. Then, the RSVP session is established between the CN and the MN. Notice that the RSVP session consists, in principle, of two sessions as mentioned previously (one between the CN and the GFA and the other between the GFA and the MN). It should be mentioned that HMRSVP messages are tunneled using RSVP tunnels, when the MN is not located in its home network, to enable resources to be reserved at the routers within the tunnel.

[1] The local proxy corresponds to the current FA in the specification of MIPRR.

[2] The MN notifies the CN of the GCoA.

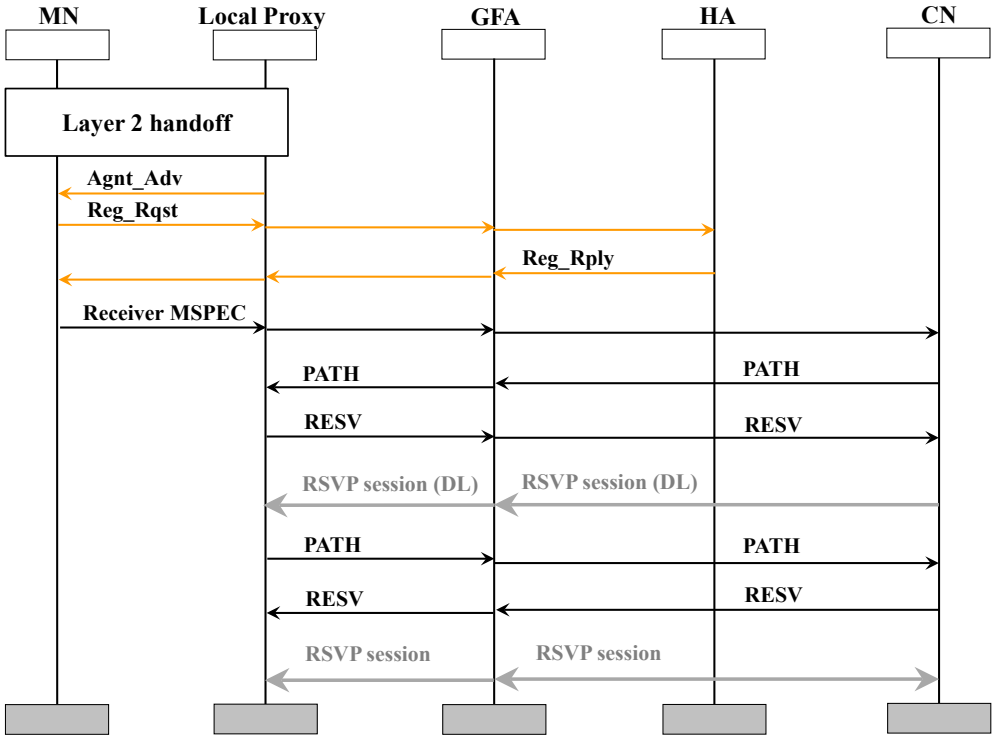


Figure 4.10: Initial registration and resource reservation employing HMRSVP

When the MN moves between proxies of the same domain, it must perform an intra-domain handoff. In this way, the MN registers itself first with the GFA by exchanging regional Reg_Rqst and regional Reg_Rply messages with the GFA. Once the GFA is informed of the new LCoA of the MN, it notifies the HMRSVP module, which exchanges active PATH and active RESV messages with the proxy serving the MN. The intra-domain handoff is depicted in Figure 4.11.

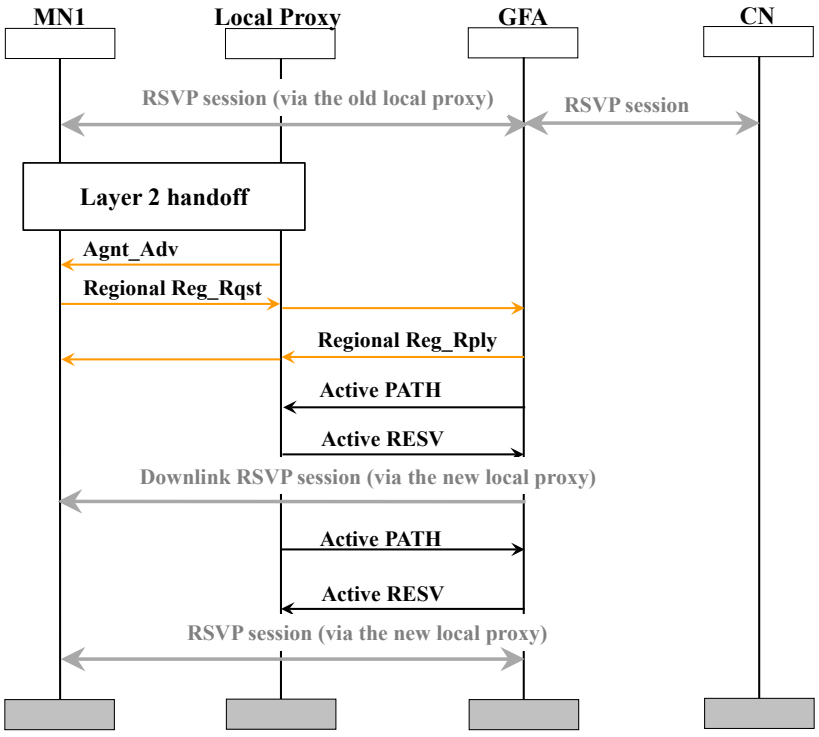


Figure 4.11: Intra-domain handoff employing HMRSVP

When the MN moves into an overlapping area between the boundaries of two different access domains, it must execute an inter-domain handoff. This is achieved as follows: the MN performs a home registration by sending a multiple simultaneous Reg_Rqst to the HA via the detected proxy belonging to the new domain. The HA, in turn, adds the newly acquired CoA to the CoAs list of the MN and replies with a Reg_Rply to the MN via the GFA and the local proxy in the new domain. The MN then sends a receiver SPEC message to the new proxy to inform it of the QoS parameters that the MN aims at having. In addition, the MN notifies the CN of the newly acquired CoA by transmitting a Receiver MSPEC message to it. Once the CN receives the Receiver MSPEC message, it proceeds with reserving resources passively in the new domain. For this purpose, the CN sends an end-to-end passive PATH message towards the new proxy via the GFA of the new domain. The new proxy, in turn, responds by sending an end-to-end passive RESV message. Notice that although the exchanged passive PATH and RESV messages are end-to-end messages, they result in two RSVP tunnels as mentioned previously, CN→GFA and GFA→local proxy. The resources reserved passively can be borrowed by other MNs as long as the MN is not located in the range of the new subnet. Clearly, this requires defining adequate policies to manage such resources borrowing. As the MN moves into the new access domain, the passively reserved resources will be switched to active. The inter-domain handoff is depicted in Figure 4.12.

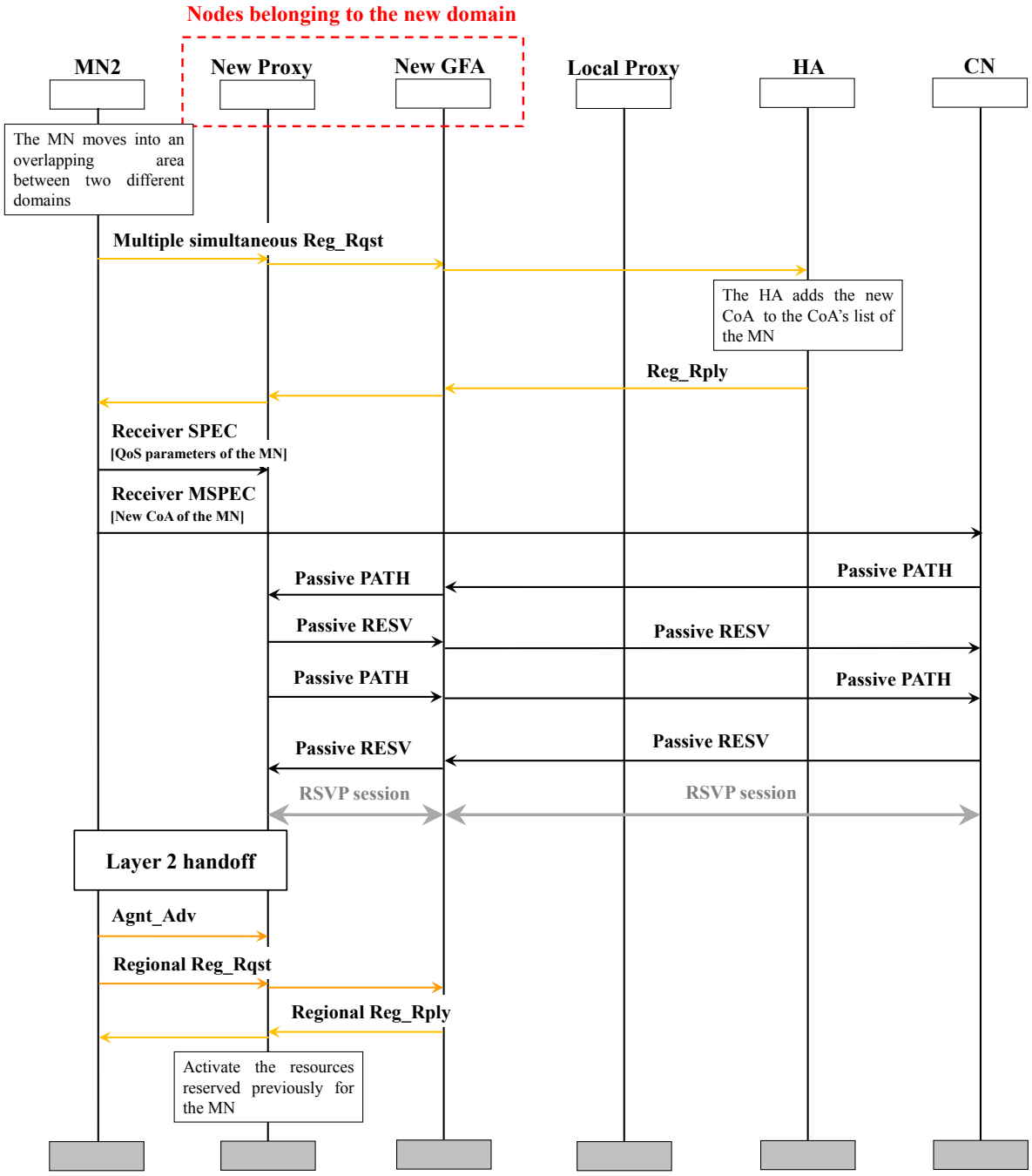


Figure 4.12: Inter-domain handoff employing HMRSVP

Performance evaluation: simulation results provided in [TLL03] show that employing HMRSVP and MRSVP enable MNs to have stable data rate while moving inside the network. Both outperform RSVP in this context since no stable data rate can be provided by RSVP. HMRSVP outperforms MRSVP, however, in terms of reservation blocking[1], forced termination[2] and session completion[3] probabilities. The main advantages of HMRSVP can be seen in

[1] Reservation blocking probability is the probability that an active MN fails in reserving resources for a new RSVP session for a new data stream.

[2] Forced termination probability is the probability that an active MN fails to reserve resources after the handoff.

[3] Session completion probability is the probability that the MN successfully completes its session after the handoff.

its ability to reserve resources very quickly and reduce the waste of resources as compared to MRSVP since passive reservations are restricted to inter-domain handoffs only.

Pros and cons: from the discussion above, one can see that HMRSVP localizes the signaling for handoffs as well as resource reservations inside the domain as long as movements of the MN are restricted to the proxies of this domain. This enables fast resource reservation to be achieved after handoffs. In addition, restricting the passive resource reservation to inter-domain handoffs increases the efficiency of resource use as compared to MRSVP. However, the MN still has to predict its possible new proxies for inter-domain handoffs, which is challenging, especially when moving at high speeds. Sure, the prediction of possible new proxies consumes the power of the MN and implies dependency on layer 2 triggers. This makes HMRSVP technology-specific.

Although HMRSVP reserves resources passively when the MN will probably move to a new domain, the success of this reservation strongly depends on the speed of the MN as well as the size of the overlapping area. A main drawback of HMRSVP is the single point of failure, since all traffic and signaling must pass the GFA controlling the domain.

HMRSVP addresses the tunneling problem inside the domains. However, in case MIP is used to support inter-domain mobility and triangular routing is used to forward data packets to the GFA that controls the domain serving the MN, tunneling as well as triangular routing problem exists since the standard RSVP is applied outside HMRSVP domains. Concerning the double reservation of resources, HMRSVP does not address it. Thus, this problem may appear when employing HMRSVP.

Concerning the employability of HMRSVP, one notices that this protocol constrains network topology since it requires a hierarchical network architecture. Furthermore, it introduces new nodes to the network since MIPRR protocol is used as a mobility management protocol inside the domain. As known, this protocol requires GFAs to be present. In addition, MNs, CNs, HA, GFA and all FAs of the domain should be updated to enable operating HMRSVP. Of course, the update of CNs negatively affects the employability of HMRSVP. Concerning security, HMRSVP does not address it.

4.4.3 Simple QoS

Network topology and basic principles: Simple QoS loosely couples RSVP and MIPv4. The network topology deployed is the same as that of MIPv4. In addition to the end-to-end RSVP session established on the triangular route between the CN and the MN (CN→HA→MN), there is another RSVP session established for the tunnel present between the HA and the FA currently serving the MN. The main goal of this RSVP session is to handle packets tunneled by MIP from the HA to the MN's new location and not handled by the end-to-end RSVP session. Figure 4.13 presents the network topology and basic principles of Simple QoS protocol.

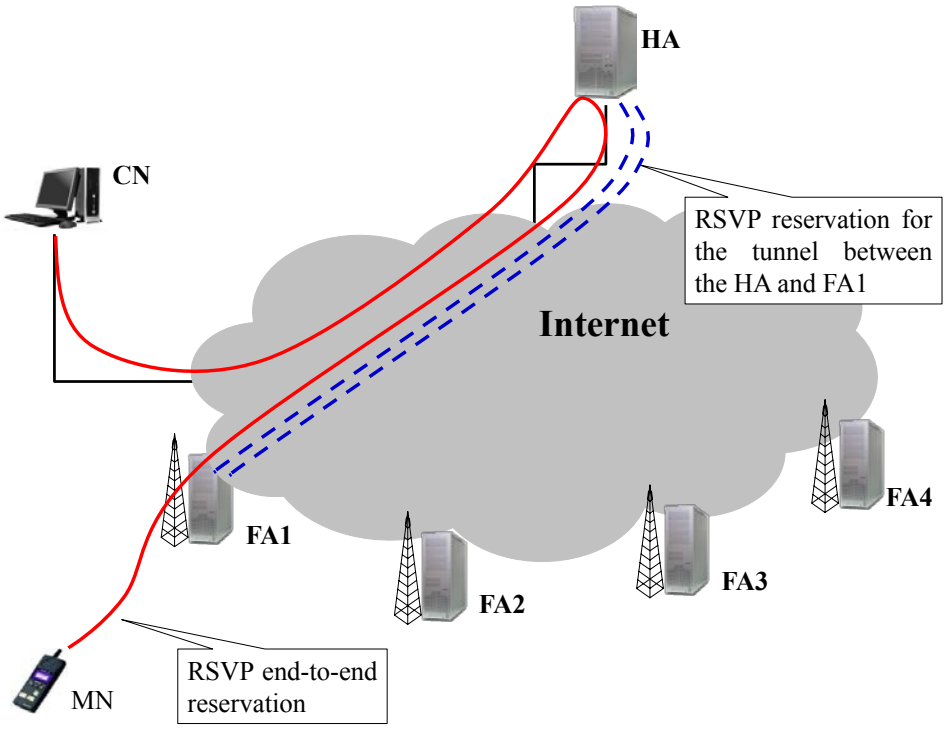


Figure 4.13: Network topology and basic principles employing the Simple QoS protocol

Operation overview: Simple QoS requires modification to MIPv4 and RSVP engines. The modification to MIP includes adding a **Q** bit to the Agnt_Adv message to advertise the support of Simple QoS protocol. A similar **Q** bit is added to the Reg_Rqst message, as well, to signal that the MN supports Simple QoS. One distinguishes between two cases of operation based on whether the MN is a receiver or a sender.

Let us first consider the case where the MN operates as a receiver and assume that the MN is located in its home network. Once a CN would like to communicate with a MN, the CN establishes an end-to-end RSVP session. As the MN moves away, it re-registers with the HA employing MIPv4. When the HA is notified of the MN's new location, it establishes a RSVP session for the tunnel between itself and the FA currently serving the MN in case such a session does not exist previously. This RSVP session will be referred to as RSVP tunnel in the following. After that data packets forwarded to the MN, they will be intercepted and tunneled to the current FA through the reserved RSVP tunnel. Moreover, once the HA receives a PATH message from a CN, it tunnels the PATH message towards the current FA. The current FA, in turn, de-tunnels and forwards the PATH to the MN. The RESV message the MN sends is handled in a similar way and tunneled through a reverse tunnel from the current FA to the HA. The authors state in [TSZ99] that the tunnel between the HA and a FA is used for all MNs registered with the HA and currently located in the range of the FA. Thus, the FLOWSPEC of this tunnel is the sum[1] of all FLOWSPECs of those MNs. When new MNs move into the range of the FA, only the FLOWSPEC of the tunnel will be adjusted.

Let us now consider the second case where the MN is a sender and assume that the MN is located in the range of a FA. In this case, the MN sends the PATH message towards the CN. Here, the PATH message may not bypass the HA, which complicates the reservation process. Therefore, a reverse tunnel should be established, in case it does not already exist, between the current FA and the HA. The reservation process is done in a similar way as when the MN

[1] The sum of FLOWSPEC can be done in various ways, e.g. add peak rates fashion, use equivalent bandwidth metric, etc.

acts as a receiver. In [TSZ99], an enhancement for the uplink reservation is proposed that does not require a reverse tunnel. Instead, the PATH message climbs up until it reaches a crossover node, which is defined as the first node located on both the old and new path between the MN and the CN. This crossover node returns a RESV message directly towards the MN. Figure 4.14 and Figure 4.15 show the two operation cases discussed above.

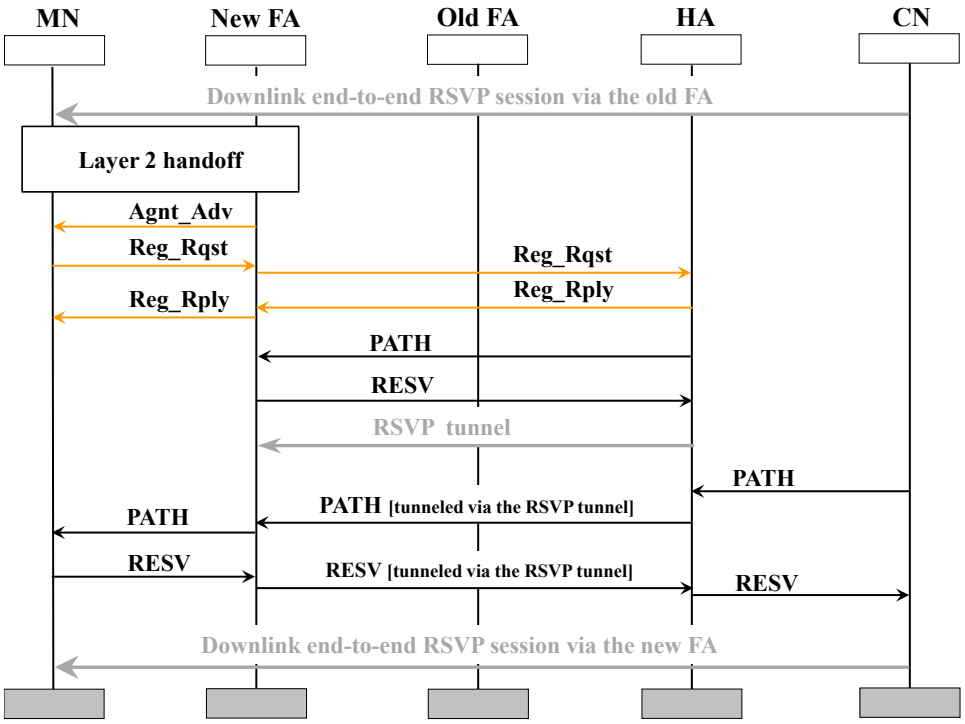


Figure 4.14: Operation of the Simple QoS protocol (the MN operates as a receiver)

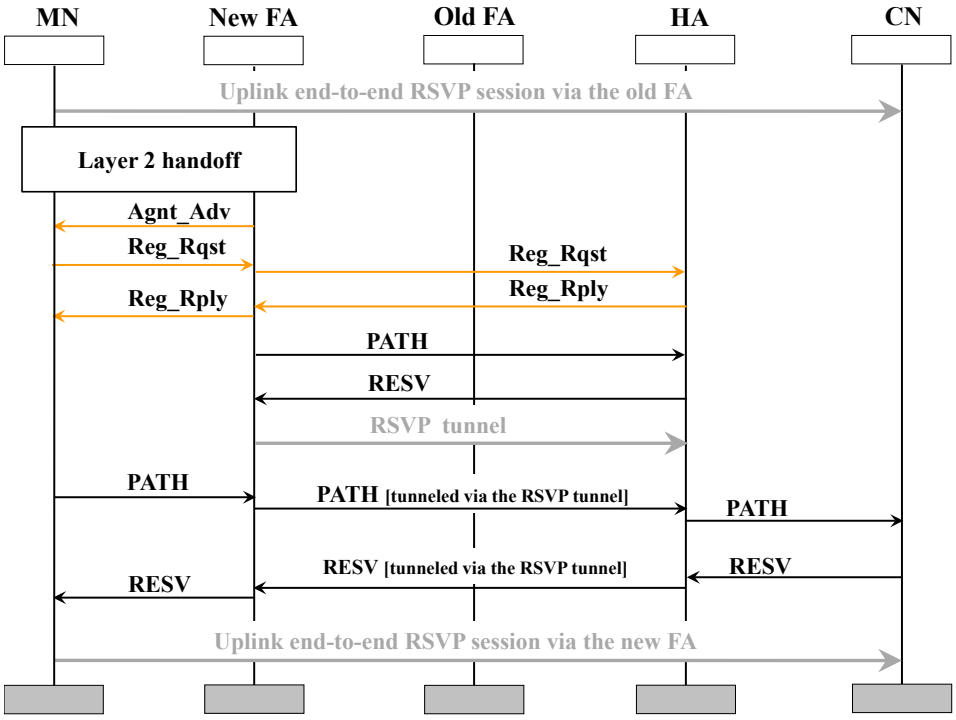


Figure 4.15: Operation of the Simple QoS protocol (the MN operates as a sender)

Performance evaluation: simulation results show that Simple QoS reduces the QoS disruption after handoffs as compared to the plain RSVP. It reduces the number of packets getting lost due to handoffs and those encountering end-to-end delays more than the play-out time.

Pros and cons: Simple QoS does not suffer from the tunneling problem since an RSVP tunnel is established between the HA and the FA currently serving the MN. This tunnel is also used for all MNs served by the specific FA, which implies efficiency in resource use. However, a considerable overhead due to the mentioned RSVP tunnel is involved. Notice that the establishment of this tunnel does not result in the avoidance of the double reservation of resources problem, since this tunnel does not consider a possible existence of crossover nodes on the path between the HA and the current FA. The triangular routing is also a drawback of this solution, since MIPv4 is applied as a mobility management protocol.

The advantages of Simple QoS are obtained without reserving resources passively or forcing the MN to track its movements, i.e. this protocol does not rely on layer 2 triggers. Concerning the employability of Simple QoS, one notices that this protocol neither constrain network topology nor introduce new nodes to the network. However, updates to MNs, the HA and all FAs of the domain are necessary.

4.4.4 Localized RSVP (LRSVP)

Network topology and basic principles: the main objective of LRSVP is to localize the RSVP signaling in an access network by introducing a new proxy (called LRSVP proxy) to split the RSVP session into two sessions. The first session is between the CN and LRSVP proxy, while the second is between this proxy and the MN. The MN's movements inside the access domain only result in updating the RSVP session to the LRSVP proxy, while the session from this proxy to the CN remains unchanged, see Figure 4.16, which shows the network topology as well as basic principles of LRSVP.

Operation overview: two new control messages are introduced to RSVP, namely PATH Request and PATH Request Tear messages. The MN sends the first message to request a PATH message from the LRSVP proxy to accelerate the reservation reestablishment for downstreams, while the second message is used to tear down a downlink reservation (i.e. request the LRSVP proxy to send a PATH TEARDOWN message). In addition to the new messages introduced, LRSVP proposes a new flag, referred to as LI flag, to be added to all RSVP messages. Setting this flag means that control messages should be processed locally and not be sent beyond the LRSVP proxy.

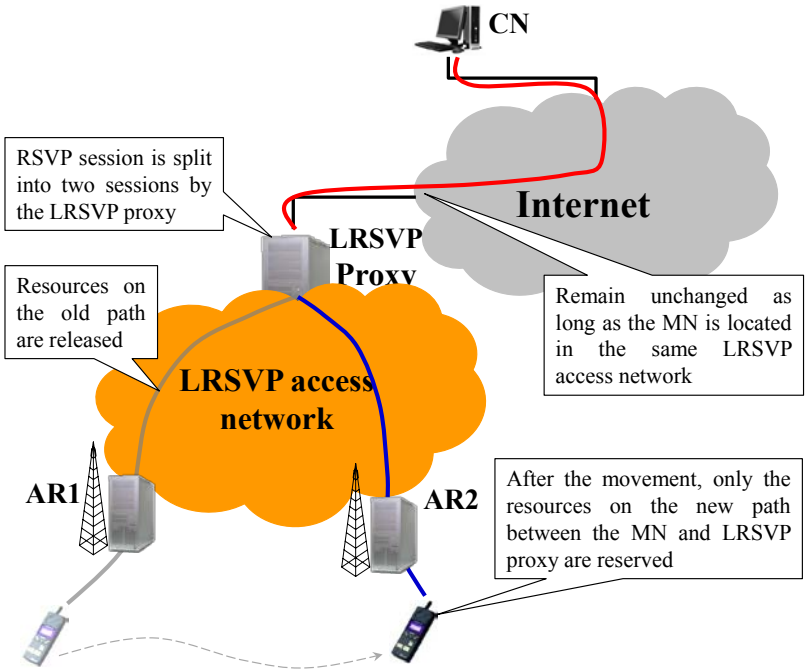


Figure 4.16: Network topology and basic principles employing LRSVP

After the MN moves into the range of a new subnet and accomplishes the handoff, it must reserve the resources either on uplink, downlink or on both directions. This depends on the session that the MN currently has . The uplink reservation is straightforward and occurs according to the standard RSVP. The difference is that the control messages involved contain the LI flag set. In detail, the MN sends a PATH message towards the CN. The message will be intercepted by the LRSVP proxy, since the LI flag is set. Subsequently, the LRSVP replies with a RESV message towards the MN. Downstream reservation is achieved in a slightly different manner, since the MN first sends a PATH Request message to force the LRSVP proxy to send a PATH message to the MN. Notice that the PATH Request message includes the IP address of the CN as the destination address. However, the message will not be forwarded beyond the LRSVP proxy due to the existence of the LI bit in this message. Once the MN receives the PATH message, it responds with a RESV message. Figure 4.17 illustrates the reservation process, where the MN reserves an uplink as well as a downlink RSVP session after the handoff. The messages colored with red are used to reserve resources on downlink, while the messages colored with blue are used for the uplink session.

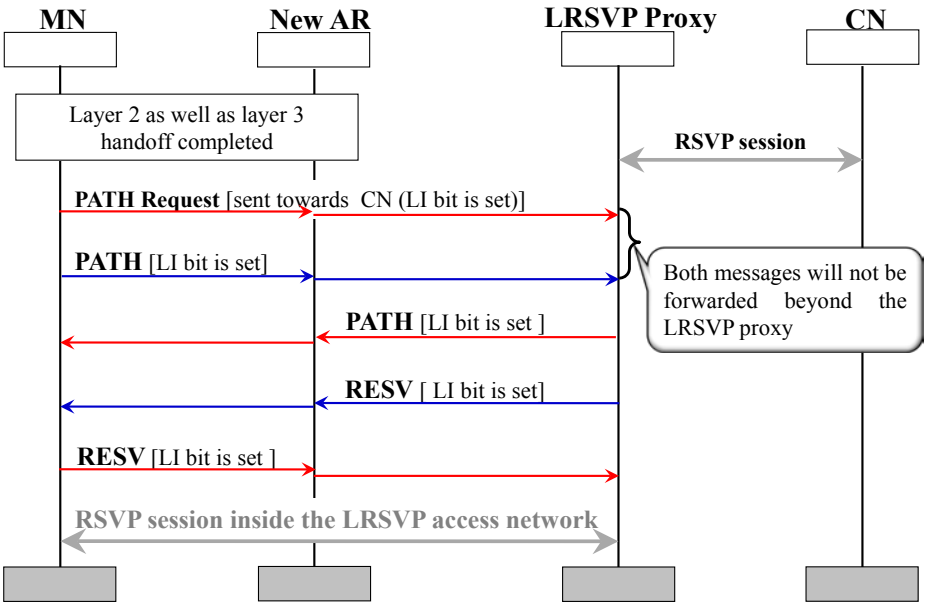


Figure 4.17: Resource reservation after the handoff employing LRSVP (the MN reserves resources for the downlink as well as uplink)

The authors in [MRa03] propose an enhancement to LRSVP by conveying the functions of LRSVP proxy in a crossover node, which is defined as the node shared by the old and new path towards the MN.

Performance evaluation: implementation results presented in [Man03] show that LRSVP is capable of quickly repairing the reservation after handoffs. The end-to-end delay by GSM-like flows could even be reduced in the applied testbed from 55 to 20 msec when packets classification and scheduling are used in ARs in addition to resource reservation.

Pros and cons: LRSVP localizes the reservation of resources and enables a quick reservation after handoffs. Clearly, this reduces the QoS degradation resulting from handoffs. However, the support of LRSVP requires, in addition to introducing a new node to the network (LRSVP proxy), updating MNs and all RSVP-enabled nodes inside the domain, which complicates the applicability of this solution. Moreover, how MNs can determine the address of LRSVP proxy is not well investigated. Note also that LRSVP constrains network topology, since a hierarchical topology is required.

Neither the tunneling nor the triangular routing problem is addressed by LRSVP, since these problems relate to the mobility management protocol applied. Moreover, the double reservation of resources problem is not addressed, as well. Keep in mind that repairing the reservation after handoffs between the LRSVP proxy and the MN does not consider a possible existence of crossover nodes on the path between the LRSVP proxy and the MN. The enhancement proposed in [MRa03] avoids, however, the double reservation of resources problem.

A main advantage of LRSVP is that it achieves the mentioned performance improvements while avoiding passive reservation of resources and tracking MNs movements. In other words, LRSVP does not rely on layer 2 triggers. Security issues are, however, a problem, since the LRSVP proxy should be able to authenticate and authorize all control messages exchanged with MNs, which is challenging.

4.4.5 Multicast-based Mobility Support Employing RSVP

Network topology and basic principles: the main idea of this [CHu00] is to utilize multicast-based mobility management principles to support mobility as well as QoS. More concrete, a multicast tree centered on the current location of the MN and additionally including subnets located in the surroundings is established. MNs' movements are then modeled as transitions of multicast group memberships, i.e. joining and leaving the multicast group. The proposed technique assumes three kinds of reservations, namely conventional, predictive and temporary reservation. Conventional reservation denotes the resources reserved on the path from the CN to the MN in the current location. These resources are active and allocated for the MN. On the contrary to the conventional reservation, the predictive reservation expresses the resources reserved on the paths from the CN to neighbor subnets to enable smooth handoffs without QoS degradations. Clearly, resources reserved predictively are inactive. Temporarily reserved resources are those reserved predictively for MNs not currently located in the subnet and used temporarily by other MNs. Once an owner of resources reserved inactively moves into the network, the temporary use of these resources must be aborted.

Figure 4.18 presents the network topology of the proposed technique along with the basic principles described above. A mobility proxy is a proxy supporting the described technique and serving a subnet. It can be compared with a FA/AR using MIP. Notice that the network topology includes a merging point, which is the point that merges between the conventional and predictive reservations for the same flow to avoid the double reservation problem.

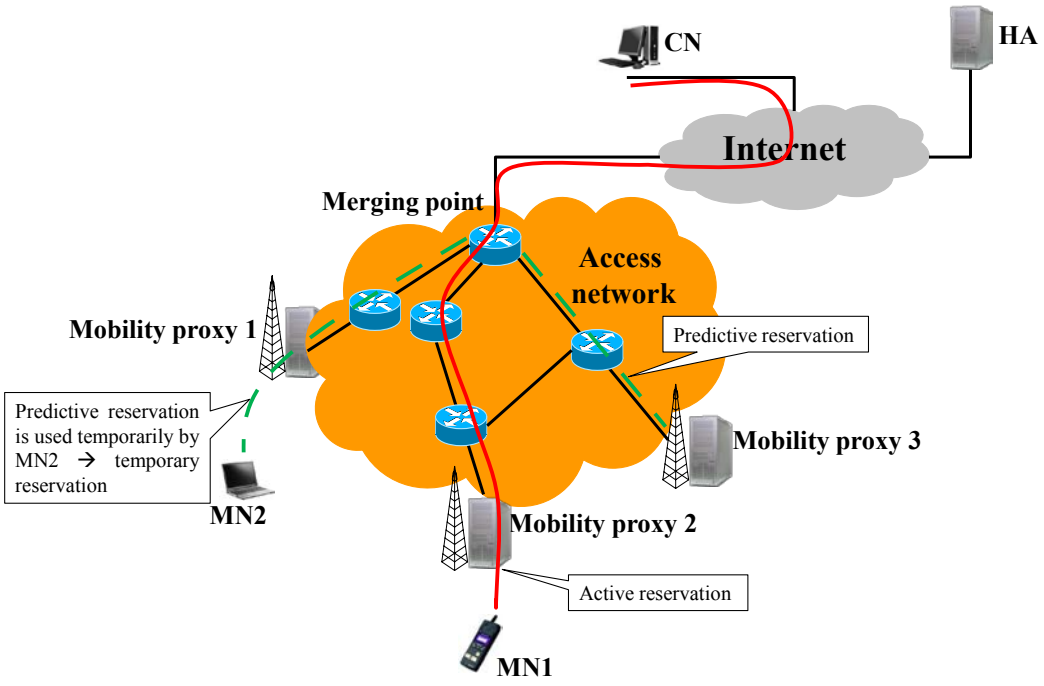


Figure 4.18: RSVP multicast-based mobility support (network topology and basic principles)

Operation overview: the handoff procedure of the proposal is pretty simple and can be summarized as follows. As the MN is switched on or moves into the range of a mobility proxy for the first time, resources are reserved on the path between the CN and the MN (conventional reservation). In addition, the MN sends a SessionSpec message including the multicast address of the MN's flow to the neighbor mobility proxies. This forces these proxies to join the multicast tree and reserve resources in advance (predictive reservation). As the MN moves into the range of a neighbor mobility proxy, the predictive reservation is switched into con-

ventional and so on. It is worth noting that when the old multicast tree branch is pruned, the MN's predictive reservation reserved there is removed.

Performance evaluation: simulation results provided in [CHu00] show that the service degradation as well as packet end-to-end delay due to handoffs when employing the proposed RSVP multicast-based mobility support is substantially lower than when employing Simple QoS protocol. Moreover, packet jitter during handoffs is minimized due to the fast forwarding of data after handoffs (using the multicast tree). However, the proposed technique requires considerable overhead due to the predictive reservation.

Pros and cons: besides the performance improvements advantages obtained from the proposal, there are some disadvantages, as well. A main disadvantage is that the MN must predict the neighbor mobility proxies and notify them to join the multicast tree. This requires MNs to be able to track their movements (i.e. relaying on layer 2 triggers) and consumes, as known, their power. Failing in the prediction of neighbor mobility proxies results in significant QoS degradation.

The tunneling and triangular routing problems appear based on the data forwarding path used. In other words, in case MIP triangular routing is used, both problems appear. Otherwise, the mentioned problems do not exist. The double reservation of resources problem is not optimally solved. This is due to the introduction of merging/anchor points to merge conventional and predictive reservations. However, resources may be double reserved after handoffs due to the usage of RSVP, which, as known, suffers from this problem.

The proposal achieves the performance improvements discussed based on predictive resource reservation, which, of course, negatively affects the efficiency of resources use. Concerning the employability of the proposal, one notices that no restrictions are made on network topology. However, mobility proxies and merging/anchor points should be introduced to the network. Moreover, MNs should be updated to operate the proposal. Security issues are not addressed in the context of the proposed technique.

4.4.6 Seamless NSIS-based QoS Guarantees with Advance Resource Reservation

The approach proposed in [LKL08] is based on the NSIS framework and aims at supporting fast resource reservation after handoffs.

Network topology and basic principles: the basic idea lies in enhancing the QoS NSLP protocol with in-advance resource reservation capabilities. More concrete, three modules are introduced to the QoS NSLP protocol, namely a crossover node discovery, advance resource reservation and a localized state update module. The crossover node discovery module is responsible for the in-advance detection of the crossover node, which is defined as the node on the old path with a route to the new subnet that will host the MN. The detection depends on a mobility prediction based on layer 2 information. The advance resource reservation module is used to achieve in-advance reservation of resources between the detected crossover node and the new location of MN. The localized state update module takes care of the states that should be created on the QNEs residing on the new path. Figure 4.19 provides the enhanced architecture of NSIS, while Figure 4.20 presents the network architecture and basic principles of the approach.

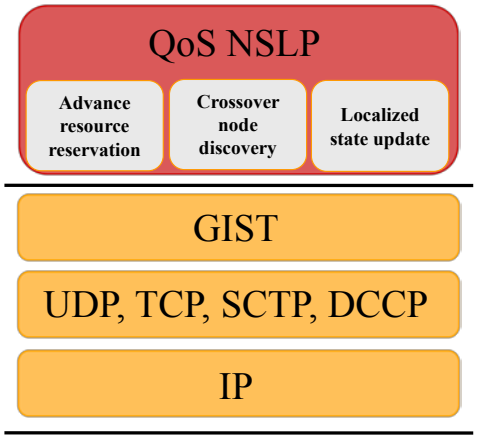


Figure 4.19: Enhanced architecture of NSIS

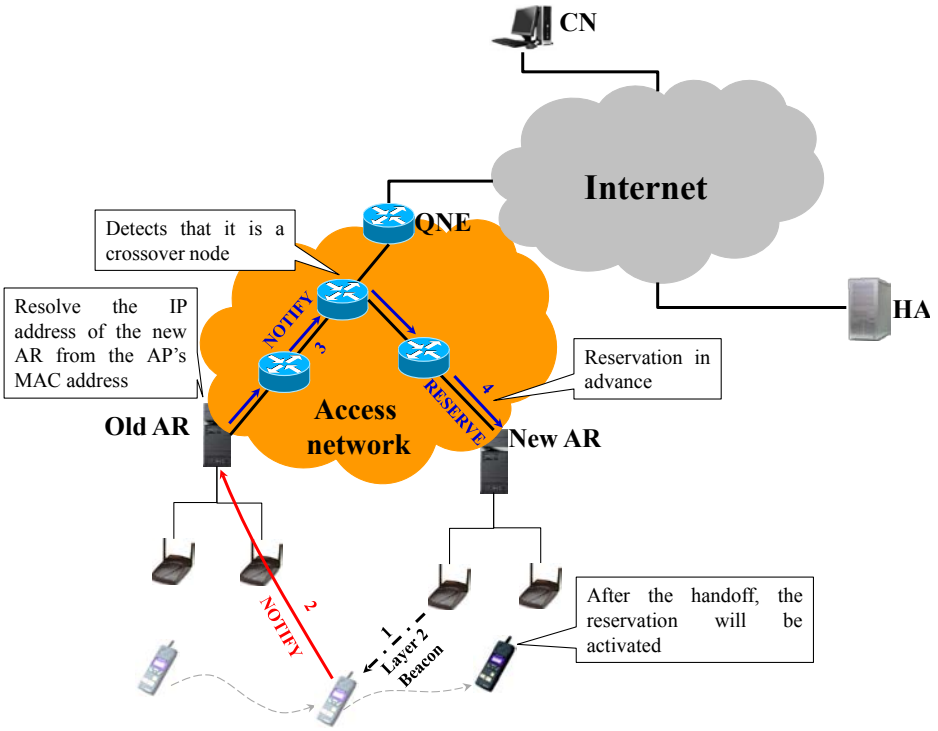


Figure 4.20: Network topology and basic principles of the proposal

Operation overview: as mentioned above, the detection of the crossover node depends on a mobility prediction based on layer 2 information. To enable such prediction, the authors of this approach rely on principles similar to those presented in [FPC05], [Mal07], [SZC04] and [TYC05]. They assume that wireless cells overlap and the MN is capable of detecting beacons from candidate APs and selecting the AP with the strongest signal. When the MN detects the new AP, it transmits a QoS NSLP NOTIFY message containing the MAC address of the detected AP and a HO_INIT flag to the old AR, see Figure 4.21. The HO_INIT flag is set to indicate that a handoff will occur in the near future. Once the old AR receives the NOTIFY message, it resolves the IP address of the new AR using a neighbor AR mapping table that contains the IP addresses of neighbor ARs and MAC addresses of APs served by them. Following this, the old AR replaces the destination address of the Message Routing Information (MRI) object present in the NOTIFY message with the IP address of the new AR. The updated NOTIFY message will be sent on the upstream path. Each QNE on this path checks if it

has a route to the new AR. If this is the case, this QNE will be selected as a crossover node. Afterwards, the crossover node responds by sending a NOTIFY message towards the MN. This message contains a flag, referred to as CRN_DCVD flag, to indicate that the crossover node has been detected. Simultaneously, the crossover node exchanges a stateless RESERVE and stateless RESPONSE with the new AR. Both messages are used to prepare reservation states on the path between the crossover node and the new AR. Notice that only reservation states are recorded and no resources are reserved at this moment. The resources will be used for other MNs until the MN moves into the range of the new AR. After the MN hands off to the new AR and accomplishes the layer 3 handoff employing MIP, it transmits a NOTIFY message including the Handoff_Done flag to the crossover router. The flag Handoff_Done indicates that the handoff has been accomplished. The NOTIFY message causes each QNE on the path towards the crossover node to activate the advance resource reservation module. Once the crossover node receives the NOTIFY message, it issues a state update message towards the CN to update the session. Simultaneously, it sends a RESERVE message with the Teardown flag set to the old AR to release the resources reserved.

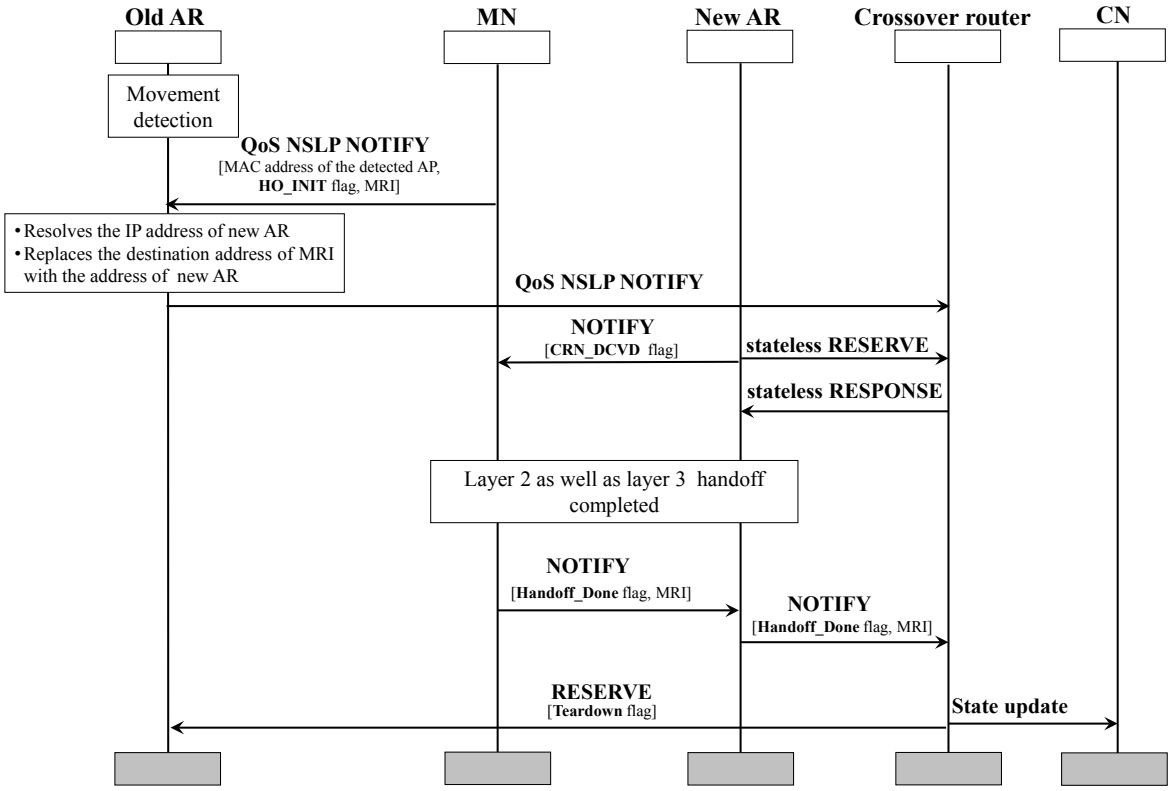


Figure 4.21: Crossover node discovery and handoff procedure employing the proposed approach

Performance evaluation: the authors have evaluated the proposed approach compared to the standard NSIS. Both schemes have been implemented in an experimental testbed and evaluated with respect to the average data transmission rate, average time required to re-establish sessions after handoffs and Peak Signal-to-Noise Ratio (PSNR). The average data transmission rate as well as the average session's re-establishment time after handoffs was measured employing UDP traffic. PSNR was measured while streaming MPEG videos from a CN to a MN moving in the testbed. Evaluation results have shown that the proposed approach outperforms the standard NSIS with respect to the average data transmission rate and average time required to re-establish sessions after handoffs. The proposal is even capable of maintaining a stable data rate after handoffs even in high-loaded network (93.5 Mbps background traffic in

the experiment). Considering PSNR, the experiments have shown that the proposed mechanism restores a stable data rate after a short service disruption due to the handoff and, thus, minimizes the PSNR. The standard NSIS suffers from significant video quality degradation after handoffs due to the high load in the network.

Pros and cons: although the proposal shows good performance due to its in-advance reservation, this mainly depends on the assumption that the MN is always capable of tracking its movements and detecting the new AR. This necessitates dependency on layer 2 triggers and makes the approach hardware-specific. Failing to correctly predict movements results in high service disruption. Moreover, the speed and size of overlapping areas play a major role. There are no studies analyzing this till now. The fact that only the crossover node should be detected, which alone will be responsible for establishing the session to the new location of the MN, may result in a suboptimal path between the CN and the MN (CN → crossover node → MN), especially in mesh-based networks. A main advantage of the proposal is that the in-advance reservation stores only the states on the new path and does not actually reserve resources. Thus, there is no waste of resources.

The tunneling as well as triangular routing problem is not addressed in the proposal. Both problems depend on the mobility management protocol applied. For instance, in case MIPv4 is applied as a mobility management protocol and data packets are forwarded via a triangular route, both problems appear. A main advantage of the proposal is the avoidance of the double reservation of resources problems, since crossover nodes are responsible of the reservation.

Concerning the network topology, no restrictions are made. However, a hierarchical topology would be better. To enable an employment of the proposal, NSIS-aware nodes should be introduced to the network. Moreover, MNs, CNs, ARs and all NSIS-enabled nodes locating in the access network should be updated. In terms of security, it is provided by the GIST, which depends on existing authentication and key exchange protocols [MKM10].

4.4.7 NSIS-based Semi-Proactive Resource Reservation

Network topology and basic principles: the approach presented in [TMT08] aims at accelerating the reservation of resources after handoffs by means of a semi-proactive resource reservation procedure. The proposal assumes an NSIS framework and uses QoS NSLP as a QoS signaling protocol. The approach's developers state that the protocols working proactively produce significant improvement in the performance (e.g. fast resource reservation, reduced lost packets, etc.) since they reserve resources in advance. This results, however, in a waste of resources, since resources are typically reserved in more than one candidate subnet. Moreover, these resources cannot be used by other MNs even if the MN has not yet started the handoff. The problem will be more critical if the handoff itself was not successful. Therefore, in order to maintain the performance improvement resulting from proactive approaches without wasting resources, a semi-proactive algorithm is proposed. The basic idea of this algorithm says that the protocol should do as much as possible proactively. Resource reservation, however, must be accomplished after the handoff. For this purpose, the developers of this approach propose that all required GIST states are created before the handoff occurs. RESERVE messages, however, are allowed to be sent after handoff completion.

Operation overview: let us first consider the receiver-initiated reservation and assume that the MN is close to the sender of the data flow, see Figure 4.22. As the MN, or possibly the QNR closest to the MN[1], notices that a handoff is about to happen, it issues a QUERY mes-

[1] In case the MN, itself, does not support NSIS and even may exist in a domain that uses different QoS mechanisms. However, the current domain contains gateway(s) capable of interacting with NSIS-based domains. In this case, the gateway(s) will be the QNR closest to the MN.

sage to the CN, or possibly the QNI closest to the CN[1]. The CN in turn waits until the handoff is completed and sends a RESERVE message to the MN/closest QNR to reserve resources on the new path. Once the MN/closest QNR receives the RESERVE message, it responds by sending a RESPONSE message.

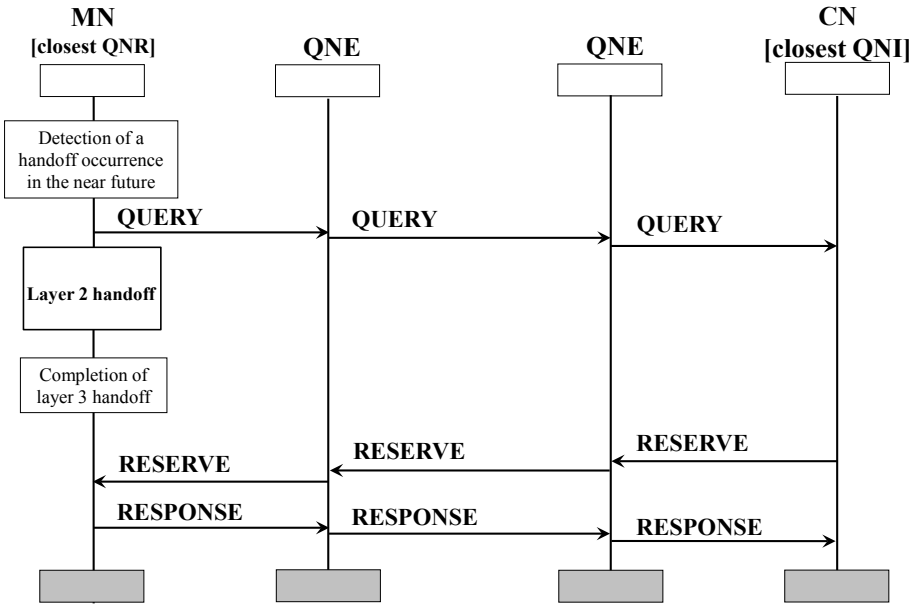


Figure 4.22: Semi-proactive reservation (receiver-initiated)

Sender-initiated reservation works in a somewhat similar manner. As the CN or possibly the QNI closest to the CN is informed that a handoff may happen in the near future, it sends a QUERY message to the MN on the new path or possibly the QNR closest to the MN's new location. After handoff completion, the MN/closest QNR waits for a RESERVE message from the CN/closest QNI. As the MN/closest QNR receives this message, it responds by sending a RESPONSE message. The sender-initiated procedure is displayed in Figure 4.23.

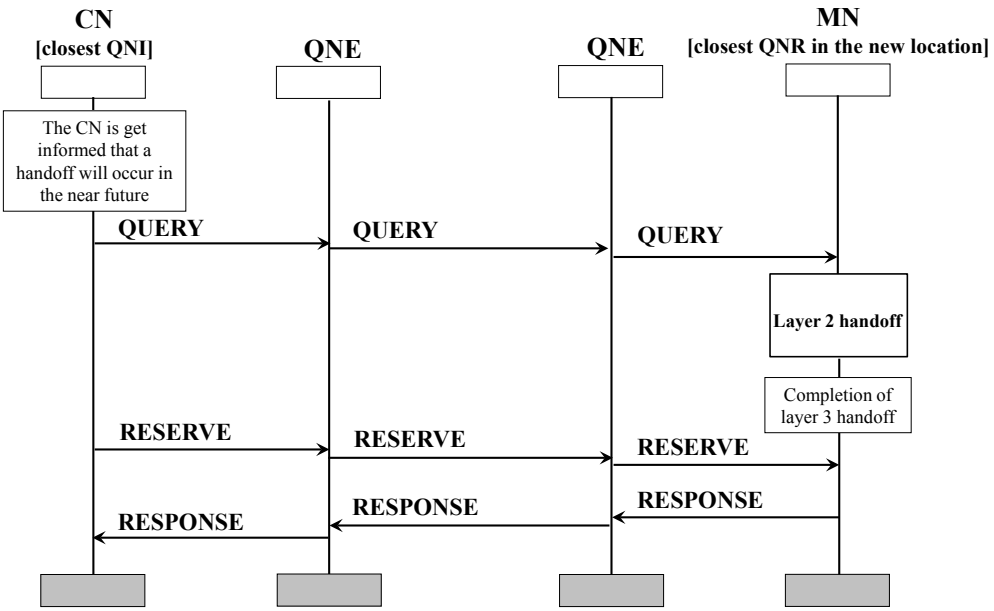


Figure 4.23: Semi-proactive reservation (sender-initiated)

[1] In case the CN, itself, does not support NSIS. However, it is located in a domain that has gateway(s) capable of interacting with NSIS-based domains. In this case, the gateway(s) will be the QNI closest to the CN.

So as to achieve seamless handoffs, data packets should be buffered in the crossover node, which is defined as the last NSIS-aware node before the path between the CN and the MN's old location diverges from the path between the CN and the new MN's location.

Performance evaluation: the analysis of the proposal has shown that the semi-proactive reservation procedure used reduces the time required to reserve resources on the new path. Moreover, the buffering in the crossover node does not significantly distrub the QoS after handoffs.

Pros and cons: the main advantage of the proposal is that the mentioned performance improvements are achieved without wasting resources, since resources are reserved only after handoffs' completion. However, the semi-proactive resource reservation procedure requires MNs to be able to track their movements and detect the possible new subnets and, consequently, possible crossover nodes (a crossover node to each new predicted subnet). Sure, this consumes MNs power and implies considerable overhead as well as dependency on layer 2 triggers.

For sender-initiated reservation, the CN/closest QNI should trigger the semi-proactive handoff procedure as it detects that the MN will make a handoff in the near future. How the CN/closest QNI gets notified of possible movements of the MN is not discussed.

Since the resource reservation is done using QoS NSLP protocol, the proposal suffers neither from the tunneling nor triangular routing nor double resource reservation problem. For the network topology required, the proposal does not constrain the topology. Furthermore, to employ the proposal, NSIS-aware nodes should be introduced to the network and MNs (or closest QNRs), CNs (or closest QNIs) and all NSIS-enables nodes locating in the domain must be updated.

Concerning security, it is provided by the GIST and, as known, depends on existing authentication and key exchange protocols.

4.5 Hybrid Approaches

As mentioned in section 4.2, hybrid techniques attempt to leave the implementations of protocols responsible for mobility management separate from those handling QoS, similar to loose-coupled solutions. From the operation point of view, however, both work as integrated solutions for mobility management and QoS, as is the case with hard-coupled approaches. It is clear, then, that hybrid solutions inherit the properties of both hard- and loose-coupled solutions.

Most known hybrid techniques are less efficient than hard-coupled ones. However, they are also less complex and, thus, more applicable to future mobile communication networks. Compared to loose-coupled techniques, hybrid approaches are, in principle, more complex and less applicable to future mobile communication networks. In order to discuss this category of solutions in more details, well-known hybrid solutions will be investigated.

4.5.1 RSVP and MIPv6 Interoperation Framework

A proposal to interoperate RSVP with MIPv6 in order to simultaneously support QoS and mobility is presented in [SSL01].

Network topology and basic principles: the basic ideas include encapsulating mobility information within RSVP control messages, on the one hand, and achieving a flow-transparent QoS model by keeping a unique flow identifier regardless of MN's movements, on the other. In order to convey mobility information, a new RSVP object named "mobility object" is in-

troduced. So as to maintain a unique flow identifier, the MN's HoA is chosen as the flow identifier for RSVP sessions. The routing of RSVP control messages, however, is done based on the CoA. The network topology and basic principles of the proposal are illustrated in Figure 4.24.

Operation overview: the operation of the approach is rather simple and distinguishes between two cases, namely whether the MN acts as a Mobile Sender (MS) or Mobile Receiver (MR). For the first case where the MN acts as a MS, the MN immediately sends a PATH message towards the CN after completing the layer 2 handoff and acquiring a new CoA. The PATH message contains the mobility information of the MN encapsulated in a mobility object. As the crossover router, termed Sender Nearest Common Router (SNCR), receives the PATH message, it responds directly by sending a RESV message to the MN. Notice that the PATH message is further transmitted to the CN, which replies with a RESV message containing the mobility information encapsulated in a mobility object. However, this RESV message only refreshes the reservation and does not allocate resources (see Figure 4.24 and ure 4.25, which show the network topology along with basic principles and the handoff procedure for the case where the MN is acting as a MS, respectively). In addition to sending the RESV message, the SNCR sends a RSVP TEARDOWN message to the MN on the old path to release the resources allocated for the MN on this path.

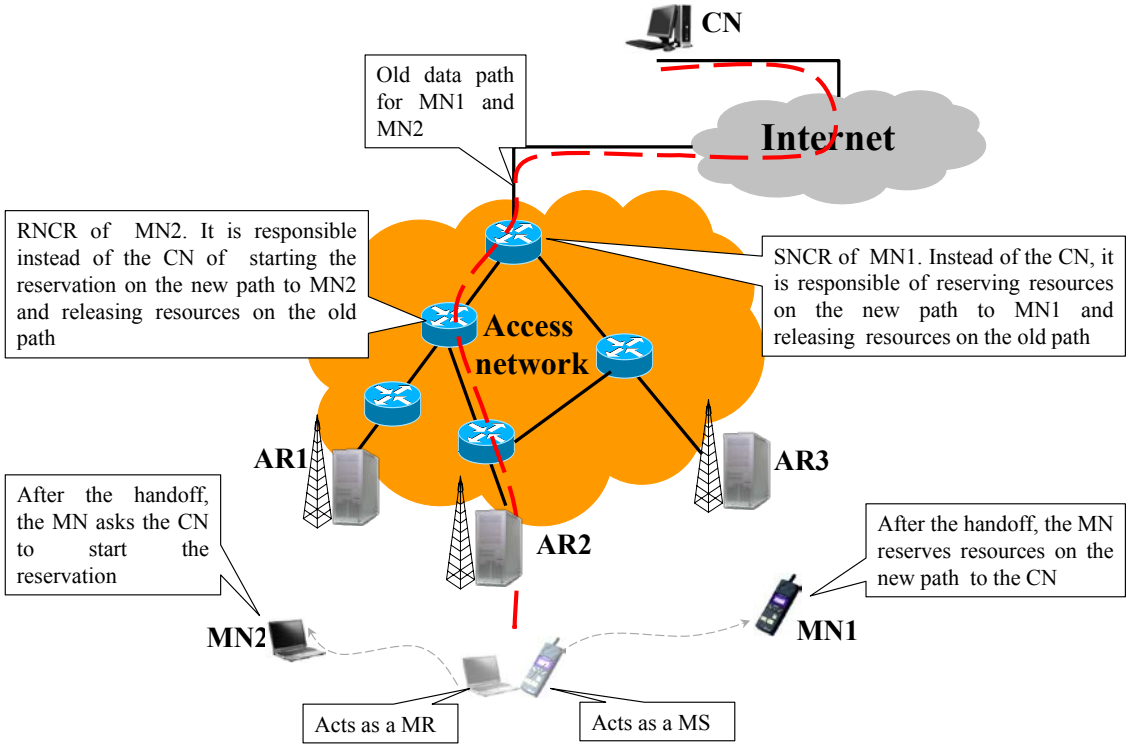


Figure 4.24: Network topology and basic principles of RSVP and MIPv6 interoperation framework (for both cases, the MN is a MS and a MR)

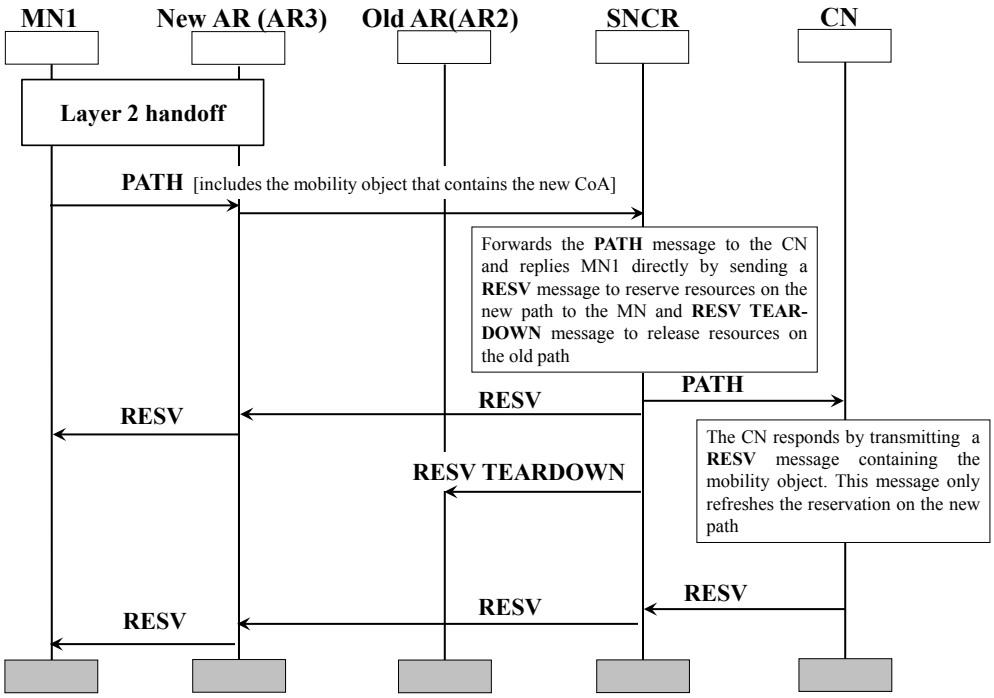


Figure 4.25: Handoff procedure employing RSVP and MIPv6 interoperation framework (the MN operates as a MS)

For the case where the MN acts as a MR, the PATH message should be sent from the CN or the crossover router, termed as Receiver Nearest Common Router (RNCR), see Figure 4.26. Of course, the MN will consume a considerable amount of time waiting for the PATH message after the handoff. To avoid this, the MN transmits a PATHREQ message towards the CN. As the RNCR receives this message, it exchanges PATH and RESV messages with the MN's new CoA. In addition, the RNCR releases the resources on the old path.

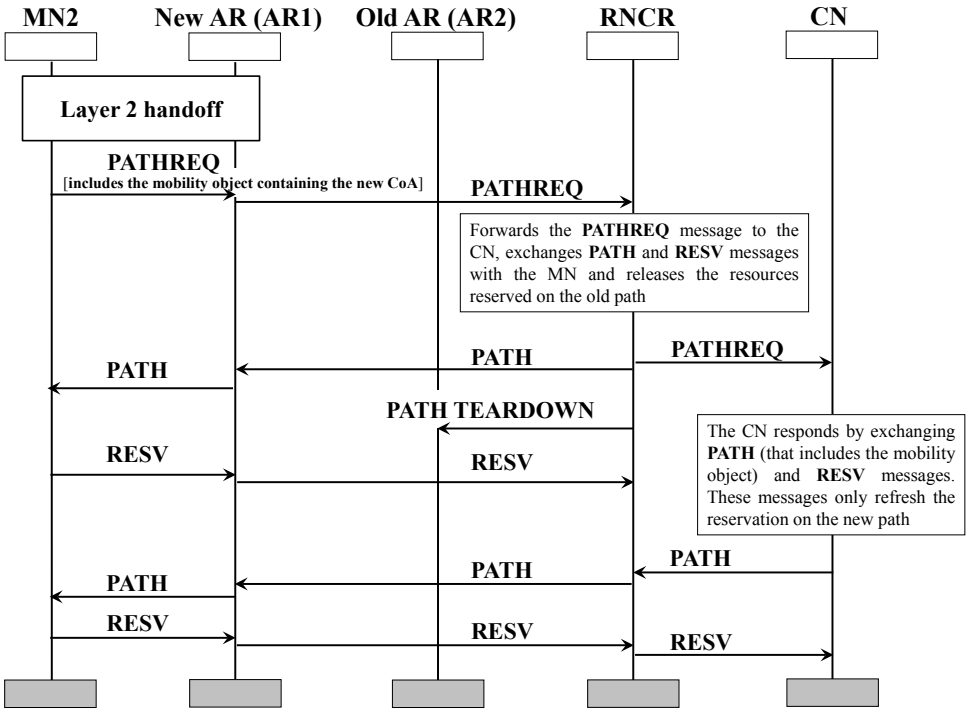


Figure 4.26: Handoff procedure employing RSVP and MIPv6 interoperation framework (the MN operates as a MR)

Performance evaluation: simulation results presented in [SSL01] show that the proposed flow-transparent QoS model minimizes the time required to reserve resources after handoffs as well as the overhead resulting from this reservation compared with the non flow-transparent QoS model[1]. Based on the deployed topology, the proposed scheme achieves a reduction of 13.9 % to 95.1 % in terms of data packet loss during handoffs as compared to the non flow-transparent QoS model. With respect to the number of data packets experiencing more delay than the allowed play-out time and, thus, considered to be lost, the proposed scheme drops 77 % to 100 % less than the non flow-transparent QoS model. Considering robustness against control messages dropping, the flow-transparent QoS model is significantly better.

Pros and cons: based on the analysis introduced above, one can see that significant performance improvements have been achieved due to the maintenance of unique flow identifiers for RSVP sessions in spite of the MNs movements. The design of the proposal enables the localization of the reservation of new resources as well as the release of old ones after handoffs, which is the main reason behind the fast handoffs achieved. In addition, the maintenance of unique flow identifiers for RSVP sessions and the dependency on crossover nodes to quickly complete reservations after handoffs eliminate the double resource reservation problem.

Because MIPv6 is applied as a mobility management protocol, the proposal does not suffer from the tunneling and the triangular routing problem as long as triangular routing is not explicitly applied.

A main advantage of the proposal is that the obtained performance improvements do not necessitate passive reservations or tracking MNs movements and prediction of possible new ARs. Furthermore, no network restrictions are required. However, all RSVP-enabled nodes should be introduced to the network and should be updated to understand the proposal. The updates must be achieved on MNs and CNs, as well, which negatively affects the employment of the proposal. A main drawback of the proposed scheme is the lack of security, since security issues have not yet been addressed.

4.5.2 QoS Extension for Next Step in Signaling in Mobile IPv6 Environment

Network topology and basic principles: the basic idea of the technique proposed in [LPN07] is to integrate the control messages of QoS NSLP protocol within MIPv6 control messages, so that resources are reserved during the registration process of MIPv6. To enable such integration, the authors propose an addition of a new IPv6 extension header to convey QoS NSLP messages.

Operation overview: the proposal is rather simple, since it employs MIPv6 only after the handoff. To clarify this, let us first consider the receiver-initiated reservation. When the MN moves into the range of a new AR, it sends a BU message carrying the RESERVE message towards the CN[2]. As a crossover node (i.e. a node that currently has a QoS NSLP state for the MN) receives this message, it responds by sending a RESERVE message with the Teardown flag set to the MN's old location to release the resources reserved on the old path. As the CN receives the BU message, it responds by sending a BA message with a RESPONSE message encapsulated inside. Keep in mind that all QoS NSLP enabled-nodes located on the path be-

[1] The non flow-transparent QoS model is the model that uses the standard RSVP after the handoff, which requires reserving resources on the whole new path between the CN and the new MN's location.

[2] We assume that route optimization is applied. Clearly, a **BU** and **BA** will be exchanged with the HA. However, these messages will be built according to standard specification of MIPv6

tween the MN and the CN must detect QoS NSLP messages encapsulated inside MIPv6 control messages and react accordingly. The above described procedure is shown in Figure 4.27.

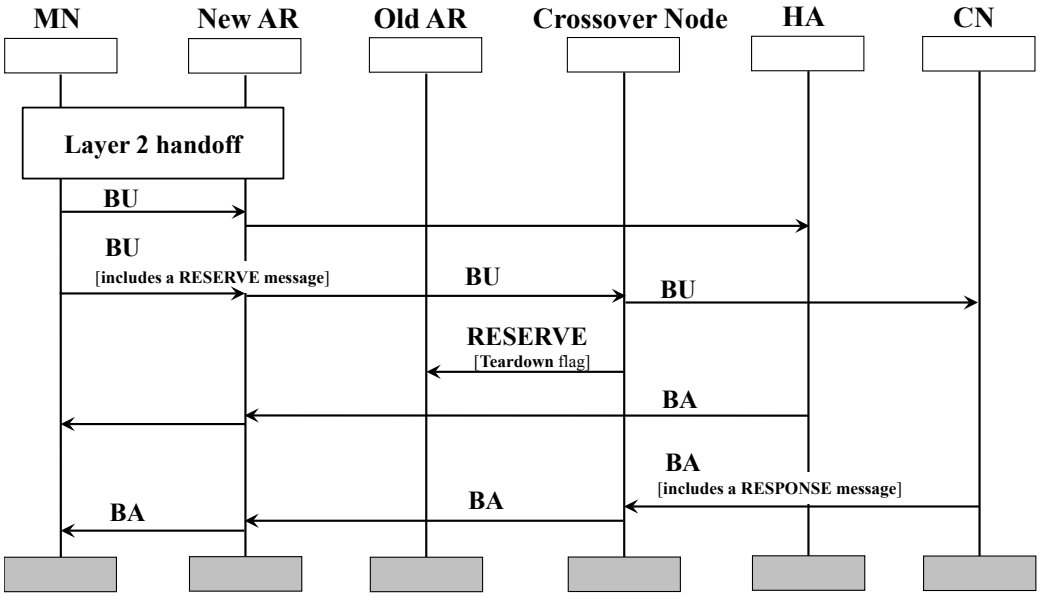


Figure 4.27: Handoff procedure (receiver-initiated reservation)

The sender-initiated reservation works in a similar way. However, the BU message does not carry a QoS NSLP message with it. As soon as the CN receives the BU message, it transmits a BA message conveying a QUERY message. Once the MN receives the BA and extracts the QUERY, it responds according to the standard QoS NSLP specification. Similar to the NSIS-based semi-proactive resource reservation, this approach buffers the packets directed to the MN at the crossover node during the handoff.

Performance evaluation: asimple performance evaluation based on simple analytical models was achieved in [LPN07]. The analysis has shown that the proposed approach significantly accelerates the reservation after the handoff as compared to the basic QoS NSLP protocol. Moreover, due to the transmission of a RESERVE message with the Teardown flag set directly after the crossover node receives the RESERVE message, the resources on the old path are released very quickly.

Pros and cons: the proposal does not suffer from the tunneling and the triangular routing problem, since MIPv6 is applied as a mobility management protocol. Of course, the route optimization is assumed as a default routing mechanism. Sure, in case the triangular routing is applied, both problems appear.

A main advantage of the proposal is that it eliminates the double resource reservation problem due to the usage of QoS NSLP. Moreover, the proposal relies neither on passive reservations nor on layer 2 triggers.

Although the proposed technique does not constrain network topology, a main drawback of it is that all QoS NSLP nodes as well as the nodes supporting MIPv6 must be updated. Notice that QoS NSLP nodes have to process MIPv6 control messages to extract QoS NSLP messages, while MIPv6-enabled nodes have to consider an extra mobility header in MIPv6 control messages. This disadvantage complicates the applicability of the proposal and negatively affects its scalability. Keep in mind also that All NSIS-enabled nodes should be introduced to the network. Similar to the proposals that depend on the NSIS framework, security is provided by GIST. Sure the integration of QoS NSLP control messages inside MIPv6 control mes-

sages also benefits from the security actions related to MIPv6, e.g. the authentication of MIPv6 control messages.

4.6 Qualitative Comparison

A detailed comparison between the schemes described in this chapter with respect to the

- tunneling problem,
- triangular routing problem,
- double reservation of resources,
- passive reservation,
- dependency on layer 2 triggers,
- network topology,
- new nodes that should be introduced to the network,
- nodes that should be updated and
- security

is presented in Table 4-1.

Note that the tunneling and triangular routing problems affect the QoS the user receives. These problems produce extra delay to packets and eventual degradation in QoS. The double reservation of resources and passive reservation problems are of a special importance from the service provider/network operator point of view, since they affect the efficiency in terms of the use of resources. Dealing with resources efficiently enables the provider to reduce the cost of operating the solution proposed and also to serve more users. The dependency on layer 2 triggers, network topology, the number of new nodes and those that should be updated determine how complex the employment of the new proposal will be. Security is an important metric since users as well as network operators and service providers will not interest in insecure solutions.

Approach	Tunneling problem[1]	Triangular routing	Double resource reservation during handoffs	Passive reservation	Dependency on layer 2 triggers	Network topology	New nodes that should be added[2]	Nodes that should be updated	Security
WLRP	Not addressed	Not addressed	Not addressed	Yes (to build Mob-profiles)	Yes (to build Mob-profile)	No restrictions	-	MNs and BSs	Not addressed
Mobile extensions to RSVP	Not addressed	Not addressed	Yes (due to the introduce of merging/anchor points, this problem is reduced)	No	Yes (to predict possible new subnets)	No restrictions	fulcrum node	MNs, all subnets, all RSVP-enabled nodes within the domain	Not addressed
MRSVP	Yes (if routing optimization is not used)	Yes (if routing optimization is not used)	Yes (due to the introduce of merging/anchor points, this problem is reduced)	Yes	Yes	No restrictions	-	MNs, CNs, HA, all proxies in the domain	Not addressed
HMRSVP	Only in case MIP is used to support inter-domain mobility and triangular routing is used to forward data packets to the GFA that controls the domain serving the MN	Only in case MIP is used to support inter-domain mobility and triangular routing is used to forward data packets	For intra-domain mobility (solved due to the splitting of the end-to-end reservation in the GFA. However, this solution is not optimal) For inter-domain mobility (not solved due to the use of plain RSVP to reserve resources)	For inter-domain mobility only	For inter-domain mobility only	Hierarchical	The nodes required to operate MIPRR (GFA)	MNs, CNs, HA, GFA and all FAs of the domain	Not addressed

[1] The tunneling problem is the problem faced mostly when the MN residing in a range of a foreign subnet and reserves an end-to-end session to the CN it communicates with (via the HA). However, due to the movements, the HA must tunnel data packets to the MN's CoA. These tunneled packets are not recognized by the reserved end-to-end session.

[2] Beyond the nodes that should be introduced to support MIP.

Approach	Tunneling problem[1]	Triangular routing	Double resource reservation during handoffs	Passive reservation	Dependency on layer 2 triggers	Network topology	New nodes that should be added[2]	Nodes that should be updated	Security
Simple QoS	No	Yes	Yes (RSVP tunnel is reserved between the HA and new FA regardless of the existence of crossover nodes on this path)	No	No	No restrictions	-	HA, MNs and all FAs of the domain	Not addressed
LRSVP	In case MIP is used to support mobility	In case MIP is used to support mobility	Yes (because the reservation inside the LRSVP domain should always be repaired between the MN and LRSVP proxy after handoffs (no crossover nodes considered))	No	No	Hierarchical	LRSVP proxy	MNs and all RSVP-enabled nodes in the access domain	Not addressed
RSVP multicast-based mobility support	Yes (in case MIP triangular routing is used)/ No (in case routing optimization is applied)	Yes (in case MIP triangular routing is used)/ No (in case routing optimization is applied)	Yes (due to the introduce of merging/anchor points, this problem is solved. However, the solution is not optimal)	Yes	Yes	No restrictions	Mobility proxies and merging points	MNs	Not addressed
Seamless NSIS-based QoS guarantees with advanced reservation	Yes (in case MIPv4 is applied as a mobility management protocol and data packets are forwarded via a triangular route)/ NO (in case MIPv6 is used)	Yes (in case MIPv4 is applied as a mobility management protocol and data packets are forwarded via a triangular route)/ NO (in case MIPv6 is used)	No	No (only states are created in advance)	Yes	No restriction.	All NSIS-enabled nodes	MNs, CNs, ARs and all NSIS-enables nodes locating in the access network	Not addressed

[1] The tunneling problem is the problem faced mostly when the MN residing in a range of a foreign subnet and reserves an end-to-end session to the CN it communicates with (via the HA). However, due to the movements, the HA must tunnel data packets to the MN's CoA. These tunneled packets are not recognized by the reserved end-to-end session.

[2] Beyond the nodes that should be introduced to support MIP.

Approach	Tunneling problem[1]	Triangular routing	Double re-source reserva-tion during handoffs	Passive reser-vation	Dependency on layer 2 triggers	Network topo-logy	New nodes that should be add-ed[2]	Nodes that should be up-dated	Security
NSIS-based semi-proactive resource reservation	No	No	No	No (only states are created in ad-vance)	Yes	No restric-tions	All NSIS-enabled nodes	MNs (or closest QNRs), CNs (or closest QNIs) and all NSIS-enables nodes locating in the domain	Not addresse d
RSVP and MIPv6 interoperation framework	No	No	No	No	No	No restric-tions	All RSVP-enabled nodes	MNs, CNs and all RSVP-enables nodes locating in the access network	Not addresse d
QoS extension for NSIS in MIPv6 envi-ronment	No	No	No	No	No	No restric-tio-ns	All NSIS-enabled nodes	MNs, ARs, CNs, all NSIS-enables nodes locating in the access network	Not address ed

Table 4-1: Comparison of the studied protocols coupling mobility management and QoS

[1] The tunneling problem is the problem faced mostly when the MN residing in a range of a foreign subnet and reserves an end-to-end session to the CN it communicates with (via the HA). However, due to the movements, the HA must tunnel data packets to the MN's CoA. These tunneled packets are not recognized by the reserved end-to-end session.

[2] Beyond the nodes that should be introduced to support MIP.

With respect to the tunneling problem, one observes that most described solutions suffer from this problem, since resources are reserved between the CN and the MN's new location, on the one hand, and data packets are mostly tunneled between the HA and the new subnet hosting the MN, on the other hand. This prevents tunneled packets after handoff from being recognized as packets for the MN that reserved the resources, thus, degrades QoS. The main reason for this problem is the aim at end-to-end QoS guarantees between the CN and the MN. In principle, the solutions that employ MIPv4 in triangular routing mode, such as LRSVP (in case MIPv4 is used to support mobility), MRSVP, HMRSVP (only in case MIPv4 is used to support inter-domain mobility and triangular routing is used to forward data packets to the GFA that controls the domain serving the MN), RSVP multicast-based mobility support and seamless NSIS-based QoS guarantees with advanced reservation, suffer from the tunneling problem. Of course, employing route optimization avoids this drawback. The techniques based on MIPv6 typically do not suffer from the tunneling problem since CNs support mobility and route optimization is applied in most cases. Simple QoS employs MIPv4 in triangular routing mode. However, it does not suffer from the tunneling problem. The reason behind this is the building of an extra RSVP session between the HA and the new subnet serving the MN to serve tunneled packets. WLRP focuses on resource reservation on wireless links and does not address issues like the tunneling problem. The triangular routing problem is related to the tunneling problem. In general, the solutions that suffer from the tunneling problem also suffer from the triangular routing problem since the tunneling problem is a result of the triangular routing. Consequently, the approaches that can cope with the tunneling problem do not suffer from the triangular routing problem.

Considering the double reservation of resources, it can be noticed that most solutions have focused on this problem and provided solutions for it. The solutions that use RSVP as a QoS signaling protocol solve this problem based on one of four principles. The first says that all RSVP nodes should be updated to be able to change the flow specification of RSVP sessions after handoffs, so that no new reservation is necessary if a node currently has a reservation for the same MN and the same flow. RSVP and MIPv6 interoperation framework is an example of such approaches. The second principle is motivated from the fact that changing all RSVP nodes is not possible in practice and will prevent the deployment of the proposed solutions in real networks. So, it would be better to introduce merging/anchor points to the network to merge passive and active reservations so that no double resource reservation will result. Clearly, this may not eliminate the double resource reservation problem. However, it minimizes the affected resources. Mobile extensions to RSVP, MRSVP and RSVP multicast-based mobility support approaches are example of such schemes. The third principle aims at localizing the resource reservation inside certain domains by splitting the session in intermediate nodes. Thus, movements of MNs affect only the sessions established between the MNs and these intermediate nodes, while the sessions between the intermediate nodes themselves and CNs are not affected. Obviously, such principles speak against end-to-end reservations. Therefore, a considerable effort should be done to make this splitting transparent to CNs and possibly to MNs, as well. Examples of such solutions include HMRSVP and LRSVP. Again, this principle does not present a complete solution, since double resource reservation may appear inside the domains due to the use of the plain RSVP. However, the resources affected due to this problem are minimized. The fourth principle is a mix of the above three principles and is utilized by Simple QoS, since this solution attempts to preserve an end-to-end session and, at the same time, uses an anchor point (the HA) to establish the reservation after the handoff until the end-to-end reservation is built. In addition, Simple QoS enables the anchor point as well as the new FA to change the specifications of the tunnel reserved between them without necessitating a new reservation. Although, Simple QoS invests major efforts to cope with double resource reservation problem, the problem may appear because of the tunnel re-

served between the HA and the new FA. Keep in mind that this tunnel is reserved regardless of the existence of nodes that are parts of the paths between the HA and the old FA as well as new FA. The solutions that employ NSIS as a QoS signaling protocol do not suffer from the double resource reservation problem, since this problem is avoided in NSIS by concept.

Regarding passive reservation, one can distinguish between three kinds of strategies. The first strategy states: to provide a guarantee of maintaining a certain QoS level, the developed scheme should work proactively and resources should be reserved passively in neighbor subnets, to which the MN is predicted to move. Clearly, this provides the best guarantee that QoS degradation will not take place. However, wasted resources are the result of such guarantees. Examples of approaches following this strategy are: WLRP, MRSVP and RSVP multicast-based mobility support. The second strategy attempts to solve the problem of wasted resources resulting from the passive reservation by working in a semi-proactive way, so that states are created in-advance in new predicted subnets. However, the reservation of resources occurs after the completion of handoffs. This provides a guarantee of a certain QoS level, as well. Obviously, the probability of maintaining this guarantee is less than that offered by the solutions employing the first strategy. Mobile extensions to RSVP, seamless NSIS-based QoS guarantees with advanced reservation and NSIS-based semi-proactive resource reservation are examples of approaches built based on the second strategy. The techniques following the first and second strategies require MNs to be capable of predicting their movements and, thus, rely mostly on link layer information. The third strategy states that passive reservation should be avoided and the QoS guarantees have to be provided in a way other than passively reserving resources, e.g. by introducing intermediate nodes to split RSVP sessions, etc. Examples include HMRSVP, Simple QoS, LRSVP, QoS extension for NSIS in MIPv6 environment and RSVP and MIPv6 interoperation framework. The schemes implementing the principles of the third strategy provide the minimum guarantees that MNs will not suffer from QoS degradation after handoffs. They do not, however, place any requirements on MNs regarding tracking of movements. Considering the dependency on layer 2 triggers, it is clear that the approaches working in a proactive or semi-proactive way depend on layer 2 triggers, while the schemes that operate in a fully reactive manner do not. It is obvious as well that the dependency on layer 2 triggers mostly makes the approaches technology-dependent and violates the separation between the layers of the TCP/IP reference model.

Let us now compare the studied approaches with respect to the network topology required. We can see that only HMRSVP and LRSVP require a hierarchical network topology, while the other techniques studied do not put any restrictions on the topology. Notice that the techniques that require deploying a hierarchical topology work properly only if such a topology is used. On the contrary, the techniques that do not put any restrictions on the topology can be used deploying both hierarchical and mesh-based topologies. In most cases, however, mesh topologies are more adequate.

Regarding the new nodes that should be introduced to the network, one can notice that only the solutions that attempt to use merging/anchor points or localize the reservation of resources inside a certain domain (as long as the movements of MNs are restricted to subnets belonging to the domain) introduce new nodes to the network, e.g. Mobile extensions to RSVP, MRSVP, HMRSVP, RSVP multicast-based mobility support and LRSVP. Clearly, the smaller the number of new nodes that should be introduced to the network, the better the approach and the smaller the cost it produces. Keep in mind that we do not consider the nodes that should be introduced to operate MIP, i.e. the HA and FA/AR.

Considering the nodes that should be updated in the network, all solutions require updates. The fewest updates are required by RSVP multicast-based mobility support, which requires updating the MNs only. Clearly, the schemes requiring few updates are better than those ne-

cessitating many updates. Notice that the new nodes that should be introduced to the network are not considered here.

Taking the security into account, one can clearly see that the proposed approaches do not consider security as a main part of their specification. Security is applied in the mobility management protocol or is ensured through extra applications. Even the schemes that depend on the NSIS framework relay on existing security protocols.

The analysis of previous work showed that there are several prior efforts to develop solutions capable of coupling mobility management and QoS techniques in a way enabling the achievement of seamless handoffs simultaneously with QoS guarantees so that real-time applications' requirements are satisfied. The approaches have attempted to achieve the mentioned goals either by working proactively or semi-proactively or by localizing mobility and resource reservation inside access networks. However, the applied mobility management protocols are stated to be unable to achieve seamless handoffs, see [Dia10] for details. Clearly, this significantly reduces performance and prohibits the capability of satisfying real-time applications. Therefore, there is a need to develop a new solution that

1. uses a mobility management solution capable of achieving seamless handoffs even for MNs moving at high speeds,
2. localizes the mobility management as well as resource reservation actions that should be achieved after handoffs without restricting the physical network topology or introducing new intermediate nodes beyond the nodes currently known from the standard mobility management solution, i.e. MIP,
3. minimizes the nodes that should be updated,
4. satisfies real-time requirements by doing some proactive tasks without relying on layer 2 information and
5. secures resource reservation after handoffs.

To meet the above mentioned requirements, a new solution named QoS-aware Mobile IP Fast Authentication protocol (QoMIFA) has been developed. The new solution integrates MIFA as a mobility management protocol with RSVP as a QoS reservation protocol [AMB06], [AMD10]. MIFA is selected due to the fast and secure handoffs it provides, while RSVP is chosen because it presents the standard solution used to support QoS in current IP networks. The specification of QoMIFA is the topic of the next chapter.

4.7 Conclusion

This section has reviewed and analyzed previous efforts to couple mobility management and QoS techniques. The main results obtained from the analysis can be summarized as follows:

- From the performance point of view, the literature states that the best solutions are the hard-coupled ones. This is because these solutions typically perform resource reservations after handoffs more quickly than the approaches of the other two strategies. Hybrid approaches come next, while loose-coupled solutions are the worst.
- From the efficiency point of view, the literature also states that hard-coupled are again the best solutions, as this kind of solution handles mobility as well as QoS at the same time. They require fewer signaling messages than when QoS and mobility are dealt with using a separate solution for each. Therefore, hard-coupled schemes consume less network resources to operate. Again, hybrid techniques come next followed by loose-coupled schemes.
- From the complexity point of view, previous studies show that hard-coupled solutions are the most complex. Hybrid solutions are not as complex as hard-coupled solutions, however, more complex than loose-coupled techniques.

- With respect to the applicability to current and future mobile communication networks, loose-coupled solutions are stated in the literature to be the best. Hybrid techniques come next, followed by hard-coupled ones.

One notices from the analysis that, although hybrid solutions inherit properties of both hard- and loose-coupled techniques, they still perform worse than hard-coupled solutions and are also less efficient. Therefore, further development of hybrid solutions to perform as well and be as efficient as hard-coupled solutions is challenging and will be of major interest. Facing this challenge was the main motivation behind the development of QoMIFA.

Chapter 5: QoS-aware Mobile IP Fast Authentication Protocol

The previous chapter has shown that hybrid solutions are promising. However, considerable work is still necessary to develop solutions capable of satisfying real-time requirements, while still as good (in terms of efficiency and performance) as hard-coupled solutions and as simple (in terms of performance) and deployable as loose-coupled schemes. Therefore, we propose QoMIFA [AMB06], [AMD10] to meet these requirements. QoMIFA is a semi-proactive hybrid protocol that integrates RSVP as a QoS reservation technique and uses MIFA as a mobility management protocol.

The new proposal is the focus of this chapter, which is structured as follows: section 5.1 provides the basic ideas of the proposal, while section 5.2 briefly presents the operation overview. The detailed operation of QoMIFA is provided in section 5.3. The error recovery mechanisms provided by our new proposal are detailed in section 5.4. Finally, the chapter is concluded in section 5.5.

5.1 *Basic Ideas*

Semi-proactive behavior: based on the L3-FHRs that MIFA employs, movements of MNs can be predicted; thus, MNs-specific data are distributed in advance. This enables a fast re-authentication of MNs after handoffs and, thus, a fast resumption of communication. This can be further utilized to discover the availability of resources on the subnets located in the current subnet's vicinity, where the MN will likely move. Notice that resources should not be reserved in advance to avoid wasting resources. Thus, it is beneficial to follow the semi-proactive behavior, which says: do as most as you can in advance without reserving resources. Reservation should occur after the handoff takes place. Therefore, the semi-proactive behavior of QoMIFA implies that the current subnet notifies the members of its L3-FHR of incoming MNs, so that these members can check in advance whether they have resources available and, if so, store RSVP states in advance, see [AMD10]. This significantly accelerates the actual allocation of resources after handoffs.

Hybrid structure: so as to retain the hybrid structure a new object called "mobility object" is introduced to RSVP. This object is used to encapsulate MIFA control messages. In this way, the MN only operates RSVP after the handoff and transmits MIFA control messages encapsulated inside the new introduced RSVP object. QoMIFA-enabled nodes handle the content of the mobility object, while standard RSVP-enabled nodes ignore this object.

As already mentioned above, RSVP states will be stored in advance in the members of the current subnets' L3-FHR. To enable the integration of MIFA and RSVP, these RSVP states are extended to include the MN-specific data, see Figure 5.1 for the basic concept of QoMIFA.

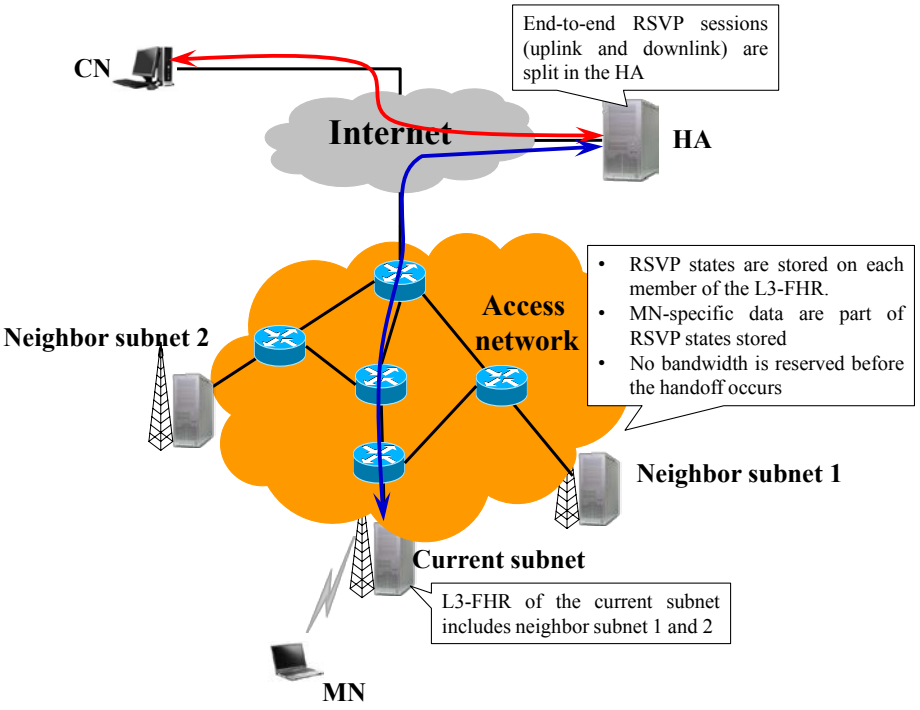


Figure 5.1: Basic ideas of QoMIFA

Split the end-to-end RSVP session: the review of the state of the art has shown that one of the basic principles used to accelerate the reservation of resources after handoffs is splitting the end-to-end session in a certain intermediate node, so that only the resources between MNs and this intermediate node are updated after handoffs. Of course, this may have negative impacts on the end-to-end session (between the MN and CN) in case the intermediate node does not accurately synchronize with the end-to-end session. Therefore, in order to utilize this principle without introducing intermediate nodes, we decided to split the end-to-end session in the HA, which must synchronize with the end-to-end RSVP session so that no break in the end-to-end RSVP session is experienced, see Figure 5.1.

QoMIFA initiatives: although the basic ideas utilized by QoMIFA are used by previous work as well, QoMIFA is not a pastiche of borrowed concepts. In addition to the distillation of the mentioned basic principles and their synthesis into a coherent architecture, QoMIFA enables achieving a significant performance improvement with neither constraining the network topology nor MNs, which is not achieved by previous work even those employing one or more of the concepts discussed. Moreover, QoMIFA is not technology-specific and does not rely on measurements achieved by the MN to determine possible new candidates as most previous work does. Important to note that MIFA itself is newly developed and has shown great performance improvements against its counterparts which motivated us to further develop this protocol towards QoS support. A major contribution in this concern is also the way MIFA is coupled with RSVP to form a hybrid solution that advances existing hybrid solutions and approaches the performance of hard-coupled ones since QoMIFA extends the old reservation with an additional bidirectional RSVP sessions between the old and new location of the MN. In this way, the time required to reserve resources on the new path is hidden and not essential anymore.

5.2 Operation Overview

QoMIFA uses the same network topology that MIFA and MIP deploy, see Figure 5.2 and Figure 5.3 for an illustration of the network topology and operation principles of QoMIFA.

Handoff procedure and uplink RSVP session: when the MN moves into the range of a new subnet, it first listens to an Agnt_Adv message from the FA serving the new subnet. The Agnt_Adv message can alternatively be solicited by broadcasting an Agnt_Sol message. After receiving an Agnt_Adv message, the MN exchanges PATH and RESV messages with the new FA, this in turn exchanges PATH and RESV messages with the old FA. Each PATH and RESV message contains a mobility object encapsulating a certain MIFA control message. The MN, new FA and old FA handle the mobility object and update the mobility binding of the MN if a successful registration occurs. Simultaneously, they reserve resources. The result of this briefly described procedure is an uplink RSVP session between the old FA and the MN, see Figure 5.2.

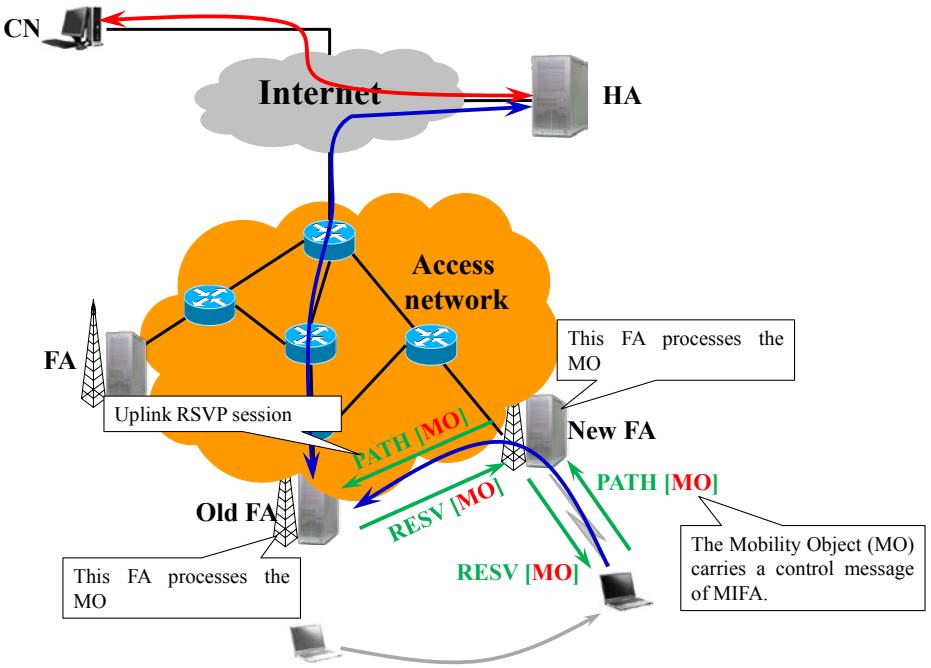


Figure 5.2: Handoff procedure and establishment of an uplink RSVP session after the handoff

Uplink data transmission with QoS guarantees: data communication is resumed on the uplink direct after the uplink RSVP session is established between the MN and the old FA. In other words, the uplink data transmission with QoS guarantees is first resumed via the old FA until the new resources between the HA and the new FA are reserved. Notice that this forwarding implies that the resources reserved between the HA and the old FA remain in use while the handoff procedure is in progress.

Downlink RSVP session: to establish a downlink RSVP session, the old FA simply exchanges PATH and RESV messages with the MN. Notice that the old FA begins establishing the downlink RSVP session when it receives notification of the MN's new location. Moreover, there is no longer a need to include a mobility object in these RSVP messages, see Figure 5.3.

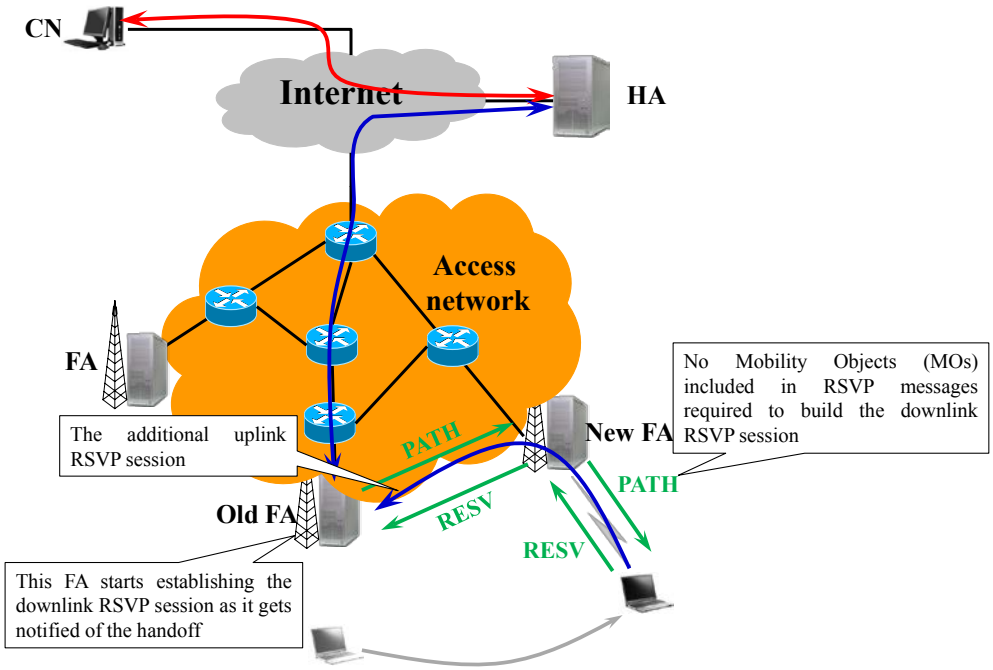


Figure 5.3: Establishment of a downlink RSVP session

Downlink data transmission with QoS guarantees: after the downlink RSVP session is established between the old FA and the MN, downlink data packets will be forwarded from the old FA to the MN's new location with QoS guarantees. Note that downlink data transmission with QoS guarantees is also resumed via the old FA until the new resources on the new path are reserved.

Handoff completion: as mentioned above, the MN resumes its communication with QoS guarantees without waiting for the handoff to be completed by the HA. However, to optimize the route between the HA and the new FA as well as to enable a continuation of the semi-proactive procedure of QoMIFA, the new FA exchanges PATH and RESV messages with the HA. This results in two RSVP sessions (one for the uplink and one for the downlink) between the HA and the new FA. Clearly, the resources reserved between the HA and the old FA as well as between the old FA and the new FA are no longer necessary. Therefore, they will be released.

Semi-proactive behavior: the last step QoMIFA should perform is the semi-proactive procedure, which implies that the new FA notifies its L3-FHR members of the possible incoming MN, so that the current members of the L3-FHR store RSVP states that include the **MN-specific data** required to authenticate and authorize the MN after the subsequent handoff, see Figure 5.4. Again no resources are reserved in advance before the actual handoff takes place. The semi-proactive procedure is explained in more details in the following.

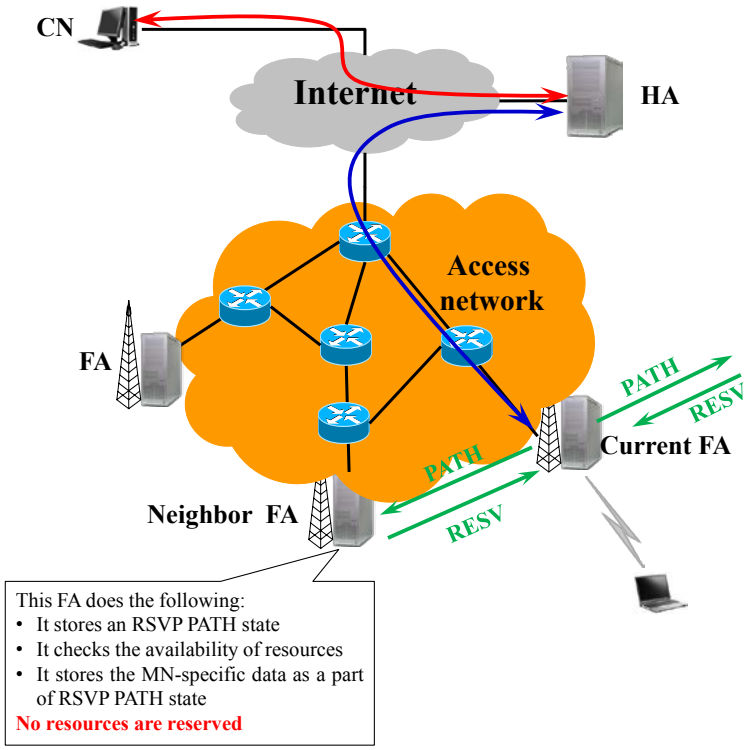


Figure 5.4: Semi-proactive procedure employing QoMIFA

5.3 Detailed Description of QoMIFA Operation

In order to operate QoMIFA, some updates are necessary. First, Agnt_Adv messages should be extended to include a QoMIFA flag (termed **QM** flag) to advertise the support of QoMIFA protocol. Agnt_Sol messages must also contain a similar flag to indicate that MNs prefer to use QoMIFA.

When the MN is powered on, it executes the initial registration and initial authentication exchange procedures known from MIFA. Then, the current FA performs the semi-proactive state creation procedure. This procedure is responsible for the establishment of RSVP states in the current L3-FHR members, thus, testing the availability of resources. When the MN moves into the range of a new FA, which is normally a member of the L3-FHR of the old FA, the MN executes a handoff procedure employing QoMIFA. Thereafter, the new FA performs the semi-proactive state creation procedure once more, and so on. The following describes in detail the above mentioned procedures.

5.3.1 Initial Registration and Initial Authentication Exchange Procedures

When the MN is switched on, it first establishes a wireless link and then waits for an Agnt_Adv message. After the MN receives an Agnt_Adv message with the **QM** flag set, it registers while initially employing MIPv4. The initial registration implies exchanging a Reg_Rqst and a Reg_Rply message between the MN and the HA.

Following the initial registration procedure, the initial authentication exchange procedure of MIFA is executed. The procedure implies the exchange of a M_P_Not and a M_P_Ack message between the current FA and the HA. This procedure aims at exchanging data that enables the use of MIFA during the subsequent registration with the new FA (e.g. authentication information, HA's features, etc.). The initial registration and initial authentication exchange procedures are shown in Figure 5.5.

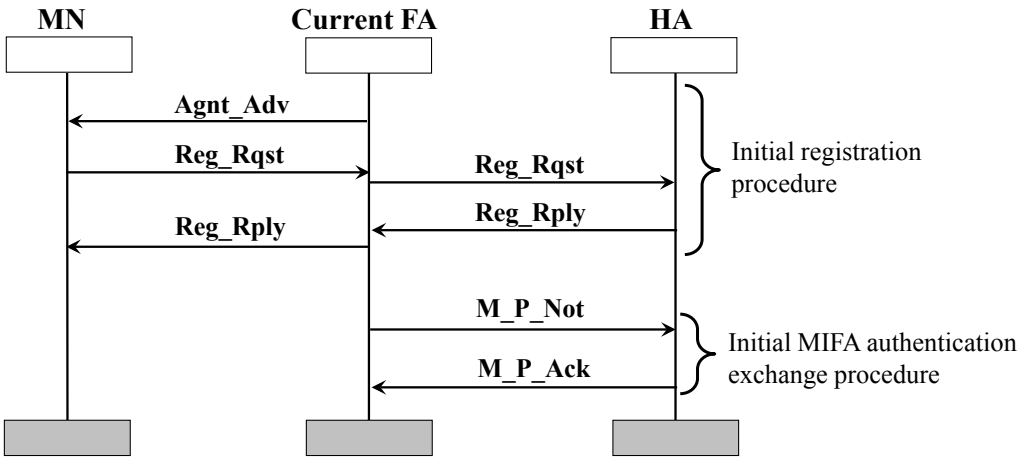


Figure 5.5: Initial registration and initial authentication exchange procedures

Notice that the initial registration as well as the initial authentication exchange procedure is executed once (only after the MN is switched on or wants to connect to the Internet for the first time).

5.3.2 Initial Reservation Procedure

The initial reservation procedure differs based on the role that the MN takes, i.e. receiver, sender or both. The following discusses all cases.

The MN operates as a receiver: when a CN wants to establish a downlink RSVP session with a MN, the CN transmits a PATH message towards the MN's home address. When the HA receives the PATH message, it first checks whether the MN is located in the home network or currently resides in a visited network. Assuming that the MN is located in the range of a FA, the HA begins establishing a downlink RSVP session with the MN. This simply occurs by exchanging a PATH and a RESV message with the MN. After the HA has established the mentioned downlink RSVP session, the HA maps this session to the end-to-end RSVP session that the CN attempts to establish and responds to the CN by sending a RESV message. In this way, the RSVP session is split in the HA without breaking the end-to-end principles. The above mentioned procedure is shown in Figure 5.6.

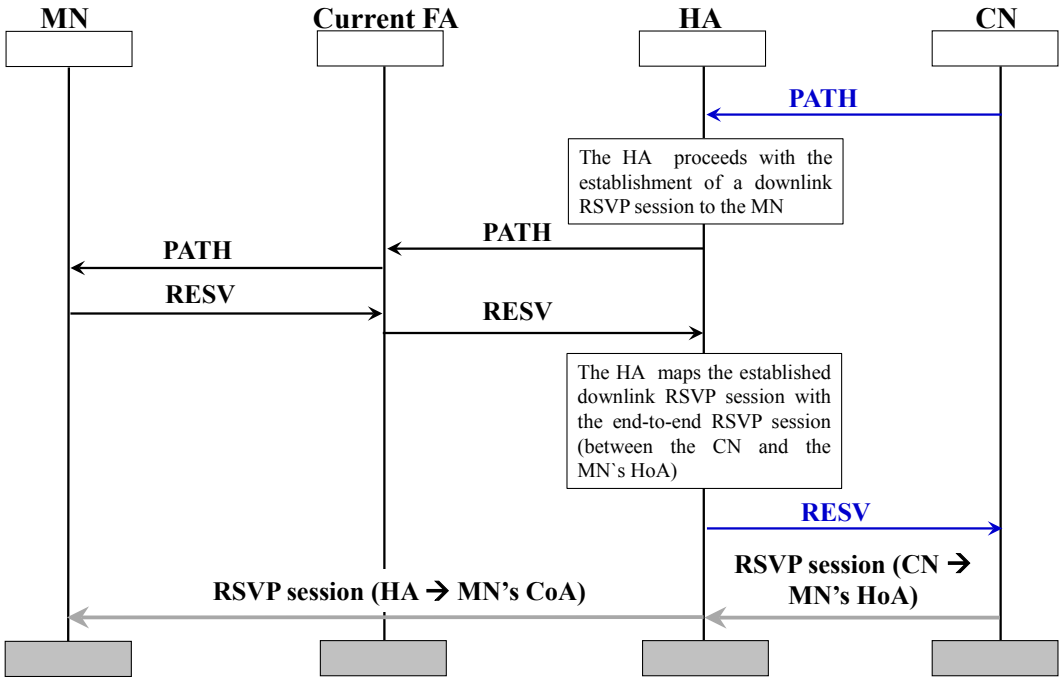


Figure 5.6: Initial reservation procedure (the MN operates as a receiver)

The MN operates as a sender: after the MN completes the MIPv4 handoff procedure, it transmits a PATH message towards the HA. To enable splitting the uplink RSVP session in the HA while taking the end-to-end uplink RSVP session into account, a new RSVP object is introduced. This object is called "**session split object**" and is used to carry the address of the CN and inform the HA that it must establish an uplink RSVP session with the CN. This object is carried with the PATH message that the MN sends to the HA, see Figure 5.7.

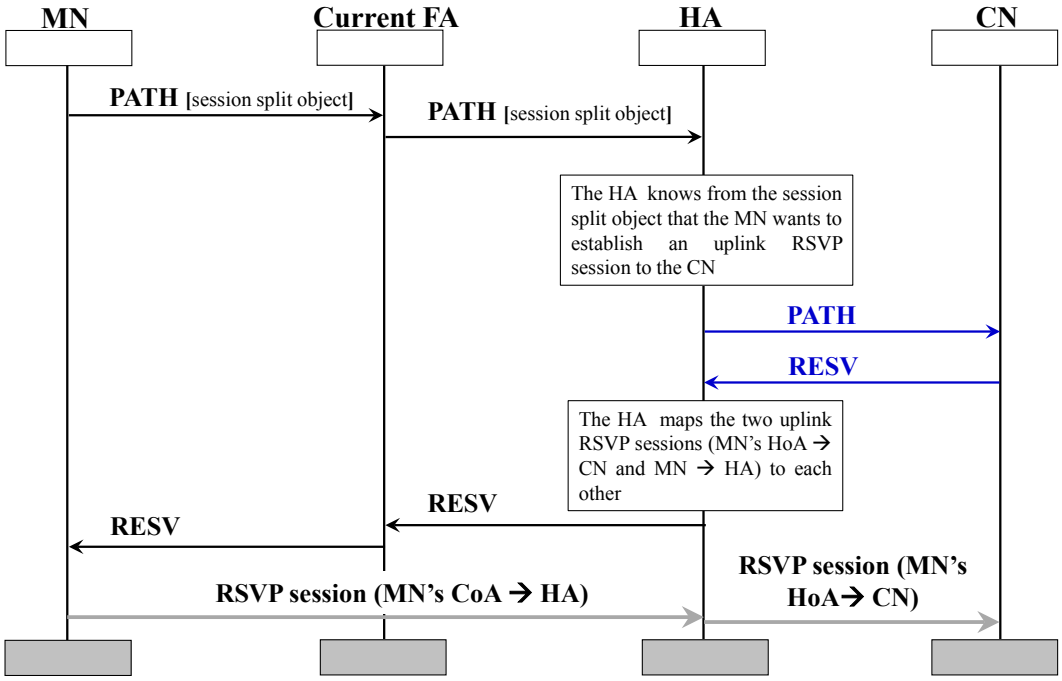


Figure 5.7: Initial reservation procedure (the MN operates as a sender)

After the HA receives the PATH message, it begins establishing an uplink RSVP session with the CN. For this purpose, the HA exchanges PATH and RESV messages with the CN. Fol-

lowing this, the HA responds to the MN by sending a RESV message. The procedure described above results in the establishment of two uplink RSVP sessions, one is between the CN and the HA, while the second is between the HA and the MN. Notice that the HA does not respond to the MN before assuring that the resources on the links to the CN have been reserved. Moreover, the PATH message that the HA sends to the CN looks like a PATH message sent from the MN's HoA (the HA uses the MN's HoA as a source address). Clearly, this aims at maintaining the end-to-end semantic.

The MN operates as a sender and a receiver: in this case, resources will be reserved on both the downlink and the uplink. For this purpose, the procedures described above are executed in parallel.

5.3.3 Semi-Proactive State Creation Procedure

After the current FA obtains all data required to authenticate and authorize the MN during the registration with the next new FA using MIFA (i.e. the MN-specific data that the current FA obtains after the initial authentication exchange procedure, see section 2.2.3), the semi-proactive state creation procedure is executed, see Figure 5.8. The goals of this procedure include the notification of the current L3-FHR's members of the incoming MN, the examination of resource availability and a prior creation of RSVP PATH states in these members. The semi-proactive state creation procedure simply implies an exchange of PATH and RESV messages between the current FA and each member of its L3-FHR. PATH messages include the MN-specific data encapsulated in mobility objects and do not require resources to be reserved. It is worth mentioning that each member of the current L3-FHR checks the availability of resources for the downlink, uplink or both. This depends on the content of the mobility objects, which must clearly indicate whether the MN acts as a receiver, sender or both. RESV messages contain mobility objects indicating the results of the procedure (e.g., states stored, resources available, resources not available, etc.).

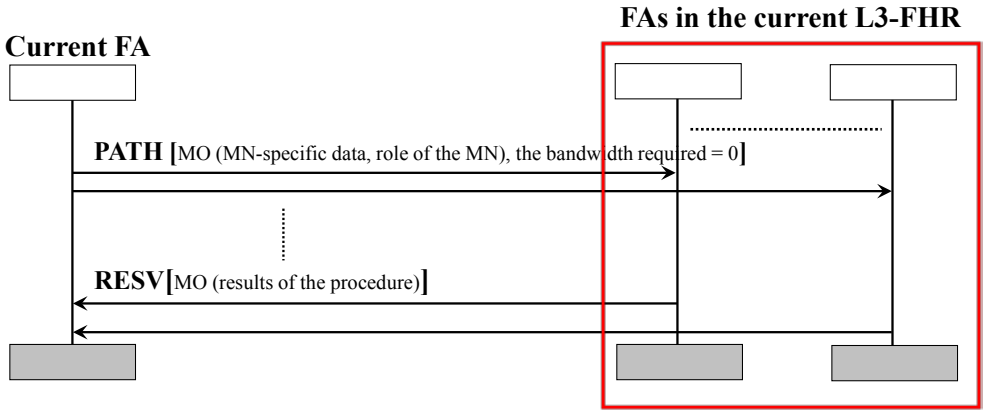


Figure 5.8: Semi-proactive state creation procedure

5.3.4 Handoff Procedure

Similar to the initial reservation procedure, the handoff procedure differs based on the role the MN carries out, i.e. receiver, sender or both.

5.3.4.1 The Mobile Node Operates as a Receiver

Movement and handoff start: once the MN moves to one of the neighboring FAs and receives an Agnt_Adv message with the **QM** flag set, the MN issues a PATH message towards the new FA. The PATH message includes a mobility object that transports the Reg_Rqst mes-

sage of MIFA inside, see Figure 5.9. Notice that because the MN acts as a receiver, no resources will be reserved for the uplink traffic.

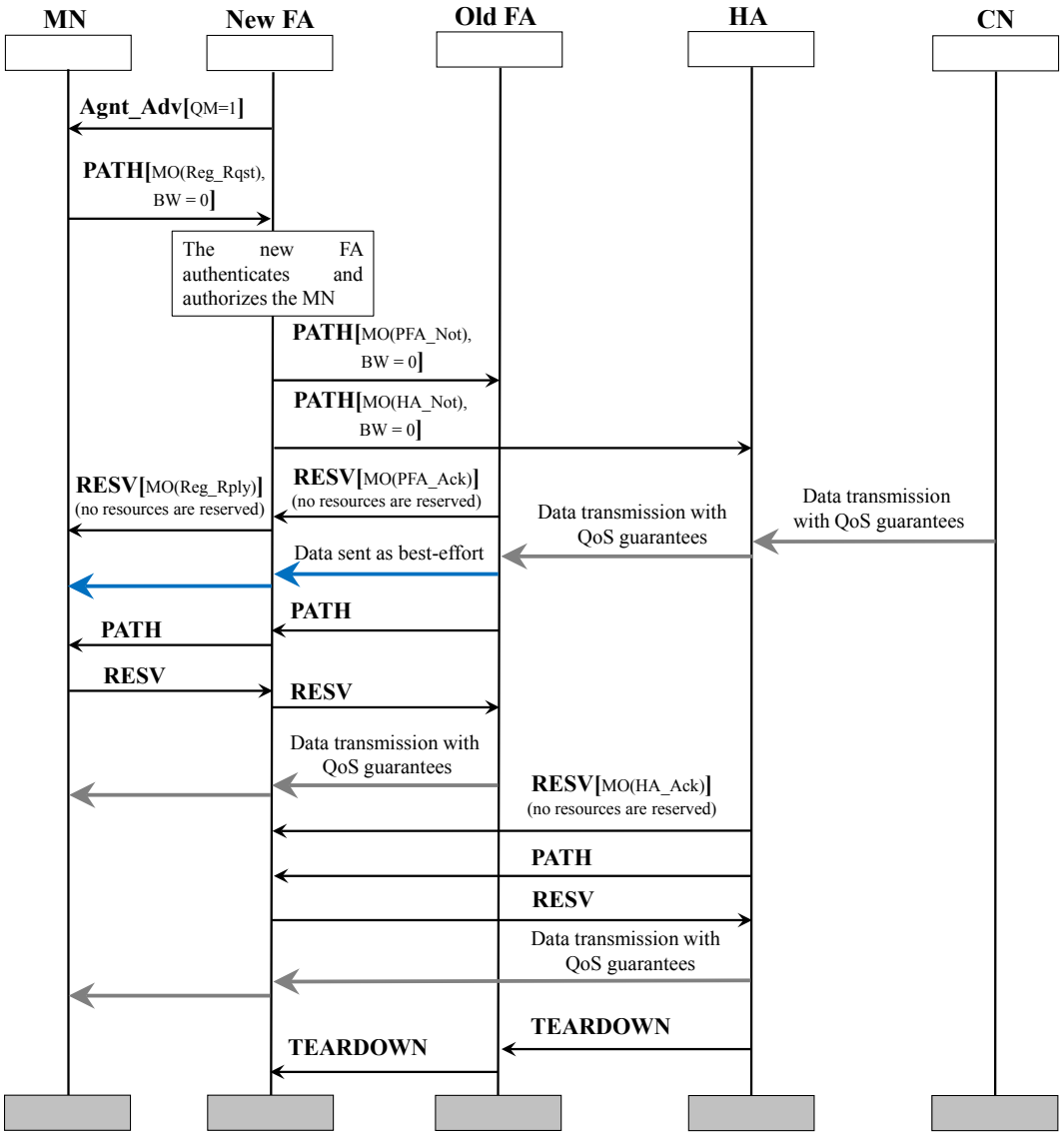


Figure 5.9: Handoff procedure employing QoMIFA (the MN operates as a receiver)

Local authentication in the new FA: the new FA in turn authenticates and authorizes the MN first. Keep in mind that the new FA received the MN-specific data during the semi-proactive state creation procedure, so it can simply authenticate and authorize the MN (more details can be found in [Dia10]).

Notifying the old FA and the HA: if the authentication and authorization of the MN are successful, the new FA sends a PATH message to the old FA with a mobility object containing the PFA_Not message of MIFA. The PFA_Not message is used to notify the old FA of the MN's new CoA. Subsequently, the new FA transmits a PATH message towards the HA. This PATH message includes a mobility object that encapsulates the HA_Not message of MIFA. As mentioned in section 2.2.3, this message is used to inform the HA about the MN's new CoA, which will result in the forwarding of data packets directly to the new CoA. Notice that both PATH messages mentioned above (the PATH message from the new to old FA and that from the new FA to the HA) do not require resources to be reserved, since no uplink traffic is issued by the MN (i.e. the bandwidth required is set to 0).

Resuming the downlink communication and completion of the handoff from the MN's point of view: when the old FA receives the PATH message from the new FA, it responds by sending a RESV message that contains the PFA_Ack message of MIFA. The PFA_Ack message is used to inform the new FA that the old FA has been successfully notified. Clearly, the RESV message does not result in reserving any resources. At this moment, the old FA begins forwarding the MN's data packets to the MN's new CoA. Notice, however, that these packets are sent as best-effort without QoS guarantees at the moment, because no RSVP downlink session has been established yet. Once the new FA receives the RESV message from the old FA, it sends a RESV message to the MN. This message contains the Reg_Rply message of MIFA conveyed in a mobility object. When the MN receives the RESV message and successfully extracts and authenticates the Reg_Rply message present in the mobility object, the handoff is declared to be completed from the MN's point of view.

Establishment of a downlink RSVP session: once the old FA is notified of the new MN's CoA, it sends a PATH message to the new FA to begin the resource reservation for the downlink traffic. Notice again that, at this moment, the old FA begins forwarding data packets to the MN's new location, as mentioned above. After the new FA receives the PATH message, it sends a PATH message to the MN, which replies with a RESV message to begin reserving resources for the downlink traffic. The new FA, in turn, sends a RESV message to the old FA. This results in a downlink RSVP session between the old and new FA as well as between the new FA and the MN. In this way, downlink data packets will then be sent using the established RSVP session. This enables the MN to receive its service with the QoS required, again without waiting for a response from the HA.

Completion of the handoff and establishment of a downlink RSVP session between the HA and the new FA: when the HA receives the PATH message with the HA_Not encapsulated in the mobility object, it responds by sending a RESV message to the new FA with a mobility object containing the HA_Ack message of MIFA. Of course, this RESV message does not reserve any resources, since there is no uplink traffic between the new FA and the HA. The HA_Ack message is used to transfer the MN-specific data required to authenticate and authorize the MN during the registration with the subsequent new FA, for more details see [Dia10]. For the downlink traffic, the HA exchanges a PATH and a RESV message with the new FA, see Figure 5.9. It is obvious that the time required to notify the HA and to establish a downlink session between the HA and the new FA is not crucial, since the MN first resumes its communication via the old FA[1].

Releasing the resources not required: after establishing the downlink RSVP session between the HA and the new FA, there is no longer a need for resources on the old path nor between the old and new FA. Therefore, these resources will be released using RSVP TEARDOWN messages.

In advance preparation for the subsequent handoff: after completing the above mentioned steps, the new FA again executes the semi-proactive state creation procedure to enable the MN to use QoMIFA during the subsequent handoff.

5.3.4.2 The Mobile Node Operates as a Sender

The handoff procedure when employing QoMIFA for the case in which the MN operates as a sender is shown in Figure 5.10.

[1] We assume, here, that the HA is far away from the new FA, while the old FA is located in the vicinity. Therefore, the old FA is notified of the MN's new CoA earlier than the HA.

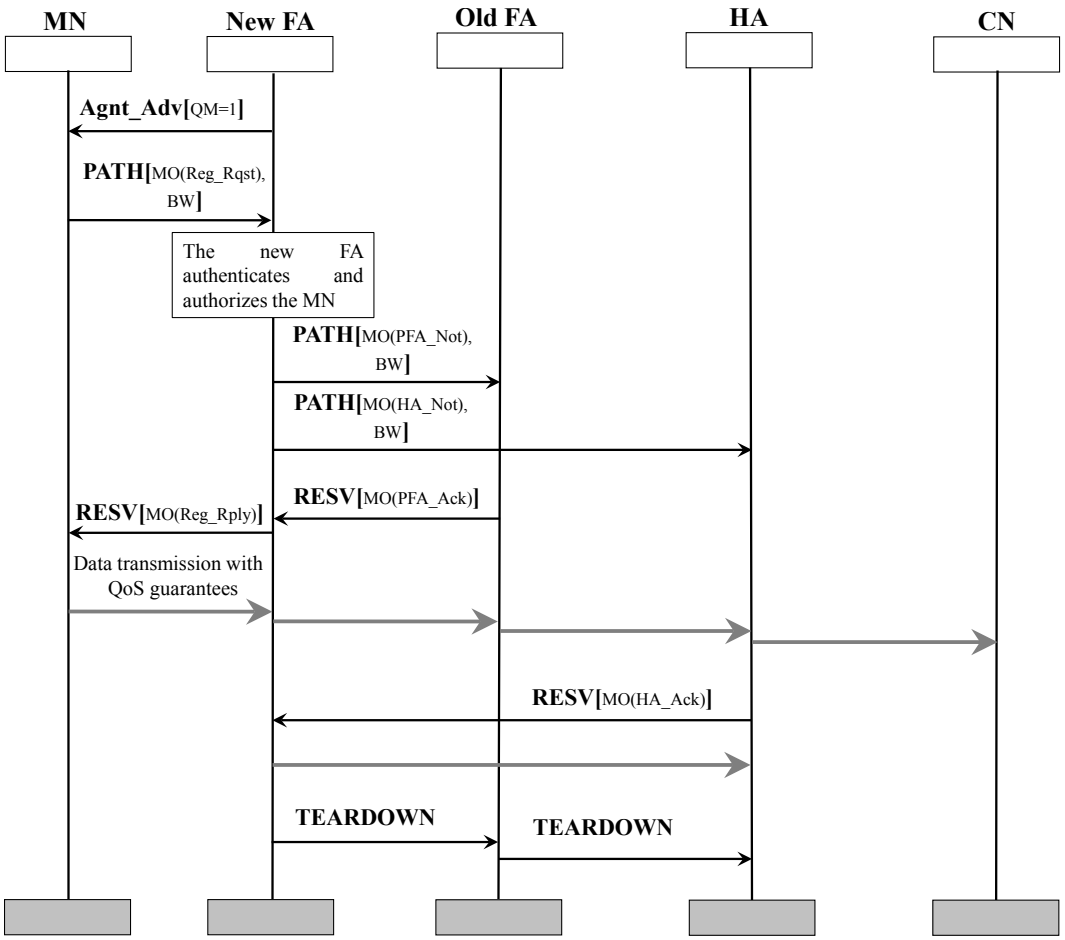


Figure 5.10: Handoff procedure employing QoMIFA (the MN operates as a sender)

Initiation of the handoff and local authentication in the new FA: again, when the MN moves into the range of a new FA and receives a new Agnt_Adv message with the **QM** flag set, the MN transmits a PATH message with a mobility object that transports the Reg_Rqst message of MIFA. When the new FA receives the PATH message and extracts the Reg_Rqst message encapsulated within, it first authenticates and authorizes the MN.

Notifying the old FA and the HA: after successful authentication and authorization of the MN, the new FA transmits a PATH message to the old FA and a PATH message to the HA. The PATH message sent to the old FA includes a mobility object containing the PFA_Not message of MIFA, while the PATH message sent to the HA contains a mobility object with the HA_Not message inside.

Establishment of an uplink RSVP session and completion of the handoff from the MN's point of view: once the old FA receives the PATH message sent by the new FA, it replies with a RESV message carrying a mobility object that encapsulates the PFA_Ack message of MIFA. The RESV message results in the reservation of the required resources on the path to the new FA. After the new FA receives the RESV message, it sends a RESV message with a mobility object that conveys the Reg_Rply message of MIFA. When the MN receives the RESV message and successfully extracts and authenticates the Reg_Rply message, it declares the handoff to be completed from its point of view. Simultaneously to the handoff, the RSVP session for the uplink traffic has been established. Thus, the MN resumes its uplink communication directly after the handoff without QoS degradations.

Handoff completion: once the HA receives the PATH message that the new FA has sent, it replies by sending a RESV message to the new FA with a mobility object conveying the HA_Ack message of MIFA. This RESV message reserves the required resources for the uplink traffic.

Releasing the resources not required: after establishing the uplink RSVP session between the HA and the new FA, the resources reserved on the old path as well as between the old and new FA are no longer necessary. Therefore, they will be released using RSVP TEARDOWN messages.

In advance preparation for the subsequent handoff: again, after completing the steps described above, the new FA executes the semi-proactive state creation procedure to enable the MN to use QoMIFA during the next handoff.

5.3.4.3 The Mobile Node Operates as a Sender and Receiver

When the MN simultaneously operates as a sender and a receiver, the handoff procedure employing QoMIFA is the same as when the MN operates only as a receiver with one exception – namely the MN reserves resources for the uplink traffic, as well.

5.4 Error Recovery Mechanisms

The errors QoMIFA handles are grouped into the following groups:

- Loss of QoMIFA support
- Control messages dropping
- Absence of the MN-specific data in the new FA
- Movement to a none-member of the current FA's L3-FHR
- Unavailability of required resources on members of the current L3-FHR

The following section discusses how QoMIFA recovers from the mentioned errors.

5.4.1 Loss of QoMIFA Support

As mentioned previously, each FA advertises its support of QoMIFA by setting the **QM** flag in its Agnt_Adv message. If the MN obtains an advertisement without the **QM** flag set, it assumes that QoMIFA is not supported. One distinguishes here between two cases, namely whether MIFA is supported or not:

- First, if MIFA is supported, the MN operates MIFA in reactive mode. Notice that the FA supporting MIFA sets a **MI** flag in the advertisement. After completing the handoff from the MN's point of view (i.e. after the MN receives the Reg_Rply message), the MN initiates an uplink RSVP session with the HA, if required. Obviously, the MN uses the standard RSVP in this case. If only downlink traffic is exchanged, the HA establishes the required RSVP session when it is notified of the handoff.
- Second, if MIFA is not supported, MIP is employed. The reservation of resources is done after handoff completion similar as in the first case.

5.4.2 Control Messages Dropping

As only RSVP messages are exchanged when employing QoMIFA, the dropping of control messages should be handled by RSVP itself. As known, RSVP relies mainly on the refresh mechanism to recover from errors resulting from the dropping of control messages. However, this is not sufficient to recover the dropping of RSVP control messages. Keep in mind that the dropping of PATH and RESV messages means a dropping of MIFA control messages, as well

– thus, disturbing the usage of mobility service. Therefore, QoMIFA addresses the reliability issue and proposes an extension to RSVP to improve it. QoMIFA extends the RSVP engine implemented in FAs, HA and MNs to include a retransmission timer for the RSVP control messages that either carry mobility objects or aim at the establishment of a new RSVP sessions. We will discuss how this mechanism works for the exchange of PATH and RESV messages between the HA and the new FA, see Figure 5.11.

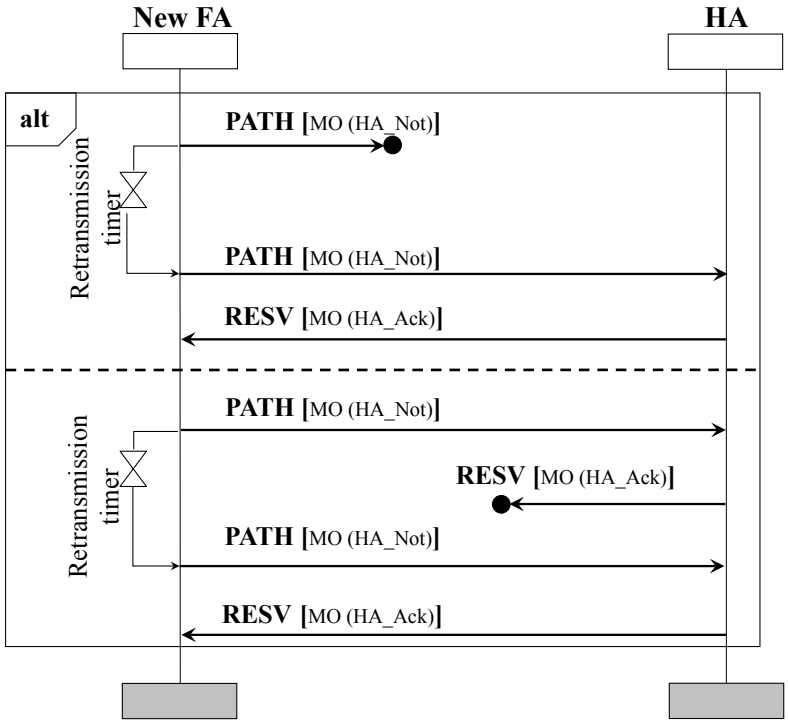


Figure 5.11: Dropping of RSVP control messages exchanged between the HA and the new FA (PATH and RESV messages carry mobility objects)

When the new FA transmits a PATH message to the HA that includes a **mobility object** or aims at the establishment of a new RSVP session, the new FA initiates a retransmission timer for this message. The value of this timer is initiated to **2*RTT**, where **RTT** is the round trip time between the new FA and the HA. The receipt of a corresponding RESV message prior to the timer expiration results in resetting this timer. Otherwise, the PATH message is retransmitted once more and the timer value doubled. The retransmission takes place a certain amount of times after the new FA declares that the HA is not reachable. Notice that the RSVP-enabled routers located on the path between the new FA and the HA treat the retransmitted PATH messages as refresh messages since RSVP does not support control messages retransmission. In addition, both the new FA and the HA handle the refresh messages based on the standard RSVP specification.

The recovery of control messages dropping between entities participating in the handoff procedure employing QoMIFA (inclusive the MN) is handled in the same way, as mentioned above.

5.4.3 Absence of the MN-Specific Data in the New FA

QoMIFA assumes that the new FA already has the MN-specific data stored in the RSVP-state for the MN. If this is not the case and the new FA receives a PATH message containing a **mobility object** with a Reg_Rqst message encapsulated inside, it first checks whether it knows the old FA (The address of the old FA is contained in the Reg_Rqst message encapsu-

lated in the mobility object) or not. If the old FA is known, the new FA assumes that the old FA could not contact it during the semi-proactive state creation procedure. This may happen, for instance, if the PATH message sent from the old FA during the semi-proactive state creation procedure had been lost due to unexpected reasons. Of course, the new FA will not be able to employ QoMIFA in this case. Therefore, it transmits the PATH message without any modification to the old FA. The old FA then authenticates the MN based on the specification of MIFA and transmits a RESV message to the new FA. The RESV message contains two mobility objects – the first conveys a Reg_Rply message, while the second includes the MN-specific data. The new FA, in turn, transmits RESV message with a mobility object containing the Reg_Rply message to the MN and proceeds with the use of QoMIFA. Figure 5.12 shows this case above.

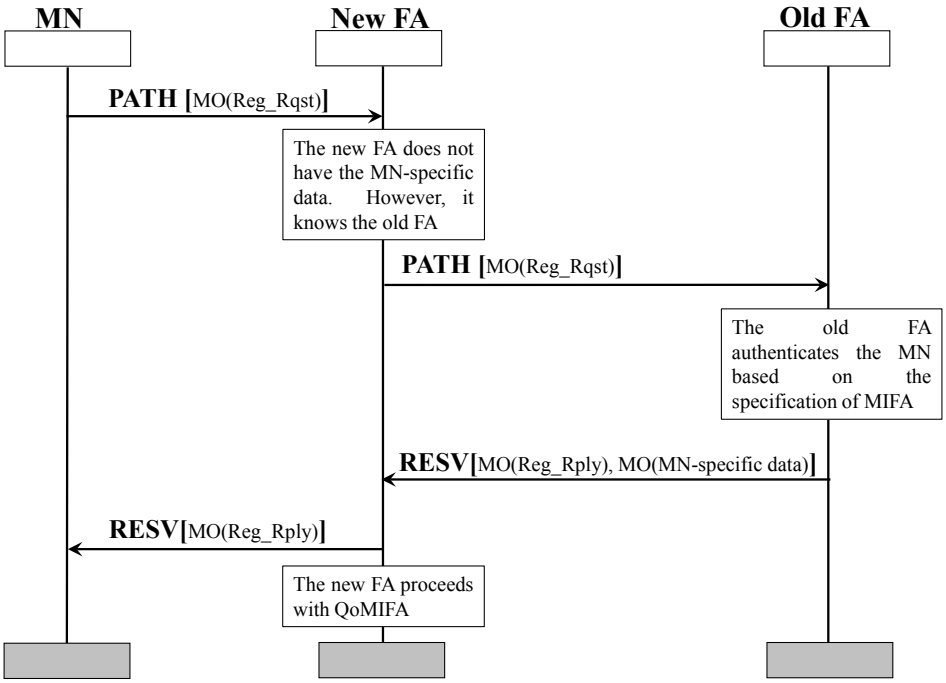


Figure 5.12: Movement to a member of the L3-FHR of the old FA. However, the new FA does not have the MN-specific data

If the new FA does not know the old FA, it behaves the same as in the case where the MN moves to a non-member of the current FA's L3-FHR. This is explained in the following section

5.4.4 Movement to a None-Member of the Current FA's L3-FHR

As often mentioned that the MN relies on the L3-FHRs in its operation. However, it is possible for the MN to move into the range of a none-member of the L3-FHR of the old FA[1]. In this case, the MN will starts operating QoMIFA after the receipt of the Agnt_Adv message which, as noted before, includes the **QM** flag set. The new FA detects that it neither knows the old FA nor has the MN-specific data that enable pushing QoMIFA forth. Therefore, the new FA replies to the MN by sending a PATHErr message with a mobility object containing a Reg_Rply message. The Reg_Rply indicates that the registration is denied and the error code will be "**MN and old FA not known**". This case is shown in Figure 5.13.

[1] This may occur in many cases. For instance, the FA is newly integrated in the network, the construction of the L3-FHR was not completed, the new FA could not be detected by the old FA because no MNs have moved there before, etc. In any case, this situation will be temporal and the new FA will be integrated in the L3-FHR of the old FA after the new FA is detected.

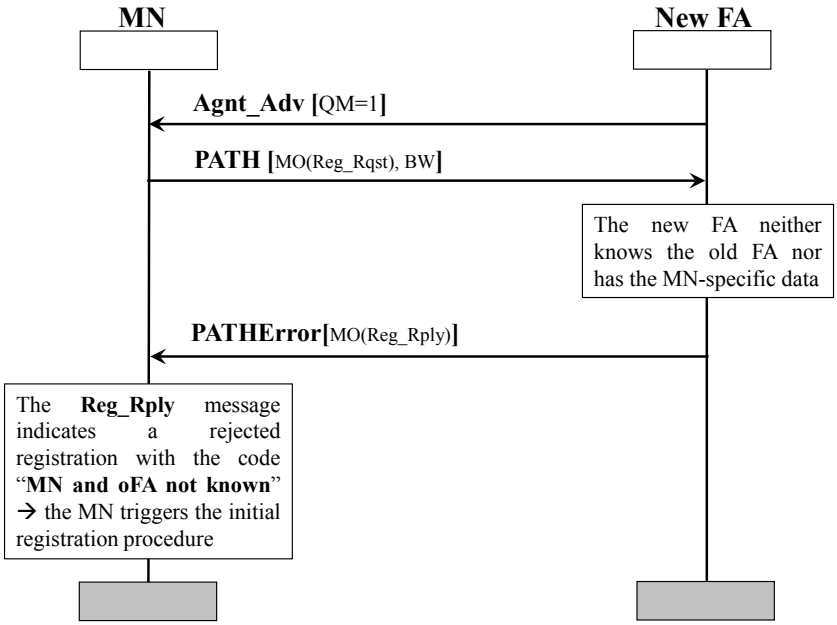


Figure 5.13: Movement to a none-member of the L3-FHR of the old FA

When the MN receives the PATHErr message and detects the error that has occurred, it proceeds with the initial registration followed by the initial reservation procedure. Of course, this will disrupt communication. However, the goal in this case is to recover the communication capability rather than achieve a seamless and fast handoff with QoS guarantees. The new FA proceeds with the initial authentication exchange and semi-proactive state creation procedures to enable operating QoMIFA in the subsequent handoff. In addition, the new FA tries to attend the L3-FHR of the old FA. This is done based on the standard specification of MIFA by exchanging a Member Join Request (*Mem_Join_Rqst*) and a Member Join Response (*Mem_Join_Resp*) message with the old FA after establishing a trust between the new and old FA, as shown in Figure 5.14. For details about how a trust can be established between the new and old FA, the reader is directed to [Dia10]

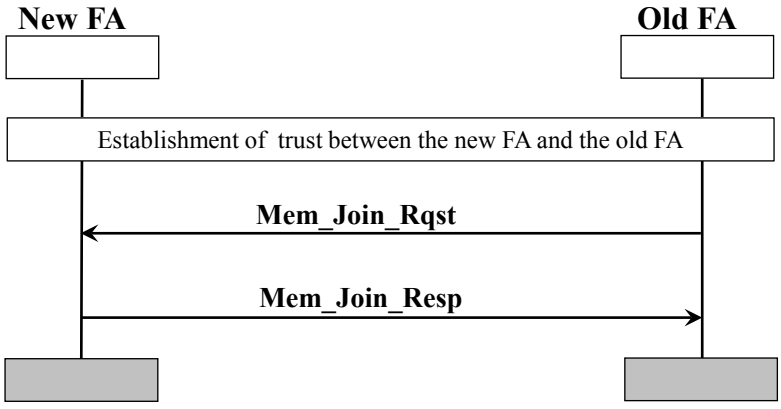


Figure 5.14: The new FA joins the L3-FHR of the old FA

5.4.5 Unavailability of Required Resources on Members of the Current L3-FHR

The major issue that affects QoS guarantees for MNs besides the fast reservation of resources in the new FA is the availability of these resources. This is the main task of the semi-proactive

state creation procedure since each member of the current L3-FHR checks whether the resources required for the MN are available and reports the results of this check to the current FA. This report shows the grade to which the resources can be provided to the MN. For instance, a grade of 100 % means that the resources are available, a grade of 80 % means that only 80 % of the resources can be offered to the MN, and so on. If the resources are available on all members of the current L3-FHR, everything is in order and nothing should be carried out. However, if some neighbors are not or only partially capable of offering the resources required, the current FA notifies the MN by sending a **resource check** message to the MN. This message contains information about the FAs that will not or only partially satisfy the MN's requirements. It is assumed that the MN implements some applications that enable it to adapt the applications running to cope with resources that are temporarily not available, e.g. reducing the throughput, freezing some applications, etc. Therefore, when the MN moves to one of the FAs listed in the **resource check** message, the MN should adapt its applications' behavior based on the resources available.

5.5 Conclusion

The chapter has described our solution (QoMIFA) for providing simultaneous support of mobility and QoS. QoMIFA is a semi-proactive solution that couples in a hybrid manner between RSVP as a protocol for resources reservation and MIFA as a mobility management protocol.

QoMIFA depends on the L3-FHRs MIFA employs. As known, each L3-FHR contains the neighbor FAs the MN may move to from the current FA. The principle of L3-FHRs enables the prediction of MNs' movements and the in-advance distribution of MNs-specific data, which enables a fast re-authentication of MNs after handoffs. Moreover, QoMIFA follows a semi-proactive behavior since the current FA notifies the members of its L3-FHR of incoming MNs (through the distribution of MNs-specific data). Therefore, these members check in advance if they have resources available and create, in case of success, RSVP states. The reservation of resources takes place, however, after handoffs.

The hybrid structure is obtained via introducing a new RSVP-object called "mobility object". This object encapsulates MIFA control messages so that the MN operates only RSVP after the handoff and transmits MIFA control messages encapsulated inside the new introduced RSVP-object. In this way, both protocols are implemented separately. However, they operate as one protocol.

QoMIFA is robust since it provides mechanisms to recover from most failures that may happen like, loss of QoMIFA support, control messages dropping, absence of MNs-specific data, movement to a none-member of the current L3-FHR, etc.

Chapter 6: Performance Evaluation

In order to evaluate QoMIFA compared to Simple QoS, both are modeled in NS2 and simulated deploying the same network topology under the same assumptions. As mentioned previously, Simple QoS is a well-known loose-coupled protocol and, as currently known, loose-coupled solutions are widely employed. Furthermore, the qualitative comparison in section 4.6 showed that this protocol presents one of the best loose-coupled protocols and is used as basis for comparison in related work. For these reasons it was selected as a candidate to compare with QoMIFA. Note that although there are other solutions in the literature that make performance improvements compared to Simple QoS, we have decided to select Simple QoS because this protocol is better when we consider other metrics like no dependency on layer 2 triggers, no restrictions on network topology, no need for passive reservations, etc. No hard-coupled solutions are evaluated in this study because they are not widely employed due to their complexity as well as their less applicability to future networks.

NS2 provides the basic Mobile IP functionalities which were extended to model MIFA in previous works, see [Dia10]. The basic RSVP implementation is developed at the University of Bonn (implementation details can be found in [Gre99]). The basic RSVP implementation was extended in the scope of this work to operate on wireless links. Following that, the extended RSVP implementation was integrated with MIFA to create QoMIFA protocol. To model Simple QoS, the extended RSVP implementation was integrated with MIP functionalities the standard NS2 provides, see [Fre09] for implementation details.

In the following, a brief overview of NS2 is introduced in sections 6.1, the applied simulation scenario is described and the performance measures of interest are presented in sections 6.2 and 6.3, respectively. Thereafter, the obtained results are discussed. Concretely, the chapter investigates the impact of network load in section 6.4 and the impact of MN speed in section 6.5. The main results of this chapter are concluded in section 6.6.

6.1 Network Simulator (NS2)

The simulator is a well-known open-source project that aims at the development of event-oriented software to be used to model, test and evaluate wired as well as wireless networks[1]. NS2 uses two programing languages, namely C++ and Object-oriented Tool control language (OTcl). The reason why two languages are used lies in the tasks NS2 has to achieve. More concrete, protocols, traffic applications, etc. require fast execution time and are rarely changed after building NS2. Thus, a language like C++ is necessary. On the contrary, tasks like the definition of network topologies, parameters, etc. are often changed during the simulation. Therefore, an interpreted language such as OTcl is necessary. Thus, the simulator provides two classes hierarchy, compiled and an interpreted classes hierarchy. Both hierarchies interact and exchange data with each other.

For analysis purposes, it is possible to write information about specific events in specific log files. These files are processed and examined manually or via tools, e.g. visualized using the Network animator Nam [Nam] or presented using the XGraph tool [XGra].

[1] The upcoming version of NS2 is called NS3 and still now under development. It is even in the initial phase. Because many protocols and features are not implemented yet, we have decided to use NS2 in our simulation studies.

6.2 Simulation Assumptions

The simulative evaluation was achieved deploying a hierarchical topology[1], see Figure 6.1. The figure depicts a domain of 4 sub-domains, each has the same structure. Each sub-domain includes 4 FAs. The assumed distance between the cells' centers of two neighbor FAs is 198 m. The coverage area of each FA is assumed as a circle with a radius of 112 m. A gateway (GW) is placed on the uppermost level of the domain hierarchy. This GW interconnects the domain with other nodes located outside the domain. The distance between the GW and each FA in the domain is 4 hops. There are 160 MNs in the domain, 10 equally distributed MNs in the coverage area of each FA. The MNs connect to the network via IEEE 802.11 wireless links with a bandwidth of 5.5 Mbps. The access points applied in the simulation fit the Lucent WaveLANs [Lucent97].

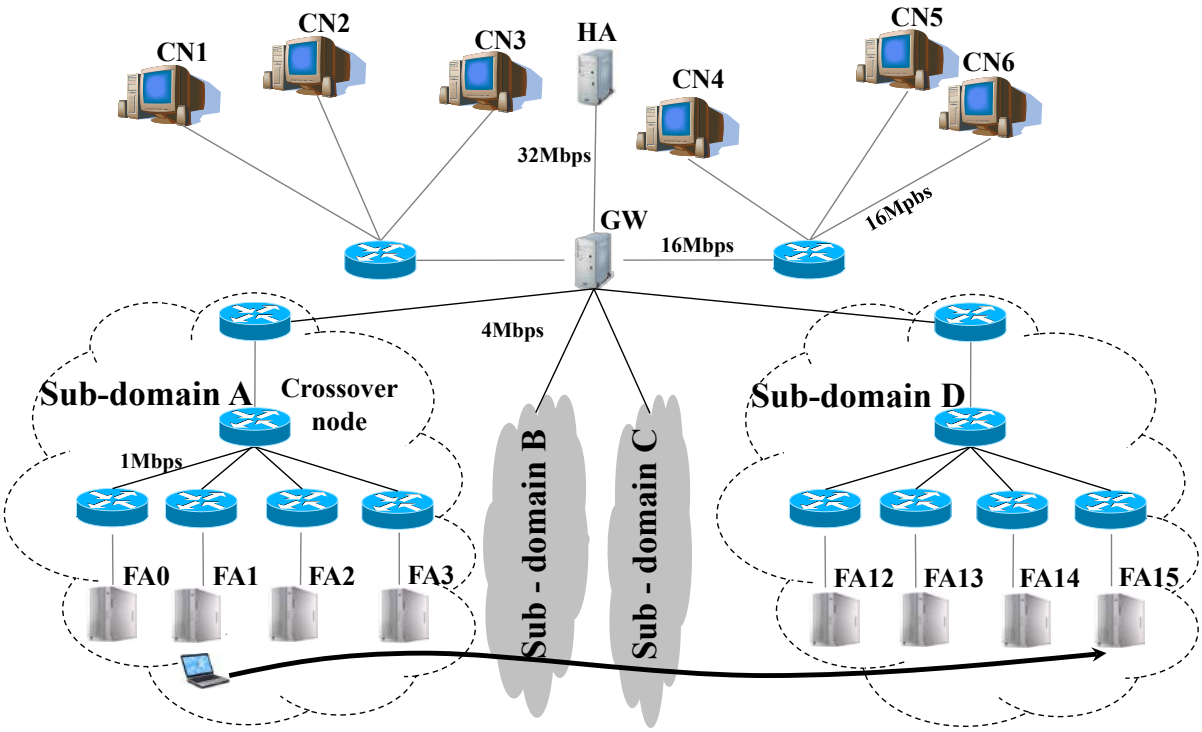


Figure 6.1: Assumed scenario

All MNs are registered with the same HA, which is located outside the domain. Active MNs communicate with 6 CNs, which are also located outside the domain. The transmission delay on each wired ethernet link inside the domain is 5 msec. Each link connecting the crossover node in each sub-domain with the FA beneath it in the sub-domain has a bandwidth of 1 Mbps, while each of the links connecting the crossover node with the GW has a bandwidth of 4 Mbps[2]. The link between the HA and the GW has a transmission delay of 150 msec and a

[1] The reason why we selected a hierarchical topology is motivated from the studies done in [Dia10]. These studies showed that a mesh topology is better adequate for both MIP (used in Simple QoS) and MIFA (applied in QoMIFA) than a hierarchical one. The hierarchical topology strongly affects the performance of MIFA, while the performance of MIP is slightly affected. Therefore, we have selected the hierarchical topology to study the performance of QoMIFA compared to Simple QoS in a topology more adequate to Simple QoS than to QoMIFA. This makes the results more meaningful. Furthermore, the topology used in our studies is similar to the topology used to evaluate MIFA in [Dia10] compared to a range of macro- and micro-mobility management protocols, where MIFA shows significant performance improvements.

[2] Note that each crossover node serves 4 routers beneath it, each is connected to a link with a bandwidth of 1 Mbps.

bandwidth of 32 Mbps[1]. The links between the GW and CN1, CN2, CN3, CN4, CN5 and CN6 have delays of 27, 23, 28, 27, 23 and 28 msec, respectively. Each of these links has a bandwidth of 16 Mbps. During the simulation, we tracked the selected active MN(s). The selected MN(s) move(s) from FA0 to FA15. The tracked MN uses Eager Cell Switching (ECS)[2] algorithm to trigger handoffs. For more clarity, the simulation assumptions are summarized in Table 6-1.

Scenario	
Number of FAs	16 FAs
Distance between neighbor FAs	198 m
Radius of the coverage area of each FA	112 m
MNs in the scenario	160 MNs (10 equally distributed in the range of each FA)
Wireless links	IEEE 802.11 (5.5 Mbps) The Access points applied in the simulation fit the 914 MHz Lucent WaveLANs
Links bandwidth	
Crossover node in each sub-domain ←→ FA beneath it	1 Mbps
Crossover node in each sub-domain ←→ GW	4 Mbps
GW ←→HA	32 Mbps
GW ←→ CN1, CN2, CN3, CN4, CN5 and CN6	16 Mbps
Links delay	
Crossover node in each sub-domain ←→ FA beneath it	5 msec
Crossover node in each sub-domain ←→ GW	5 msec
GW ←→ HA	150 msec
GW ←→ CN1, CN2, CN3, CN4, CN5 and CN6	27, 23, 28, 27, 23 and 28 msec

Table 6-1: Simulation assumptions

The main goal of our analysis is to investigate how network load and MN speed affect the performance of the studied protocols in the selected scenario. To study the impact of network load, the number of active MNs in the range of each FA is changed between 4, 6 and 8. An active MN is tracked while moving at a speed of 36 km/h from the first to last FA in each studied scenario. Notice that each active MN reserves resources in the network and exchanges a constant bit rate UDP stream with its communication partner on uplink and downlink. Each UDP stream has a packet arrival rate of 10 packets per second and a packet size of 1000 bytes.

[1] Note that the GW interconnects 4 crossover nodes (each with a link of 4 Mbps → 16 Mbps) from the domain with the CNs and the HA. Because the traffic from CNs should bypass the HA, the link between the GW and the HA is made 32 Mbps.

[2] A movement detection algorithm. It assumes frequent location changes and, thus, initiates a handoff direct after the receipt of a new Agnt_Adv message [Dia10].

In addition to the traffic of active MNs, an uplink and a downlink constant bit rate UDP best-effort traffic is exchanged between a CN and a MN (different from the active MNs) located in the range of each FA. This traffic represents background traffic only and, clearly, does not reserve resources. The packet arrival rate of each stream of this best-effort traffic is also 10 packets per second and the packet size is 1000 bytes. Remaining MNs that are neither active nor exchange best-effort traffic are considered idle and only generate signaling traffic.

Considering the loads mentioned above, one derives that the links inside the sub-domains are loaded to approximately 42 % when each FA hosts 4 active MNs in addition to the MN that exchanges best-effort traffic. When each FA serves 6 and 8 active MNs in addition to the MN that exchanges best-effort traffic, wired links in each sub-domain are loaded to 58 % and 74 %, respectively. Sure, the tracked MN generates extra load in the networks. Note also that we excluded the signaling traffic resulting from all MNs, which is not negligible. As a result, the loads mentioned express low-, middle- and high-loaded network.

To investigate the impact of MN speed on the performance of QoMIFA and Simple QoS, similar assumptions as those mentioned above are considered. However, each FA serves 6 active MNs, one MN with best-effort traffic and 3 idle MNs. It is worth noting that the observed MN moves from the first to last FA with a speed alternating between that of a pedestrian (3 km/h), a cyclist (20 km/h), a car driver inside a city (50 km/h) and a car driver on an autobahn (120 km/h).

In order to ensure reliable simulation results, several measurements were conducted. That is, each scenario was repeated 20 times, resulting in 300 handoffs for each measurement. Finally, it is worth mentioning that we simulate handovers only. We assume that L3-FHRs are truly constructed and the MN-specific data are successfully distributed in advance. Thus, the MN always moves to a member of the current L3-FHR. Recovering from control messages dropping is included in the simulation.

6.3 Performance Measures

The performance measures we are interested in comprise the following:

1. Resource reservation latency: this metric includes the handoff latency[1] employing the used mobility management protocol and the time required to complete the reservation of resources. This metric will be studied for uplink and downlink sessions separately.
2. Number of dropped packets per handoff: the packets dropped per handoff on downlink are the packets getting lost between the time after which the MN loses the link with the old FA and the time at which the MN begins receiving packets on downlink. The packets that get lost per handoff on uplink are those the MN sends after losing the link with the old FA and before the handoff is completed employing the mobility management protocol used. The simulation studies conducted do not include buffering of data packets due to handoffs neither on downlink nor on uplink.
3. Number of packets sent per handoff as best-effort until the reservation is accomplished: best-effort packets sent per handoff on downlink are those transmitted after the HA (for Simple QoS) or the old FA (for QoMIFA) is notified of the MN's new CoA and before resource reservation for downlink traffic is completed. On uplink, best-effort packets are sent between the time after which the MN completes the handoff and the time at which the MN finishes the reservation of resources for the uplink session. The number of packets sent as best-effort until resources are reserved is a very important performance measure since this parameter determines how long the us-

[1] The handoff latency includes a movement detection time, which represents the duration between the time after which the MN loses the link with the old FA and the time at which the MN discovers the new FA's IP address.

er obtains his service without QoS guarantee. One may argue that best-effort packets are sent during a short time duration, which makes this metric not that important. Although this is true in general, best-effort packets will result in a considerable QoS degradation if their amount exceeds certain thresholds. Therefore, it is important to study this metric.

4. Probability of dropping sessions: this metric represents the likelihood that RSVP-sessions reserved for active and moving MNs are dropped due to handoffs. To study this performance measure, the number of active and also observed MNs is varied between 4 and 8 MNs in a low network load scenario (4 active MNs in the range of each FA). It should be noted that these extra observed MNs are added to the previous scenario which originally contains 160 MNs, as written earlier.

For clarity, the performance measures of interest are illustrated in Figure 6.2.

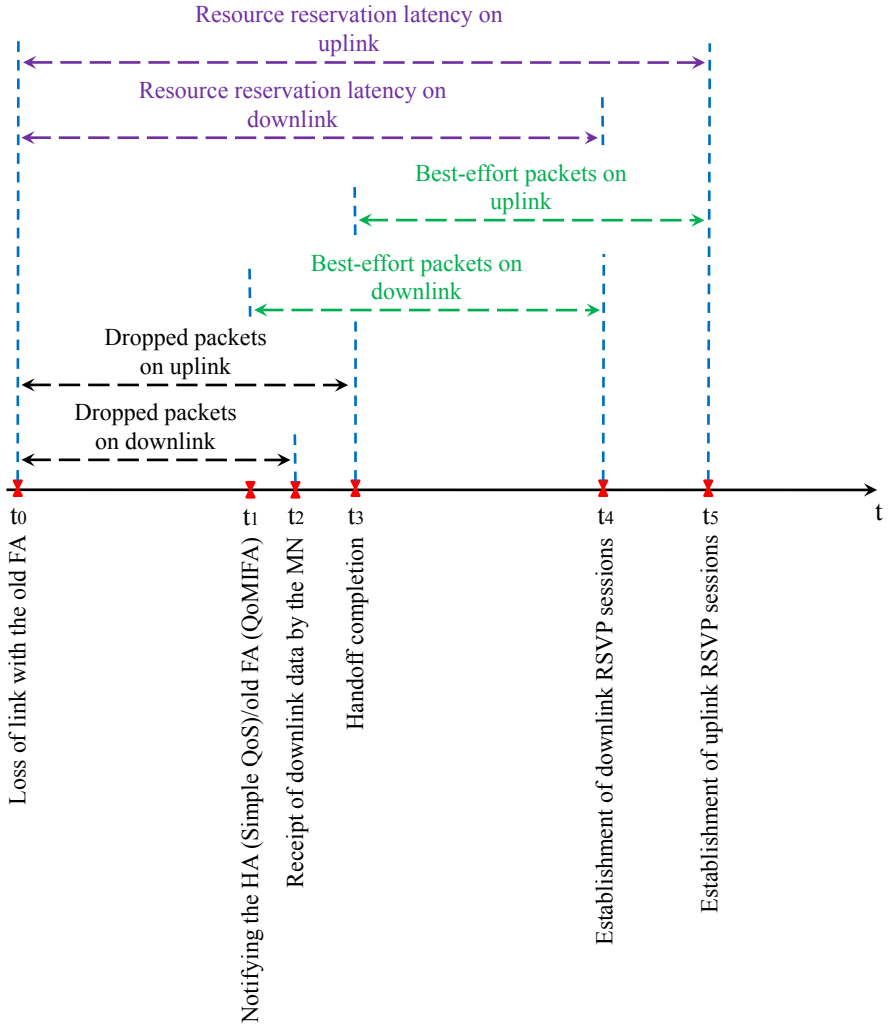


Figure 6.2: Performance measures of interest

6.4 Impact of Network Load

6.4.1 Resource Reservation Latency

6.4.1.1 Uplink

Figure 6.3 displays the distribution function of the resource reservation latency on uplink when employing both protocols in the studied scenario under the above mentioned loads.

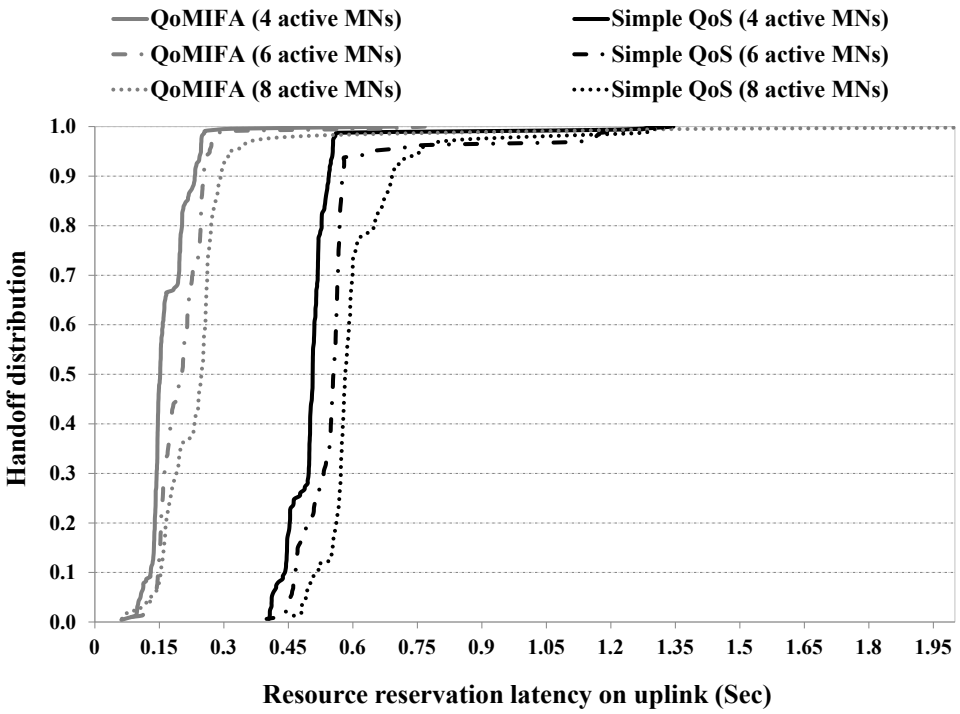


Figure 6.3: Resource reservation latency on uplink with QoMIFA and Simple QoS under different network loads

This figure shows that the time required to reserve resources on uplink employing both QoMIFA and Simple QoS follows, in general, similar behavior under the studied loads. Concerning performance, QoMIFA is significantly better than Simple QoS. The reason lies in the way QoMIFA combines mobility and RSVP control messages. This coupling results in simultaneously executing handoffs and reserving resources. This is not the case for Simple QoS, which completes the handoff first and then reserves resources.

Let us first consider the situation in which each FA serves 4 active MNs. Our simulation results show that while resources on uplink are reserved in less than 232 msec in 90 % of handoffs employing QoMIFA, the reservations are accomplished in less than 544 msec when employing Simple QoS for the same number of handoffs. The reason is mentioned above and the result is expected since QoMIFA manages mobility and QoS simultaneously, while Simple QoS handles mobility first and QoS after that. Based on that, Simple QoS will take the double of the time QoMIFA consumes and even longer, which is proven by our simulation results. The difference in the performance of both protocols when the number of active MNs in each cell goes up to 6 remains comparable to that when each FA hosts 4 active MNs. Increasing the number of active MNs hosted by each FA to 8 results in a clear deterioration in performance for Simple QoS. Figure 6.4 shows that QoMIFA requires no more than 294 msec to reserve resources on uplink for 90 % of the handoffs. Simple QoS, however, reserves resources on uplink in less than 693 msec for the same number of handoffs. Again, the reason is mentioned above.

Similar results are displayed in Figure 6.4, which shows the average latency required to reserve resources on uplink when employing both studied protocols under the mentioned loads.

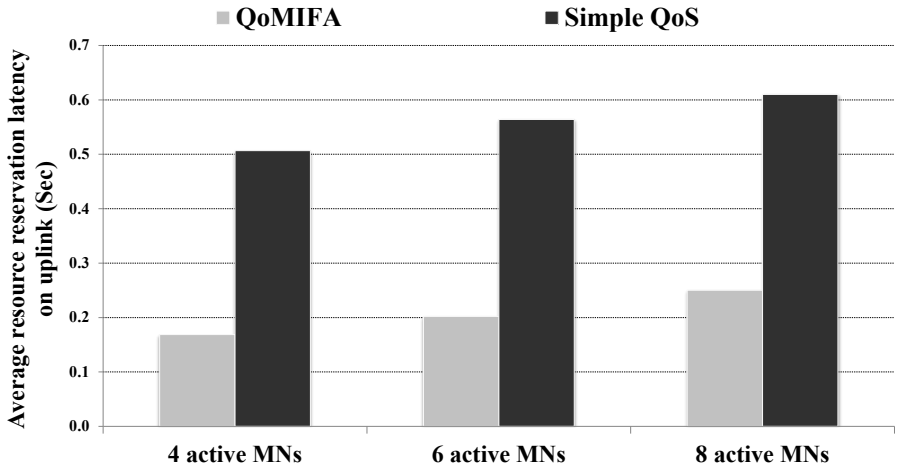


Figure 6.4: Average resource reservation latency on uplink when employing QoMIFA and Simple QoS under different network loads

The simulation results show that QoMIFA is 67 %, 64 % and 59 % better than Simple QoS when each FA hosts 4, 6 and 8 active MNs, respectively.

6.4.1.2 Downlink

Let us now study the time required to reserve resources for the downlink traffic, see Figure 6.5.

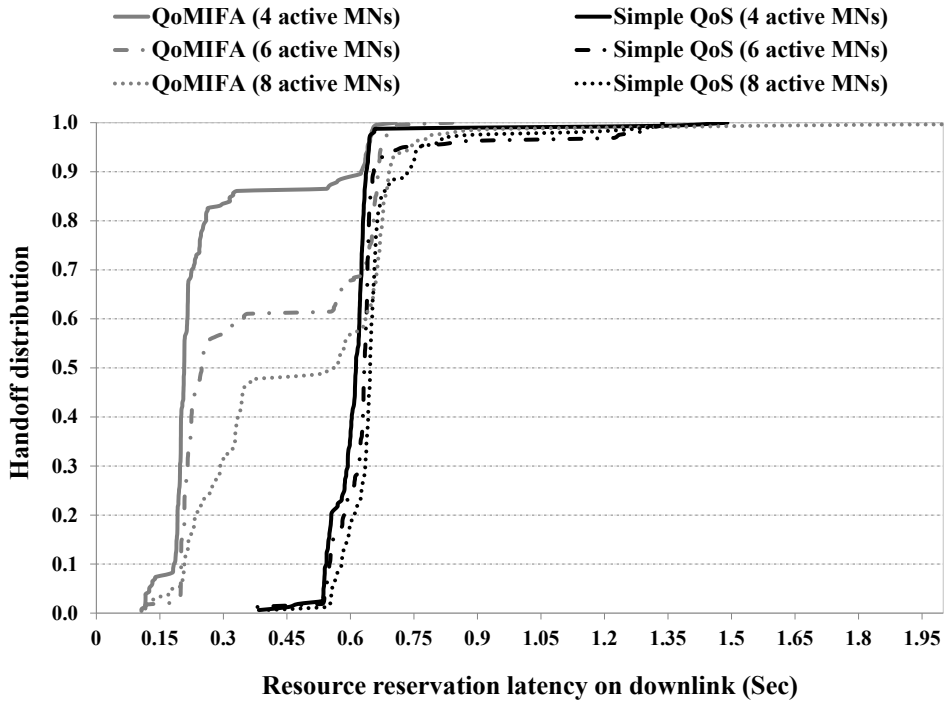


Figure 6.5: Reservation latency on downlink with QoMIFA and Simple QoS under different network loads

The first result that can be obtained from this figure is that QoMIFA is significantly faster than Simple QoS with respect to the time required to reserve resources on downlink. This is because QoMIFA requires only contacting the old FA, which reserves a bidirectional RSVP

session between the old CoA and the new one. In contrast, Simple QoS must first contact the HA and then establish an RSVP tunnel between the HA and the new FA. It can also be observed that the performance of QoMIFA with respect to the time required to reserve resources on downlink follows a similar pattern as its performance in terms of the time required to reserve resources for uplink traffic in about 30 % of the handoffs. In the remaining 70 % of the handoffs, one notices that QoMIFA is on downlink more affected by the load than on uplink. Of course, this is because, uplink sessions are first established between the old and new FA. After that, downlink sessions are built. The time required to reserve resources for Simple QoS follows similar behavior as that QoMIFA follows. The reason for this is that the PATH message sent to the HA from the MN operating Simple QoS is intercepted by the crossover router located on both the path between the HA and old FA and the path between the HA and the new FA. The crossover router answers directly by sending a RESV message. Notice that there is no reverse tunnel used between the new FA and the HA for uplink traffic, see [TSZ99]. On the contrary, resource reservation on downlink when operating Simple QoS is triggered before starting the reservation for uplink (as the HA receives the Reg_Rqst message). The reservation for uplink starts even before the completion of that for downlink (as the MN receives the Reg_Rply message).

Similar results to those shown in Figure 6.4 can be observed in Figure 6.6, which presents the average time required to reserve resources on downlink employing QoMIFA and Simple QoS under the loads mentioned in section 6.3.

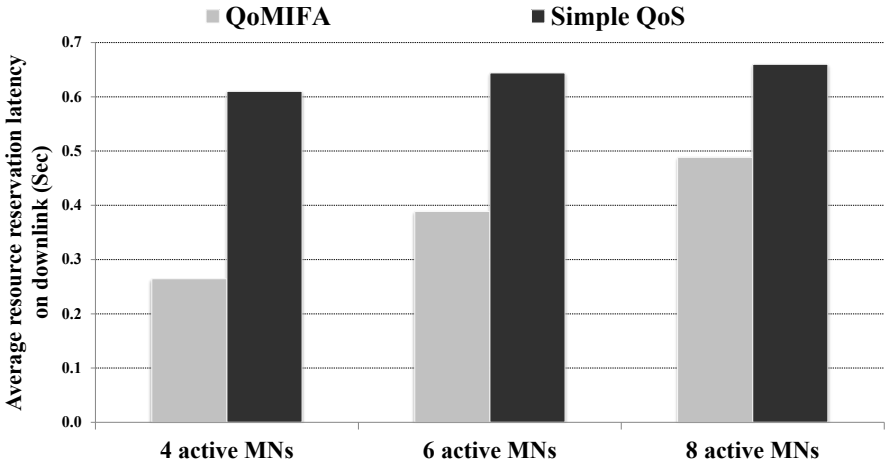


Figure 6.6: Average resource reservation latency on downlink when employing QoMIFA and Simple QoS under different network loads

The figure shows a clear performance improvement when employing QoMIFA compared to Simple QoS. The reason for this was mentioned above while discussing Figure 6.6. According to the simulation results, QoMIFA outperforms Simple QoS by 57 %, 40 % and 26 % when the number of active MNs in the range of each FA varies between 4, 6 and 8, respectively.

In addition to the results discussed above, one notes that the resource reservation latency on downlink is longer than that on uplink when employing QoMIFA. This is not the case, however, for Simple QoS. The reasons are discussed above.

6.4.2 Number of Dropped Packets per Handoff

6.4.2.1 Uplink

Figure 6.7 shows the distribution function of the number of dropped packets per handoff on uplink when employing QoMIFA and Simple QoS in the studied topology under the mentioned loads. The figure shows that QoMIFA results in significant performance improvements. The figure also depicts approximately similar behavior as that observed in Figure 6.2. Our simulation results show that QoMIFA and Simple QoS drop no more than 6 and 7 packets per handoff in 99 % of the handoffs in networks containing 64 (4 in the range of each FA) and 96 (6 in the range of each FA) active MNs, respectively. Increasing the number of active MNs in the network to 128 strongly deteriorates the performance of both QoMIFA and Simple QoS.

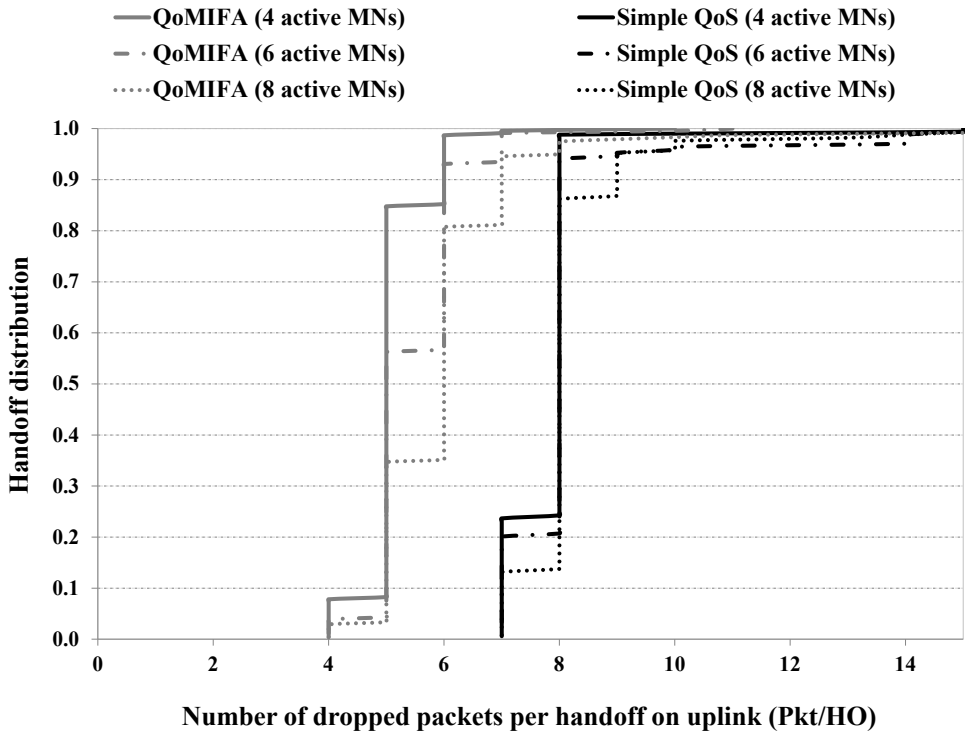


Figure 6.7: Number of dropped packets per handoff on uplink with QoMIFA and Simple QoS under different network loads

Let us now consider the average number of dropped packets per handoff for both protocols, see Figure 6.8.

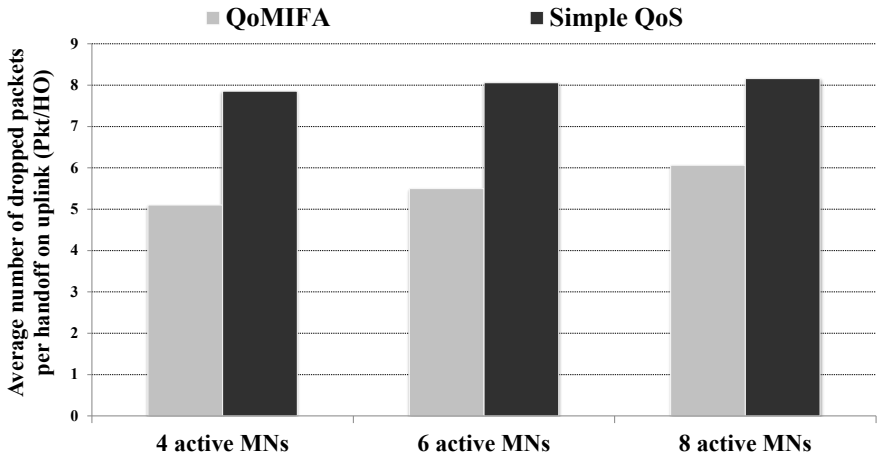


Figure 6.8: Average number of dropped packets per handoff on uplink when employing QoMIFA and Simple QoS under different network loads

The results show that QoMIFA drops 35 %, 32 % and 26 % fewer packets than Simple QoS when the number of active MNs in the range of each FA is 4, 6 and 8, respectively. This is because QoMIFA executes registration and reservation simultaneously. However, Simple QoS waits until the HA is notified and reserves the resources following that.

6.4.2.2 Downlink

The distribution function of the number of dropped packets per handoff on downlink when employing both studied protocols in the deployed scenario under the mentioned loads is shown in Figure 6.9. Let us first discuss a network that serves 64 active MNs (4 active MNs in the coverage area of each FA). One can see that while no more than 1 packet gets lost in 86 % of the handoffs when using QoMIFA, Simple QoS may drop till 4 packets in 99 % of the handoffs. Increasing the number of active MNs in the range of each FA to 6 does not result in a considerable change in the behavior of both protocols, see the figure. As the load further increases, the performance of both protocols becomes more and more closed to each other. In this case, both QoMIFA and Simple QoS suffer from significant packet loss. However, QoMIFA still performs better. Note that the figure also shows a comparable behavior to that displayed in Figure 6.5.

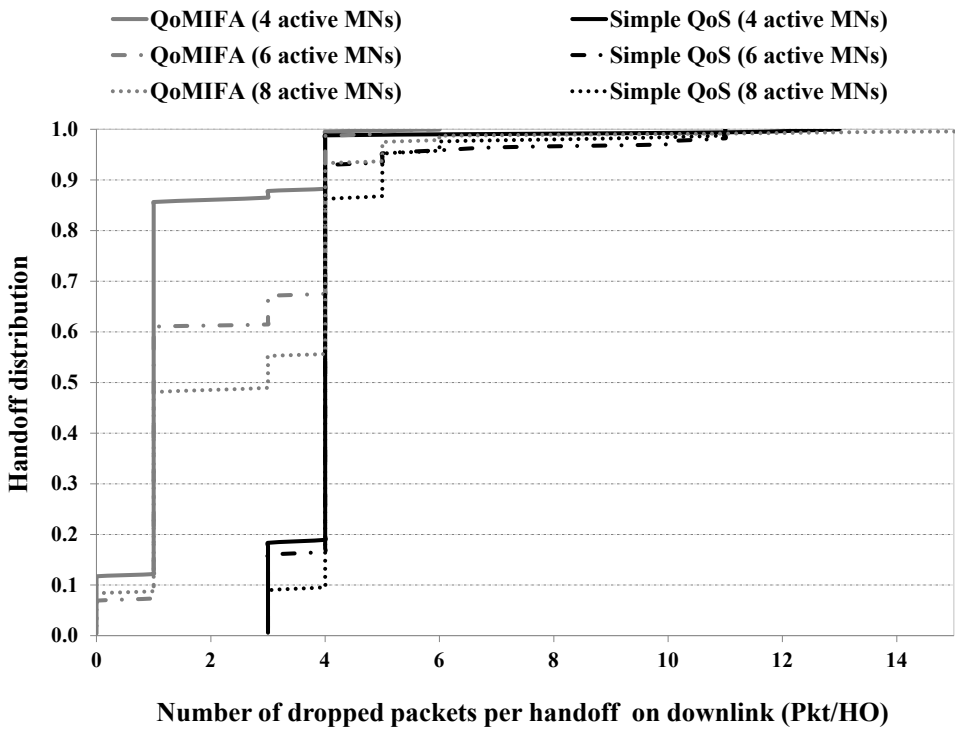


Figure 6.9: Number of dropped packets per handoff on downlink with QoMIFA and Simple QoS under different network loads

Let us now consider the average number of dropped packets per handoff on downlink, see Figure 6.10.

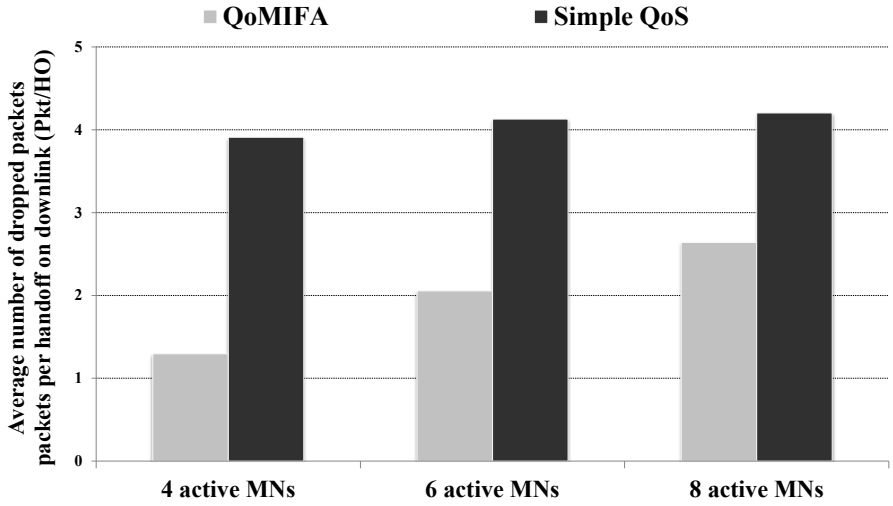


Figure 6.10: Average number of dropped packets per handoff on downlink when employing QoMIFA and Simple QoS under different network loads

One can clearly see that the average number of dropped packets per handoff increases with increasing network load. This is expected, since handoff latency increases, thus, more packets are lost during the handoff. The figure also shows that Simple QoS results in significantly more dropped packets than QoMIFA under all studied loads. The reason behind this behavior is the fast handoffs achieved by QoMIFA, which only requires, as mentioned previously, con-

tacting the old FA. On the contrary, Simple QoS requires a registration with the HA each time the MN moves in the network. This registration normally takes more time, especially if the network is high-loaded. According to the achieved results, QoMIFA reduces the average number of dropped packets by 67 %, 50 % and 37 % as compared to Simple QoS when the number of active MNs in the range of each FA is 4, 6 and 8, respectively.

6.4.3 Number of Best-Effort Packets Sent per Handoff

6.4.3.1 Uplink

A main advantage of QoMIFA is that it never forwards packets as best-effort on uplink regardless of the network load. This means that the user directly obtains his service with a QoS guarantee. The reason for this is simply the hybrid-coupling feature of QoMIFA, since it exchanges PATH and RESV messages conveying MIFA control messages with the new and old FAs after the handoff. After these messages, the handoff is completed from the MN's point of view and resources are reserved for uplink traffic. As opposed to QoMIFA, Simple QoS always sends packets as best-effort on uplink. This is because of the loose-coupling principle that Simple QoS follows, since the handoff is first executed and then resources are reserved. Clearly, the number of packets sent as best-effort per handoff increases as network load increases, see Figure 6.11.

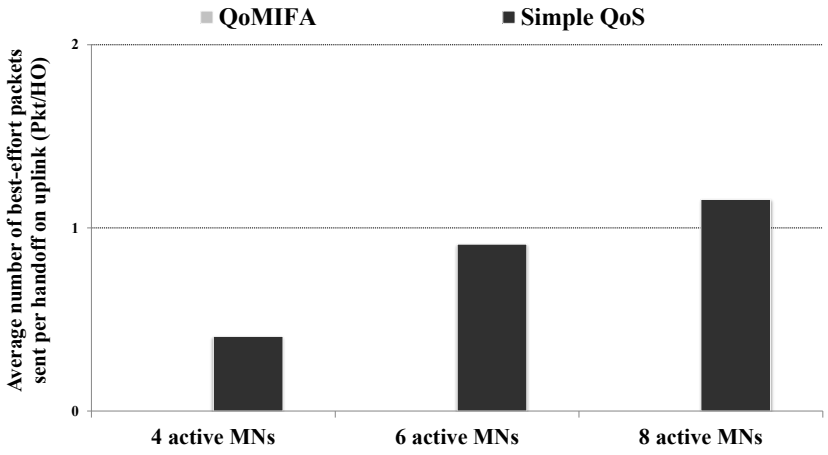


Figure 6.11: Average number of best-effort packets sent per handoff on uplink when employing QoMIFA and Simple QoS under different network loads

6.4.3.2 Downlink

In contrast to the performance of QoMIFA on uplink, some packets are sent as best-effort on downlink. The reason for this is that the old FA responds after it gets notified of the new MN's CoA (via a PATH message containing mobility information encapsulated in a mobility object) with a RESV message reserve resources on uplink. Following this, the old FA begins forwarding downlink data packets and exchanges simultaneously PATH and RESV messages with the MN to reserve resources on downlink. Therefore, some packets are sent as best-effort until the reservation is completed.

Figure 6.12 provides the average number of packets sent as best-effort per handoff when employing both studied protocols under the assumed loads. The figure shows that QoMIFA sends fewer packets as best-effort as compared to Simple QoS under all studied loads.

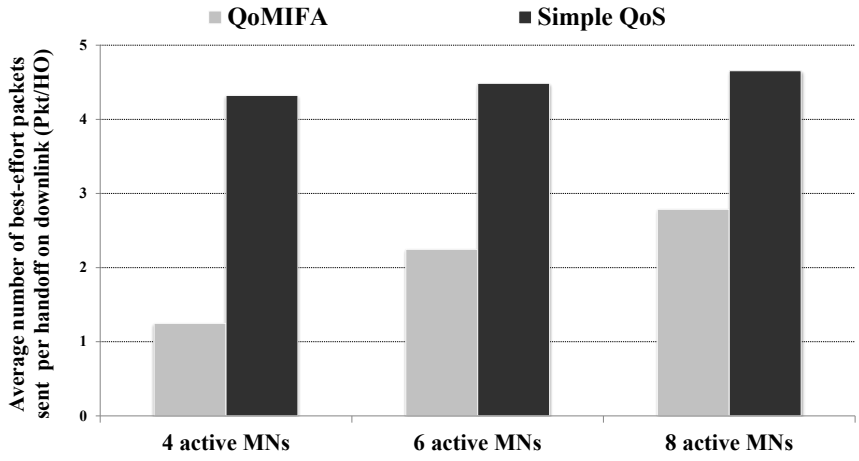


Figure 6.12: Average number of best-effort packets sent per handoff on downlink when employing QoMIFA and Simple QoS under different network loads

The reason for this is that the MN operating Simple QoS first registers with the HA. After the HA is notified, it begins forwarding data packets as best-effort to the new CoA and proceeds with reserving resources for the downlink session. This, of course, requires a considerable amount of time, which results in the forwarding of a considerable number of data packets without QoS guarantees. Keep in mind that the HA is normally far a way from the new FA. Clearly, this degrades the performance and is undesirable. According to our simulation results, Simple QoS increases the average number of packets sent as best-effort by 40 to 71 % as compared to QoMIFA under the studied network loads.

6.4.4 Probability of Dropping Sessions

As we mentioned in section 6.3, to study the probability of dropping sessions, the number of tracked MNs is varied between 4 and 8 in a scenario where each FA hosts 4 active MNs. The probability of dropping of sessions is calculated for both uplink and downlink sessions. Figure 6.13 shows the distribution function of the probability of dropping sessions employing QoMIFA and Simple QoS in the studied scenario when the number of tracked MNs is 4, 6 and 8. The first result one derives is that QoMIFA performs significantly better than Simple QoS. The more the number of MNs trying to establish sessions, the more the probability of dropping sessions. This is clearly deduced from the figure, since the sessions are dropped due to the loss of RSVP control messages. Of course, more control messages drop with increasing load.

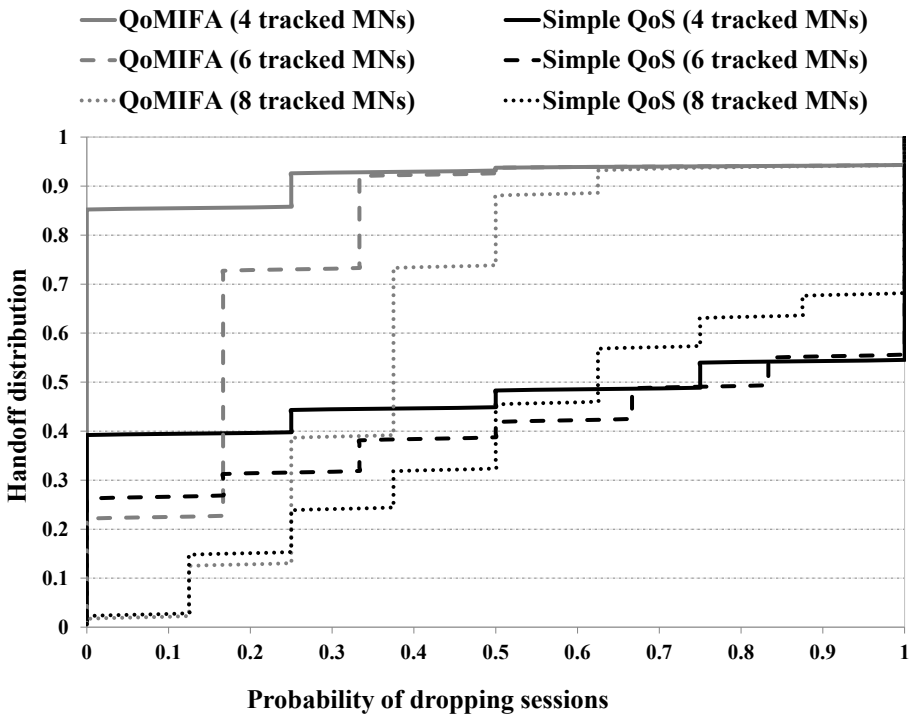


Figure 6.13: Probability of dropping sessions with QoMIFA and Simple QoS in the studied scenario

Let us first consider 4 tracked MNs in a network serving 64 active MNs. One can clearly see that Simple QoS is outperformed by QoMIFA. Our simulation results show that the probability that QoMIFA reserves sessions successfully reaches 0.852. This probability reaches, however, only 0.426 by Simple QoS. The figure shows that 52 % performance improvement in terms of no sessions dropping results when employing QoMIFA as compared to Simple QoS. Furthermore, the number of handoffs with unsuccessful reservation requests during or after handoffs is minimized by 82 % when employing QoMIFA.

It is expected that the performance of both protocols deteriorates as the number of tracked MNs increases to 6. The figure provides that the number of handoffs that do not face any session dropped due to handoffs is 22 % and 23 % with QoMIFA and Simple QoS, respectively. The number of handoffs in which all sessions are dropped due to MNs movements is 6 % using QoMIFA. These handoffs are, however, 38 % of all handoffs when employing Simple QoS.

Sure, a further increase in the number of tracked MNs results in further deteriorating the performance for both protocols, see the figure. Note that the reason for the well performance of QoMIFA is the usage of old RSVP sessions temporarily until new ones are established. This makes any dropping in new sessions between the new FA and the HA not so crucial.

Figure 6.14 presents the average probability of dropping sessions resulting from using both protocols in the studied scenario when the number of tracked MNs changes from 4 to 8.

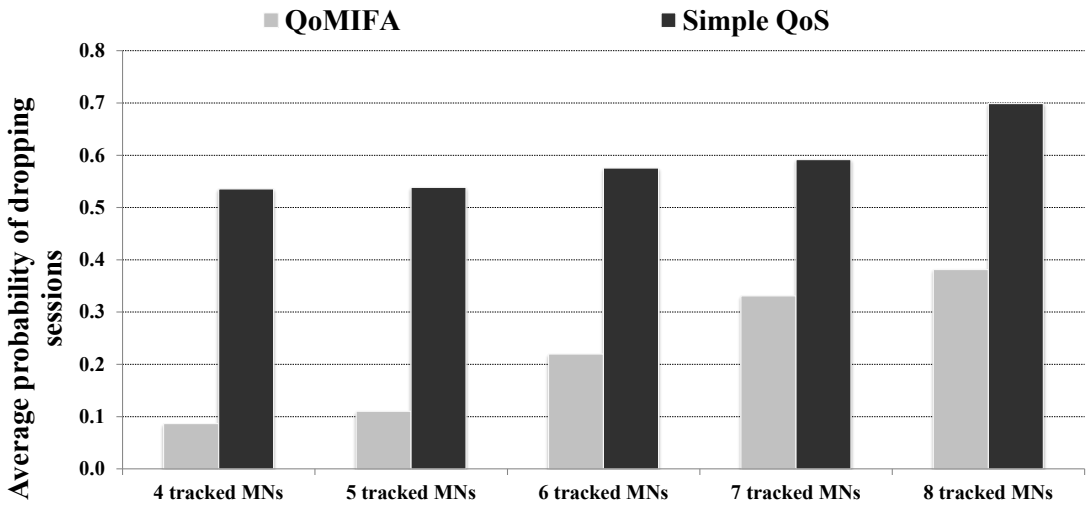


Figure 6.14: Average probability of dropping sessions when employing QoMIFA and Simple QoS in the studied scenario

The figure shows that the average value of the probability of dropping sessions increases for both protocols as network load goes up. Again, the figure shows that Simple QoS is outperformed by QoMIFA. An improvement in performance of 84 %, 80 %, 62, 44 % and 45 % results when the number of tracked MNs is 4, 5, 6, 7 and 8, respectively.

6.5 Impact of Mobile Node Speed

6.5.1 Resource Reservation Latency

6.5.1.1 Uplink

Figure 6.15 shows the distribution function of the resource reservation latency on uplink when employing QoMIFA in the studied scenario under the speeds mentioned in section 6.2.

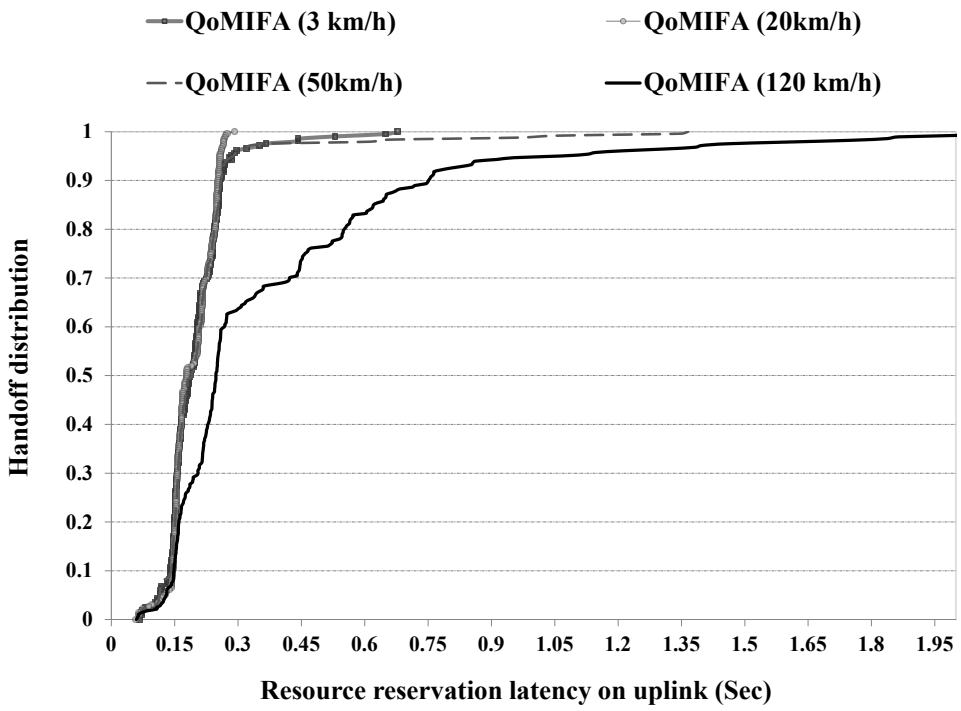


Figure 6.15: Resource reservation latency on uplink with QoMIFA under different MN speeds

As the figure shows, the speed of MNs affects the resource reservation latency for uplink traffic. Concrete, when the observed MN moves at a speed of 3 km/h, the resource reservation latency is higher than that required to reserve resources when the MN moves at a speed of 20 km/h. As the MN moves at a speed of 50 km/h, it takes longer time to reserve resources than at a speed of 20 km/h in all handoffs. Moving at a speed of 120 km/h significantly deteriorates the performance.

At first glance, the results obtained at a speed of 3 km/h looks unexpected. This is due to the fact that the ping-pong effects often appear at slow speeds. The MN spends 5.61 sec inside the overlapping area while moving at a speed of 3 km/h (the length of the MN's path inside the overlapping area between each two neighbor FAs is 5 m). Thus, the MN often switches between the old and new FA. Taking into account the fact that the quality of the old wireless link quickly deteriorates while moving inside the overlapping area away from the old FA as well as the poor quality of the new wireless link, many handoffs will not be accomplished successfully. Of course, this will result in a higher latency. Increasing the speed of MNs reduces the time the MNs spend inside the overlapping area and, with that, the ping-pong effects. This, in turn, results in improved performance. Increasing the speed to 120 km/h results in a long movement detection time, since the MN crosses the cells very quickly, this produces a considerably high latency, as the Figure 6.15 shows.

Figure 6.16 displays the distribution function of the resource reservation latency on uplink when employing Simple QoS in the studied scenario under different MN speeds.

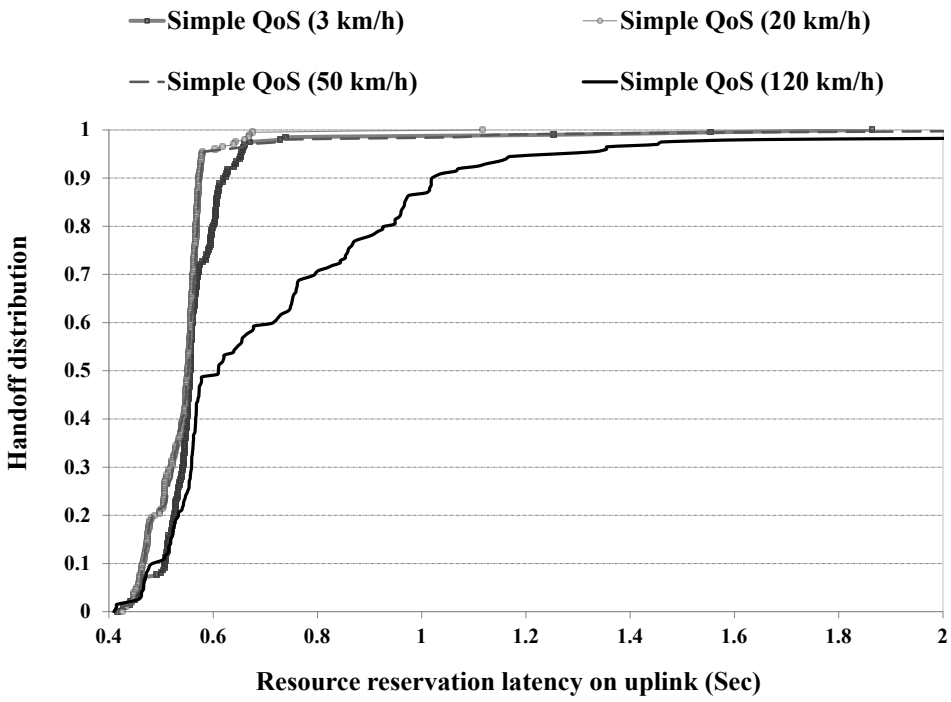


Figure 6.16: Resource reservation latency on uplink with Simple QoS under different MN speeds

From the figure, it is apparent that Simple QoS follows approximately similar behavior as QoMIFA does. Simple QoS takes at 3 km/h more time to complete the resource reservation on uplink than at 20 km/h and 50 km/h in about 92 % of the handoffs. The reason for this is the ping-pong effects discussed above. Increasing the MN speed to 120 km/h results in a clear performance degradation, since the MN crosses cells very quickly and requires more time to detect the movement at this speed than at lower speeds.

Let us now compare both protocols to each other, see Figure 6.17 which provides the average resource reservation latency on uplink when employing QoMIFA and Simple QoS under the studied MN speeds.

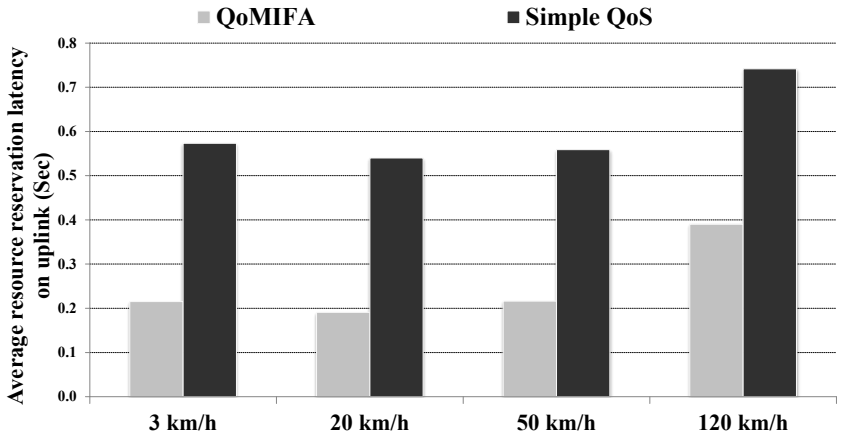


Figure 6.17: Average resource reservation latency on uplink when employing QoMIFA and Simple QoS under different MN speeds

The figure shows a significant performance improvement when employing QoMIFA as compared to Simple QoS. QoMIFA at 3, 20, 50 and 120 km/h is faster than Simple QoS by 62 %, 65 %, 61 % and 47 %, respectively. Note that the average time required to reserve resources on uplink increases by 12 % as MN speed decreases from 3 to 20 km/h when using QoMIFA. The reason for this is reducing the ping-pong handoffs discussed earlier. Thereafter, we see an increase in the average time required to reserve resources on uplink (increase by 12 %) as MN speed increases from 20 to 50 km/h. This is simply because of increasing the probability of dropping packets while increasing the MN speed to 50 km/h. Moving faster results in a significant deterioration in performance as mentioned earlier (long movement detection time). However, one sees 6 % decrease by Simple QoS when the speed of the tracked MN increases to 20 km/h. Following that, the resource reservation latency increases by 3 % and 25 % at speeds of 50 and 120 km/h, respectively.

6.5.1.2 Downlink

The distribution function of the resource reservation latency on downlink when employing QoMIFA in the studied scenario while the MN moves at the mentioned speeds is shown in Figure 6.18.

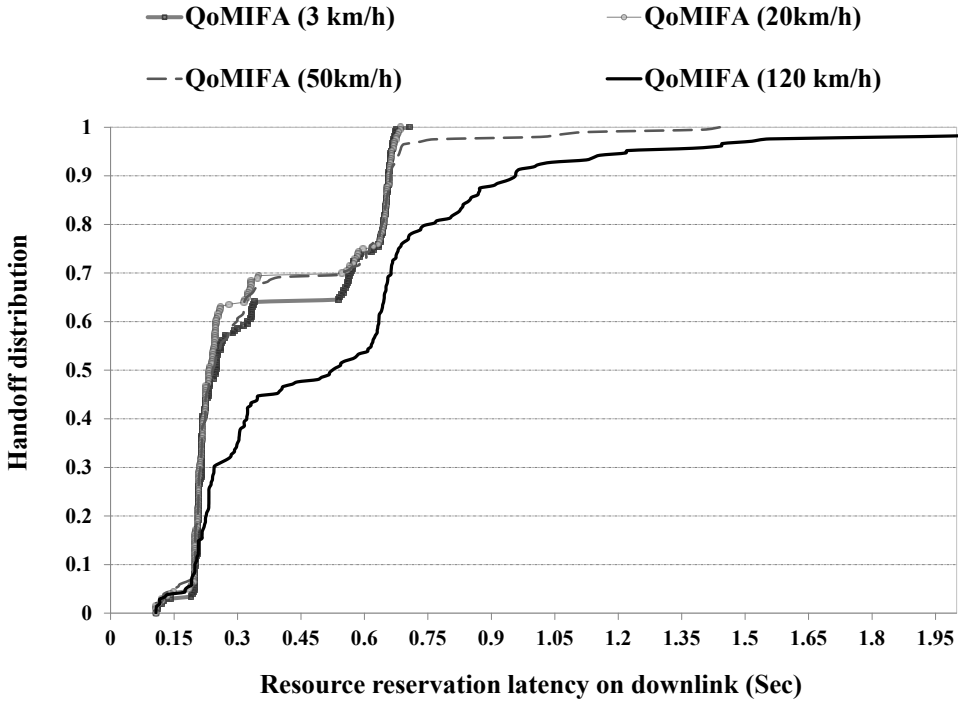


Figure 6.18: Reservation latency on downlink with QoMIFA under different MN speeds

Again, the figure depicts that the MN speed significantly affects the amount of time required to reserve resources for the downlink traffic. One notices that the behavior observed in Figure 6.15 is similar to that observed in this figure. Notice that ping-pong effects appear in about 57 % of the handoffs when the MN moves at a speed of 3 km/h. Similar behavior is observed when comparing the results when the observed MN moves at a speed of 50 km/h to that at 3 km/h. Increasing the MN speed to 120 km/h results in a considerable deterioration in performance.

The main result observed is that the ping-pong effects are not that clear visible when speaking about the resource reservation latency on downlink as the case we saw on uplink. It is agreed upon that ping-pong effects appear while moving inside overlapping areas at slow speeds. The

MN often switches between the old FA and the new one. Note, however, that the uplink RSVP sessions are built first and downlink sessions after that. So, in many cases, a new ping-pong handoff will be triggered before the reservation of downlink session is initiated. The network continues the reservation for the uplink RSVP session and starts a new handover following that. These cases are excluded from our calculation.

Figure 6.19 shows the distribution function of the resource reservation latency on the downlink when employing Simple QoS in the studied scenario under different MN speeds.

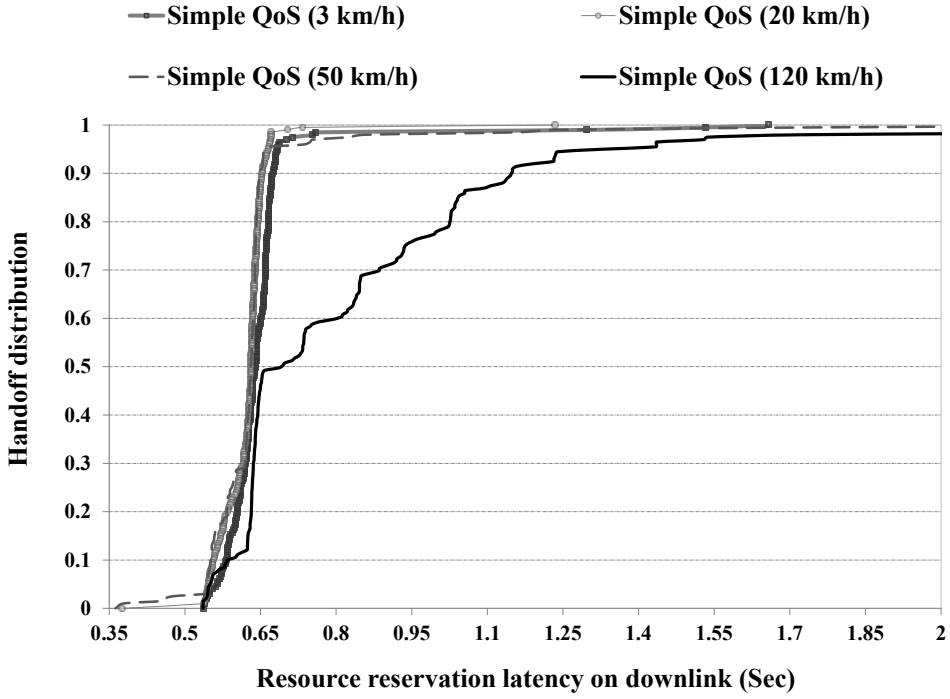


Figure 6.19: Resource reservation latency on downlink with Simple QoS under different MN speeds

Again, similar behavior to that seen in the last discussed figure can be observed in this figure, as well. When the MN operating Simple QoS moves at a speed of 20 km/h, the resource reservation latency on downlink almost less than when the MN moves at a speed of 3 km/h. The reason for this is the ping-pong effects discussed earlier. When the MN increases its speed to 50 km/h, the resource reservation latency on downlink goes in general up. However, ping-pong effects remain observable. Increasing the speed to 120 km/h deteriorates the performance. Although the ping-pong effects disappear at this speed, the performance gets down due to the increased movement detection time.

Figure 6.20 presents the average resource reservation latency on downlink when employing QoMIFA and Simple QoS when the MN moves at different MN speeds.

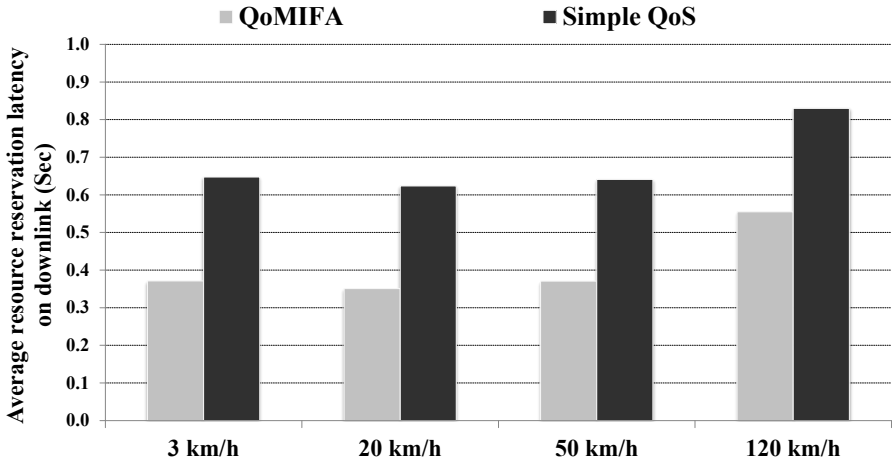


Figure 6.20: Average resource reservation latency on downlink when employing QoMIFA and Simple QoS under different MN speeds

Similar results to those derived from Figure 6.17 can be seen in Figure 6.20. QoMIFA is significantly better than Simple QoS. It reserves resources 43 %, 44 %, 42 % and 33 % faster than Simple QoS when the speed of the MN is 3, 20, 50 and 120 km/h, respectively. The average resource reservation latency decreases by 6 % when using QoMIFA as the MN speed increases from 3 to 20 km/h. Thereafter, this latency increases by 5 % when moving at 50 km/h. The reason for this was highlighted earlier. At a speed of 120 km/h, the resource reservation latency increases by 33 % once more due to the increased movement detection time. Similar behavior can be seen by Simple QoS, the average resource reservation latency decreases by 4 % as the MN speed increases from 3 to 20 km/h and increases by 3 % when the MN speed goes up to 50 km/h. The resource reservation latency experiences an increase by 23 % as the speed of the MN increases to 120 km/h. The reasons for this were investigated in section 6.5.1.1.

6.5.2 Number of Dropped Packets per Handoff

6.5.2.1 Uplink

Figure 6.21 shows the distribution function of the number of dropped packets per handoff on uplink when employing QoMIFA under the studied MN speeds. For the same reasons highlighted while discussing Figure 6.15, the number of dropped packets per handoff at a speed of 3 km/h is slightly more than that at a speed of 20 km/h. As the MN speed increases to 50 km/h, the number of packets get dropped per handoff goes up. However, the ping-pong effect remains observable. At a speed of 120 km/h, the performance deteriorates clearly. It should be noted that the curves shown in Figure 6.21 are not gradual (like staircase) since the number of dropped packets is an integer value, not real as by the resource reservation latency.

Let us now discuss the obtained results in more details. QoMIFA drops more than 5 packets per handoff at all studied speeds in about 33 % of handoffs. This implies that even at high speed it is expected that QoMIFA will operate well. The number of dropped packets per handoff does not exceed 6 packets in 89 % of the handoffs when the MN moves at a speed of 3, 20 and 50 km/h. Due to the expected performance deterioration at high speeds, the number of packets get dropped per handoff while moving at 120 km/h reaches 11 packets per handoff for the same number of handoffs, see Figure 6.21.

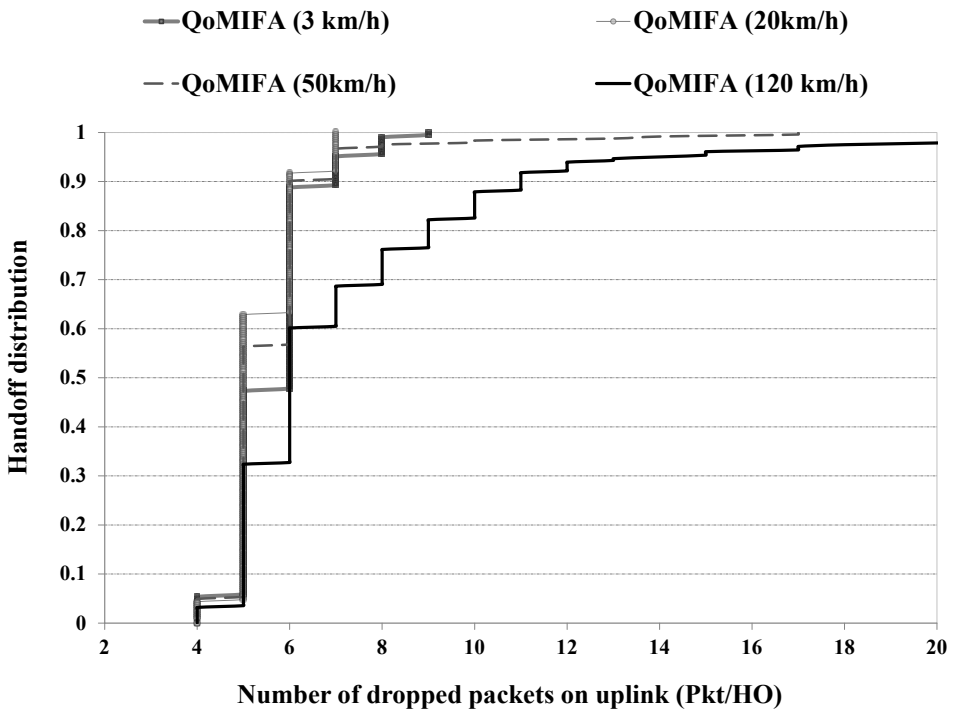


Figure 6.21: Number of dropped packets per handoff on uplink with QoMIFA under different MN speeds

The behavior of Simple QoS is similar to that observed in Figure 6.16, see Figure 6.22. This is expected since there should be a relation between the latency required to reserve resources and the number of dropped packets, see Figure 6.2. Our simulation results show that the MN operating Simple QoS drops no more than 8 packets per handoff at a speed of 3, 20, 50 and 120 km/h in approximately 52 % of the handoffs. This means that even at slow speed, one may expect that Simple QoS will not perform well. While the number of dropped packets per handoff does not exceed 8 packets at MN speeds of 3, 20 and 50 km/h in approximately 96 % of handoffs, the number of dropped packets per handoff reaches 15 packets when the MN speed goes up to 120 km/h for the same amount of handoffs.

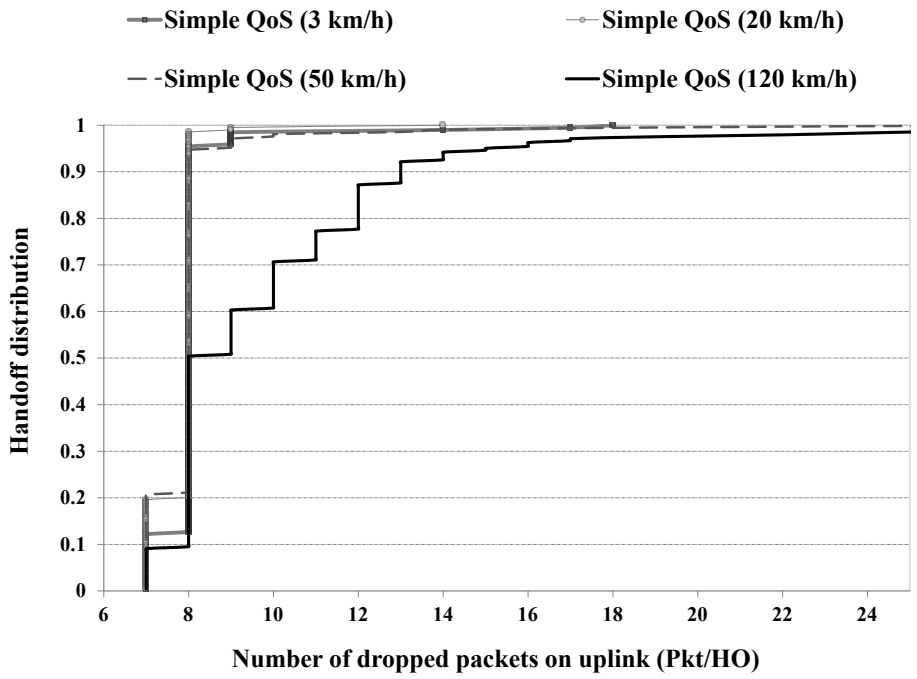


Figure 6.22: Number of dropped packets per handoff on uplink when employing Simple QoS under different MN speeds

Figure 6.23 presents the average number of dropped packets per handoff on uplink when employing QoMIFA and Simple QoS under the studied MN speeds.

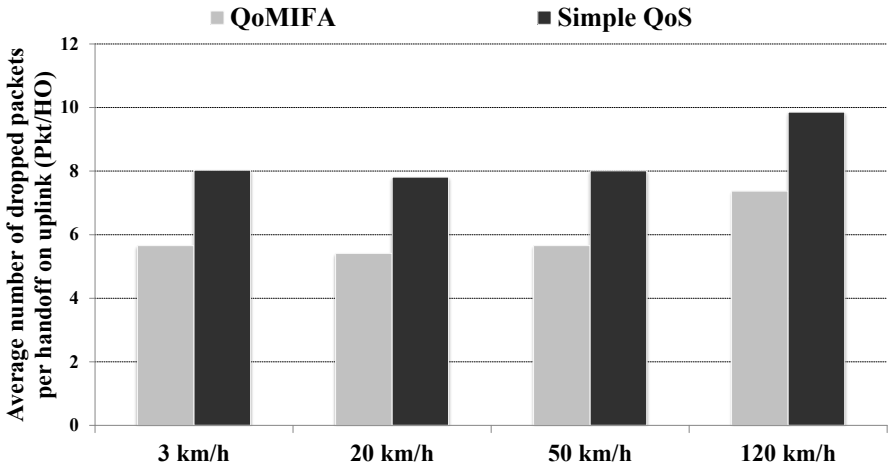


Figure 6.23: Average number of dropped packets per handoff on uplink when employing QoMIFA and Simple QoS under different MN speeds

It is expected that the same behavior observed when analyzing the time required to reserve resources on uplink (see Figure 6.15) will be seen here, as well. According to our simulation results, QoMIFA outperforms Simple QoS by 30 %, 31 %, 29 % and 25 % when the speed of the MN is 3, 20, 50 and 120 km/h, respectively.

6.5.2.2 Downlink

Figure 6.24 plots the distribution function of the number of dropped packets per handoff on downlink when employing QoMIFA under the studied speeds.

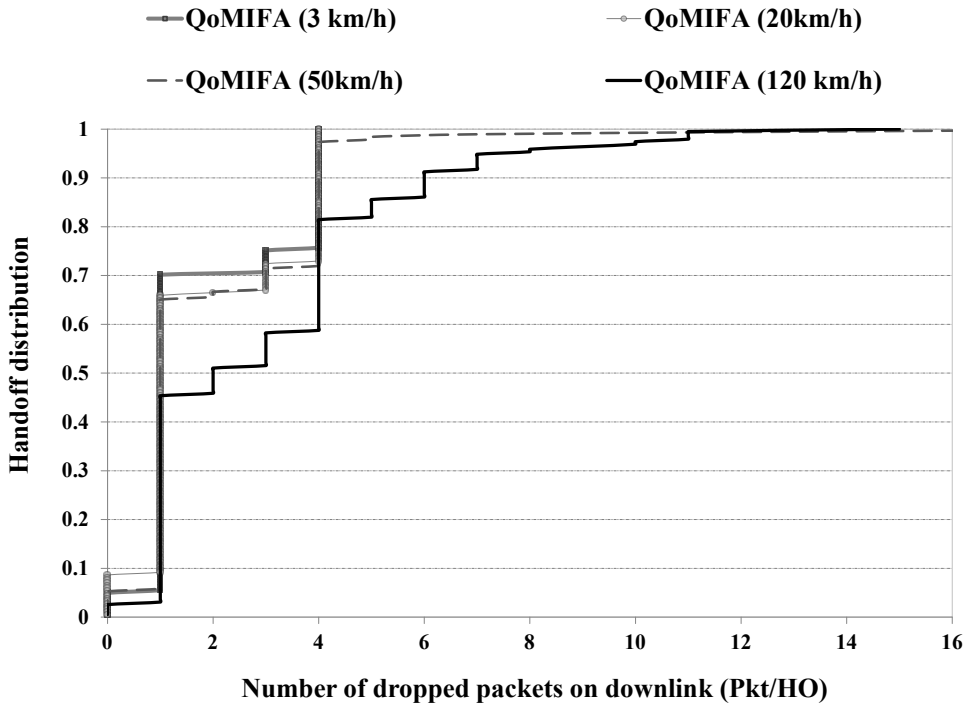


Figure 6.24: Number of dropped packets per handoff on downlink with QoMIFA under different MN speeds

The figure shows similar behavior to that observed in Figure 6.18. One observes that QoMIFA drops no more than 1 packet in approximately 42 % of the handoffs when the tracked MN moves at a speed of 3, 20, 50 and 120 km/h. Again, this implies that there is a good probability that QoMIFA will serve MNs moving at high speeds as good as those moving at slow speeds. In approximately 74 % of the handoffs, no more than 4 packets are dropped per handoff when the tracked MN moves at a speed of 120 km/h. In the remaining handoffs, the performance of QoMIFA clearly deteriorates at this speed.

In Figure 6.25, Simple QoS follows similar behavior to that seen in Figure 6.19. One can notice that, in 50 % of the handoffs, Simple QoS drops no more than 4 packets at all studied speeds. The worst performance is seen in the remaining handoffs when the MN moves at a speed of 120 km/h due to the long movement detection time, as discussed earlier.

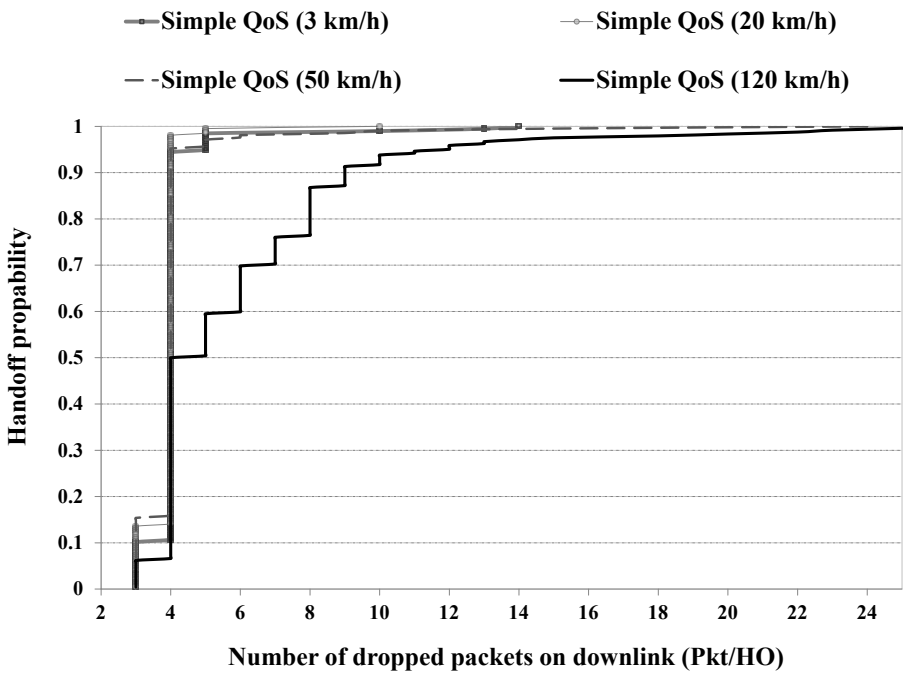


Figure 6.25: Number of dropped packets per handoff on downlink with Simple QoS under different MN speeds

Let us now compare the average number of packets get lost per handoff on downlink when employing QoMIFA to that resulting when using Simple QoS, see Figure 6.26.

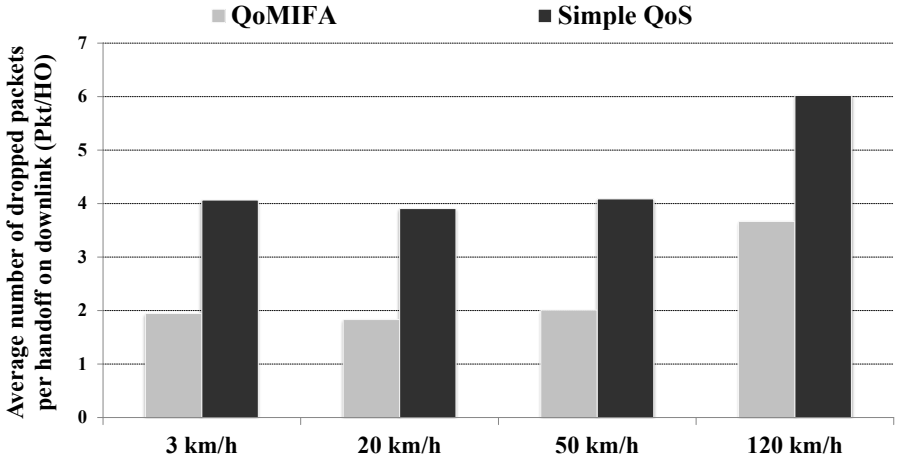


Figure 6.26: Average number of dropped packets per handoff on downlink when employing QoMIFA and Simple QoS under different MN speeds

The figure shows that QoMIFA reduces the number of dropped packets per handoff by 52 %, 53 %, 51 % and 39 % compared to Simple QoS when the speed of the MN is 3, 20, 50 and 120 km/h, respectively. It is worth noting that the same behavior is observed on downlink when employing QoMIFA and Simple QoS on uplink, see Figure 6.23.

6.5.3 Number of Best-Effort Packets Sent per Handoff

6.5.3.1 Uplink

Figure 6.27 presents the average number of packets sent as best-effort per handoff until the MN reserves resources for uplink traffic. The figure shows that QoMIFA sends no packets as best-effort on the uplink. The reason is the same highlighted while discussing the impact of network load in section 6.4.3.1. Considering Simple QoS, the number of best-effort packets decreases by 16 % as the MN speed increases from 3 to 20 km/h. The reason behind this behavior is the ping-pong effect mentioned earlier. Increasing the MN speed to 50 km/h results in increasing the number of best-effort packets by 3 %. A further increase in MN speed to 120 km/h results in increasing the number of best-effort packets by 1 %.

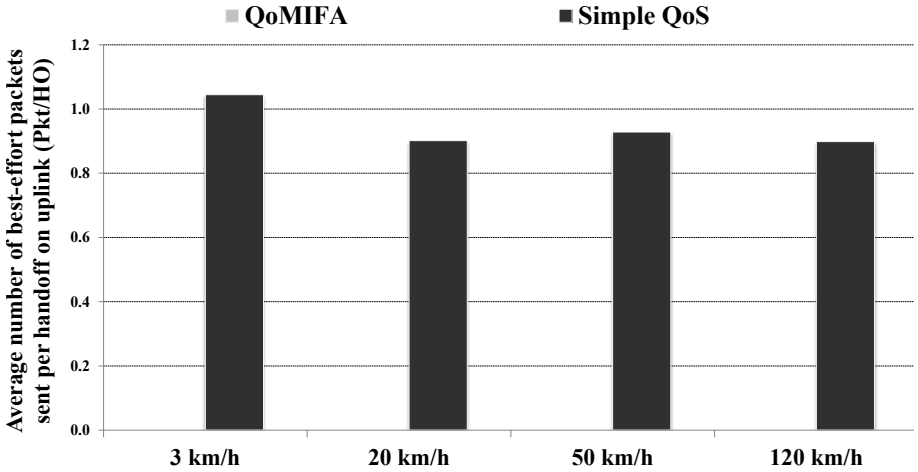


Figure 6.27: Average number of best-effort packets sent per handoff on uplink when employing QoMIFA and Simple QoS under different MN speeds

6.5.3.2 Downlink

Figure 6.28 displays the average number of packets sent as best-effort per handoff on downlink and shows similar results to those derived from Figure 6.20. QoMIFA reduces the number of best-effort packets by 45 %, 52 %, 55 % and 49 % when the speed of the MN is 3, 20, 50 and 120 km/h, respectively. The reasons for this were already discussed in sections 6.4.3.2 and 6.5.1.2.

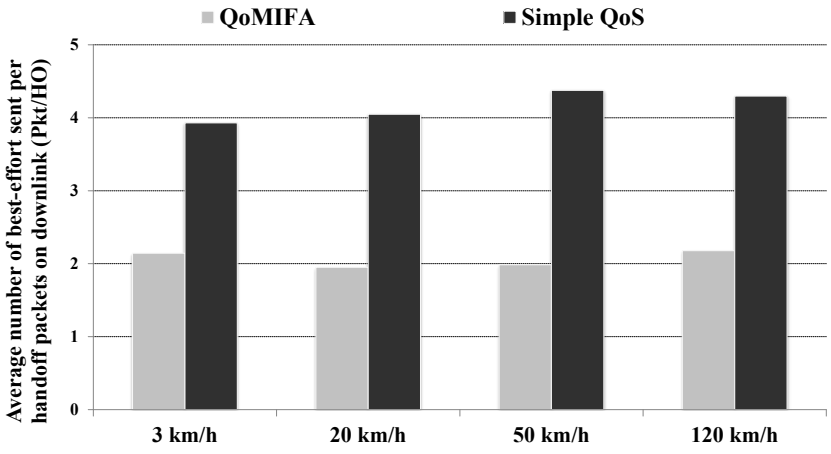


Figure 6.28: Average number of best-effort packets sent per handoff on downlink when employing QoMIFA and Simple QoS under different MN speeds

6.6 Conclusion

The main results of our simulation studies can be summarized as follows:

1. **The impact of the network load**

Increasing the network load results in an increase in the resource reservation latency, number of dropped packets per handoff, number of packets sent as best-effort per handoff and probability of dropping sessions.

- For resource reservation latency, QoMIFA is up to 67 % and 57 % faster than Simple QoS on uplink and downlink, respectively.
- Considering the number of lost packets per handoff, QoMIFA reduces the dropped packets by 25 to 35 % on uplink and 37 to 67 % on downlink as compared to Simple QoS.
- For best-effort packets, QoMIFA reduces the number of packets sent as best-effort per handoff by 40 to 71 % on downlink as compared to Simple QoS and completely eliminates these packets on uplink.
- Regarding the probability of dropping sessions , QoMIFA reduces this probability by 44 to 84 % compared to Simple QoS when the number of tracked MNs is changed between 4 and 8 in networks, where each FA hosts 4 active MNs and, one MN with background traffic.

2. **The impact of MN speed**

The impact of the ping-pong effect is seen by both protocols and causes higher resource reservation latencies, more dropped packets per handoff and more best-effort packets per handoff at low speeds than at higher ones. The worst impact of the ping-pong effect is seen at a speed of 3 km/h when employing QoMIFA and Simple QoS. It is worth mentioning that the ping-pong effect is heavily affected by metrics such as the size of overlapping area, the movement detection method applied and parameters relating to the handoff, e.g. the handoff threshold. The author in [Dia10] sates that the ECS method is not appropriate for slow speeds, the Lazy Cell Switching (LCS)[1] method would be better. At high speeds, however, ECS method should be used. Therefore, it will make sense to adapt the movement detection algorithm as well as handoff parameters to the speed of the MN. Furthermore, a good network planning will, for sure, have major contribution. The results of our simulations can be summarized as follows:

- Considering the resource reservation latency, QoMIFA is up to 65 % faster on uplink and up to 44 % faster on downlink as compared to Simple QoS.
- For the number of dropped packets per handoff, QoMIFA reduces the dropped packets by 25 % to 31 % on uplink and by 39 to 53 % on downlink as compared to Simple QoS.
- QoMIFA reduces the number of best-effort packets sent per handoff by 45 to 55 % on downlink as compared to Simple QoS and completely eliminates such packets on uplink.

All in all, there is a great performance improvement when employing QoMIFA as compared to Simple QoS. Our new proposal is capable of serving MNs well in low- and high-loaded networks. Furthermore, QoMIFA achieves its work quickly even when operating on MNs moving at high speeds.

[1] This method follows the philosophy: do not achieve a handoff until it is absolutely necessary. The MN tries to detect a new FA after the lifetime of the last Agnt_Adv message received expires.

Although the comprehensive simulation studies done in this chapter are meaningful, further evaluation is necessary and even from other points of view, namely the cost QoMIFA induces compared to other well-known counterparts. Such evaluation is the topic of the next chapter.

Chapter 7: Analysis of Signaling Cost

The simulation studies presented in Chapter 6 have showen that QoMIFA presents a solution capable of simultaneously handling mobility and QoS in a faster and smoother manner than the well-known loose-coupled protocol, Simple QoS, in low-, middle- and high-load networks as well as for MNs moving at low and high speeds. Naturally, this is desirable from a performance point of view and certainly appreciated by users and service providers. However, a comprehensive study must consider an additional point of view, namely the signaling cost that both protocols generate. Signaling cost has a special importance for service providers and network administrators, since it determines the cost to be considered when employing one of both protocols within a certain backbone.

Achieving such a comprehensive study is the goal of this chapter, which is structured as follows: section 7.1 discusses the generic mathematical model used in this study. Section 7.2 discusses the application of the mathematical model to both studied protocols, QoMIFA and Simple QoS. This section introduces the assumptions that we use in our analysis, the network topology that we deploy and the movement models that we apply. This is followed by the parameterization of QoMIFA and Simple QoS. Section 7.3 presents the results obtained, while section 7.4 concludes the chapter.

7.1 Review of the Applied Generic Mathematical Model

In [Dia10], a generic model for the analysis of mobility management protocols was introduced. The parameters of the model were selected on the basis of the studied protocols, deployed network topologies and applied mobility scenarios. The model is used to analyze the performance and signaling cost of mobility management protocols while taking into account control message dropping[1]. Performance is analyzed with respect to the average handoff latency and expected average number of dropped packets per handoff, while signaling cost is estimated regarding the location update, packet delivery and total cost. The model proposed in [Dia10] is used in the study achieved in this chapter.

Since the protocols we aim to analyze are also mobility management protocols, the model is used to estimate the cost resulting from both QoMIFA and Simple QoS in addition to analyzing their support for QoS. The location update cost is the cost resulting from the mobility binding updates after movements. The packet delivery cost is the cost required to forward data packets along the path from the CN to the new location of the MN. The total cost is the sum of both location update and packet delivery costs using an adequate weighting factor. Note that we did not use the model to analyze performance, since only average handoff latency and expected average number of packets dropped per handoff can be calculated via the mathematical model. Metrics such as resource reservation latency, sessions dropping probability, etc. cannot be calculated using this model. Note also that we decided to investigate the signaling using the mathematical model rather than the simulation since simulations analyzing signaling costs only consider counting control messages, which is not enough from our point of view. Sure, other metrics can be built in the simulation to increase the accuracy of the results con-

[1] Dropping of control messages has considerable impact on the performance. It is, however, negligible when we consider signaling since the dropping of a control message results in waiting for a timeout before retransmitting the control message again. This results in more dropped packets, higher latency, etc. From the signaling point of view, this results in only retransmitting the control message, which does not result in a significant impact on the signaling cost. Therefore, it was neglected in the model.

cerning signaling. This is, however, complicated and time consuming, which makes the mathematical model more adequate for such measurements.

The following briefly introduces the generic mathematical model developed in [Dia10], beginning with the basic assumptions followed by the applied mobility models as well as the assumed network topology. After that, we briefly present the manner in which location update, packet delivery and total cost are calculated.

7.1.1 Basic Assumptions

The model assumes that the MN moves only within one domain. The domain is interconnected with the global Internet via a gateway (abbreviated as GW in the model). IP-connectivity is offered by means of nodes referred to as Mobility Agents (MAs). In real scenarios, a MA can be an AR, a FA, etc. In addition to MAs, the domain contains nodes with mobility support called Mobility Routers (MRs) that are neither MAs nor the GW, see Figure 7.1, which shows an example domain. The figure shows that each MR interconnects an average number of MAs (referred to as ν) with the GW.

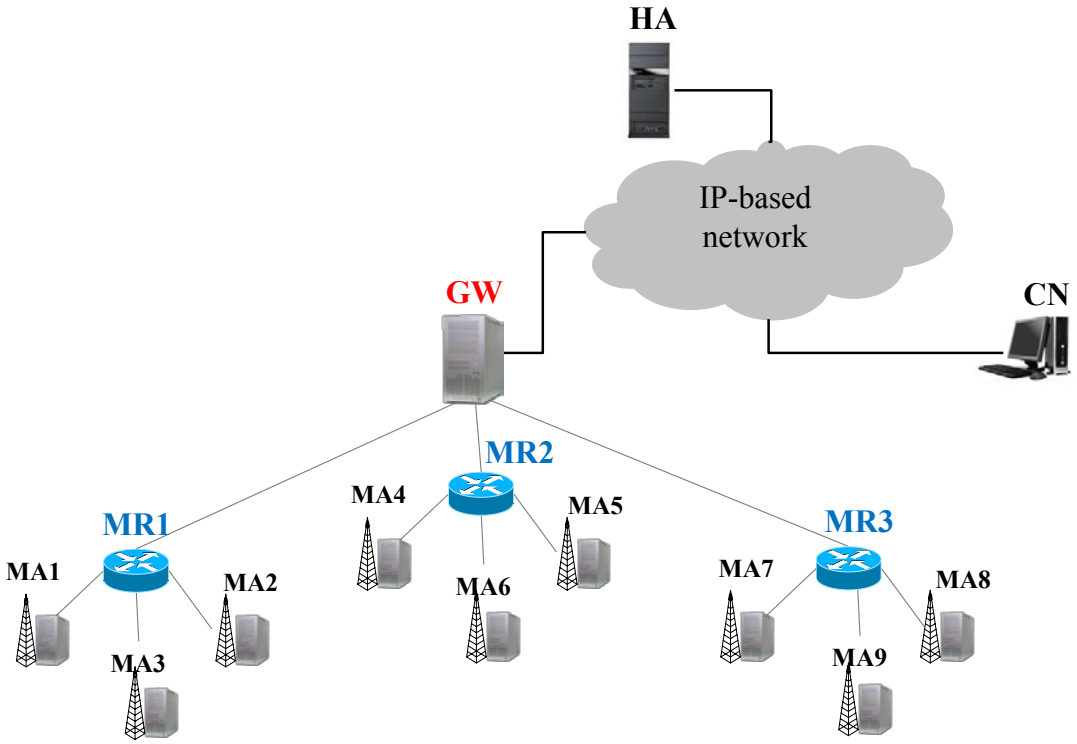


Figure 7.1: An example domain (z =9, ϑ= 3 and ν = 3)

The input parameters that the model uses are listed in Table 7-1.

Parameter	Definition
z	Number of nodes that offer IP-connectivity inside the studied domain.
N	Number of neighbors of a node that offers IP-connectivity inside the studied domain.
ϑ	Number of MRs inside the studied domain.

v	Average number of MAs each MR interconnects with the gateway.
$D_{x,y}$	Number of hops between two nodes, x and y (also termed the distance between node x and y).
$\bar{D}_{x,y}$	Average number of hops between two nodes, x and y.
T_r	Average residence time inside the coverage area of each MA. This term determines how long the MN is located inside the region of each MA.
$Tc^s_{x,y}$	Transmission cost resulting from transmitting a control message on the path from node x to y.
$Tc^D_{x,y}$	Transmission cost resulting from transmitting a data packet on the path from node x to y.
$a`_x$	Processing cost of a control message in node x.
d_x	Processing cost of a data packet in node x.
λ	Packet arrival rate.

Table 7-1: Parameters assumed in the generic mathematical model introduced in [Dia10]

In addition, the model defines the parameters listed in Table 7-1.

Parameter	Definition
luc_x	Location update cost resulting from updating the location of the MN at node x.
$pdc_{x,y}$	Packet delivery cost resulting from forwarding data packets on the path from node x to y.

Table 7-2: Parameters defined in the generic mathematical model introduced in [Dia10]

Most macro mobility management protocols update their mobility bindings at the HA (e.g. MIP). Some macro mobility management protocols actualize their mobility bindings at the old MA (e.g. MIFA in reactive mode) or even at the new MA (e.g. MIFA in predictive mode). Most micro mobility management protocols register with MRs and/or the GW as long as MNs remain moving within the same domain (e.g. MIPRR updates its mobility bindings at MRs and the GW, while HMIPv6 registers only with the GW). Some mobility management protocols choose one or more specific nodes and always update their mobility bindings at those nodes, e.g. AFA. Thus, the node(s) at which a mobility management protocol updates its mobility bindings can be either the HA, MR, GW, old MA, new MA, a specific node in the domain or a subset of the nodes we already mentioned. In the following, we refer to the control message that the MN transmits to update its mobility bindings as an **update message**. The node at which the MN updates its mobilitybindings after movements is referred to as the Binding Update node ($BUnode$). Any specific node that is chosen to be a $BUnode$ and is different from the HA, GW, MR, old MA and new MA is called an Anchor Point (ANP). Any node inside the domain with mobility support and is different from the HA, GW, MR, old MA, new MA and ANP is called an Intermediate Node ($InNode$).

In order to capture the variation in the $BUnode$ from protocol to protocol, a vector $B = [J_{MR}\ J_{GW}\ J_{HA}\ J_{MA-MR}\ J_{MA-GW}\ J_{ANP}]$ is assumed. J_x represents the probability that mobility bindings are updated at node x. J_{MA-MR} is the probability that mobility bindings are updated at the old or new MA[1] when one of the MRs is the crossover router on the old and new path between the CN and the MN[2]. J_{MA-GW} has a similar meaning to J_{MA-MR}. However, the crossover router is the GW. A distinction between these two terms is necessary to model the hierarchical topologies, for more details see [Dia10]. Finally, the last assumption of the study is a downlink and an uplink constant bit rate UDP stream exchanged between the CN and the MN with a packet arrival rate (λ) for each.

7.1.2 Movement Models

The generic mathematical model does not restrict the movement models that can be applied. The model only necessitates calculating the probabilities that MN's move between the MAs interconnected with the GW via the same or different MRs. Note that the velocity of MNs is not considered in the model.

Although no restrictions on movement models are made, as mentioned above, the model introduces a probabilistic model to further simplify its application. The probabilistic model first assumes two probabilities, namely q_i and $P_{i,j}$. q_i is the probability that the MN is turned on inside the coverage area of MA_i. $P_{i,j}$ is the transition probability between MA_i and MA_j (i.e. the probability that the MN moves from MA_i to MA_j). P denotes the matrix that collects the transition probabilities between the MAs in the domain. This matrix has the form given below. Notice that the probability that the MN moves from a MA to the same MA is equal to 0.

$$P = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,z} \\ \ldots\ldots & \ldots\ldots & \cdots & \ldots\ldots \\ p_{z,1} & p_{z,2} & \cdots & p_{z,z} \end{bmatrix} = \begin{bmatrix} 0 & p_{1,2} & \cdots & p_{1,z} \\ \ldots\ldots & \ldots\ldots & \cdots & \ldots\ldots \\ p_{z,1} & p_{z,2} & \cdots & 0 \end{bmatrix}$$

As known, the network topology in use does not affect movement patterns of MNs. In reality, movements of MNs from a certain MA are restricted to one of the MAs located in the geographical neighborhood of the MA. An example movement pattern is plotted in Figure 7.2. The probabilities not listed are assumed to be zero.

[1] Note that we use one term to express this probability, since the mobility management protocol under study will update its mobility bindings either at the old or the new MA.

[2] Eventually, there may exist more than one crossover router on the old and new path between the CN and the MN. However, we mean here that a MR is the first crossover router we pass through when going upstreams from the MN to the CN.

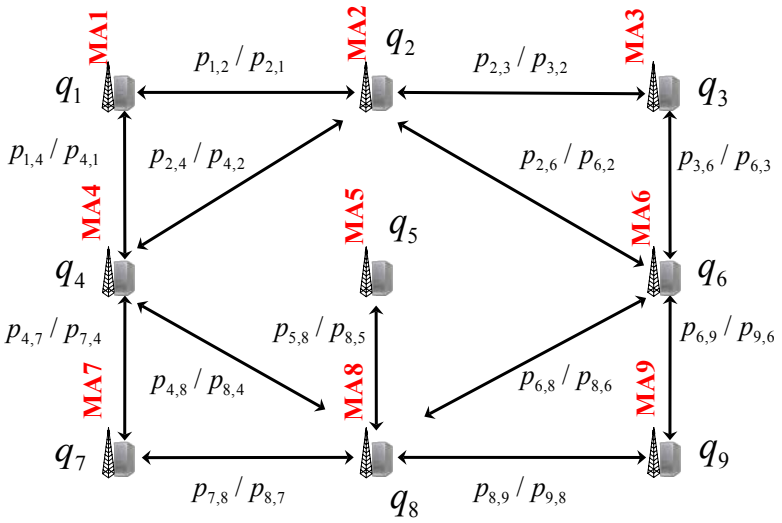


Figure 7.2: An example movement pattern

Based on a start vector $Q^0 = [q_1 \ q_2 \ \ldots \ q_z]$ and a recursive equation, $Q^{n+1} = Q^n * P$ the probabilities that a MN is attached to each MA in the steady state of the system can be written as in equation (1).

$$Q = \lim_{n \to \infty} Q^n = [P_1 \quad P_2 \quad \ldots \quad P_Z] \tag{1}$$

Based on the assumption stating that the MN moves only within one domain, one expects that the MN moves from a certain MA to a new MA that is either connected to the same MR as the old MA or to a different MR. As mentioned above, the model requires the probabilities that MNs move between MAs interconnected with the GW via the same MR (termed R) or different MR's (termed G). These probabilities can be derived from equations (2) and (3), where $P(MR_n)$ stands for the probability that the MN moves between the MAs connected to MR_n and can be written as in equation (4), where $I(MR_n)$ expresses the set of MAs located beneath MR_n.

$$R = \sum_{n=1}^{\vartheta} P(MR_n) \tag{2}$$

$$G = 1 - R \tag{3}$$

$$P(MR_n) = \sum_i \sum_j P_i * p_{j,j} \quad where \quad i, j \in I(MR_n) \quad and \quad i \neq j \tag{4}$$

7.1.3 Network Topology

As mentioned in 7.1.1, the domain contains z MAs capable of offering IP-connectivity, ϑ MRs with mobility support residing somewhere inside the domain, standard IP routers and a GW to interconnect the domain with other external networks. Nodes such as the HA and CNs are assumed to be located outside the domain under study. Based on this assumption, a 3-level logical hierarchical topology - namely MAs, MRs and a GW - is constructed, see Figure 7.3, which plots an example network topology and the corresponding logical hierarchical structure. It is worth mentioning that ϑ can be set to 0 if no MRs are contained in the studied topology. However, z is never equal to 0.

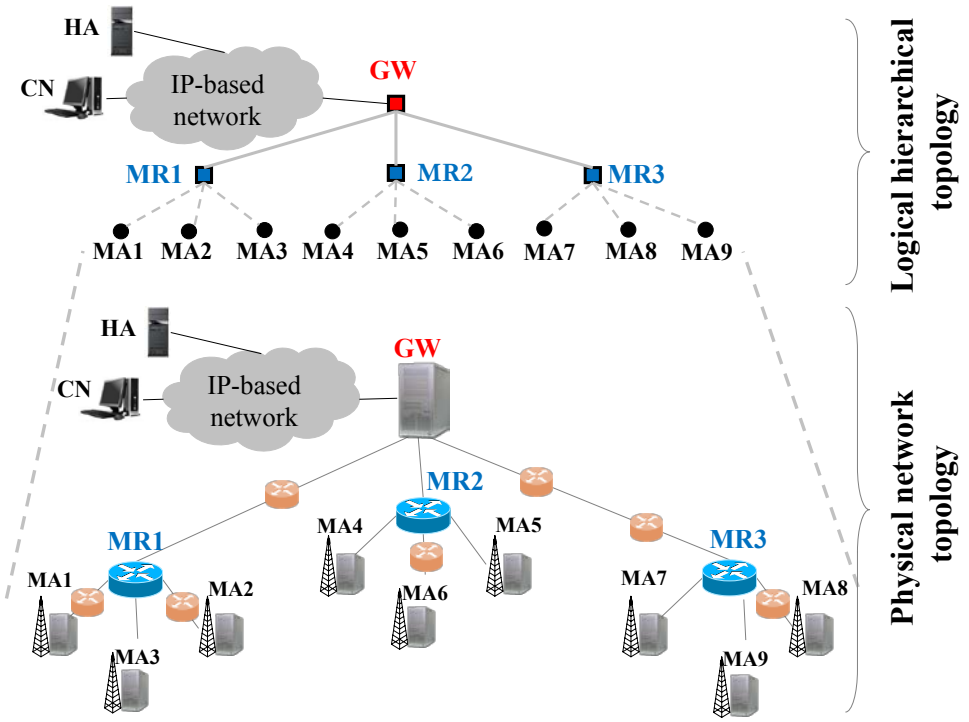


Figure 7.3: An example network topology and the corresponding logical hierarchical structure (z =9, ϑ = 3)

To model a network topology, both the wireless and wired parts of the network should be considered. The wireless part depends on the radio access technology in use (TDMA, CDMA, FDMA, OFDMA, etc.). In other words, the delay and cost resulting from transmitting a message or a data packet over the air is assumed based on the used radio access technology. The distance between a MN and a MA is 1 hop. To consider the wired part, the analysis must also consider the access technology, since this affects the delay and the cost produced when a control message or a data packet is transmitted over a wired link. Of course, other factors should not be neglected, e.g. bandwidth, load of the link, etc. All metrics mentioned strongly depend on the distance between various network nodes, e.g. $D_{MA,MR}$, $D_{MA,GW}$, $D_{newMA,oldMA}$, etc. These distances can be derived by counting the number of hops on the shortest path when employing a symmetrical hierarchical topology. However, in networks with asymmetrical hierarchical or mesh topologies, these distances vary from movement to movement and, therefore, should be calculated as average values. Of course, movement patterns should be considered in this context, as well.

Based on the discussion introduced above, the average distance between MAs and the GW can be written as in equation (5).

$$\overline{D}_{MA,GW} = \sum_{i}^{Z}\sum_{j}^{Z} P_i * p_{i,j} * D_{MA_j,GW} \qquad where \quad i \neq j \tag{5}$$

The average distance between the new MA and old MA is derived from equation (6).

$$\overline{D}_{oldMA,newMA} = \sum_{i}^{Z}\sum_{j}^{Z} P_i * p_{i,j} * D_{MA_j,MA_i} \qquad where \quad i \neq j \tag{6}$$

In order to calculate the average distance between MAs and a MR, one must first consider each MR individually, see Figure 7.4. For MR_n and the set of MAs located beneath it

($I(MR_n)$), the average distance between the MAs contained in the set $I(MR_n)$ and MR_n can be written as in equation (7).

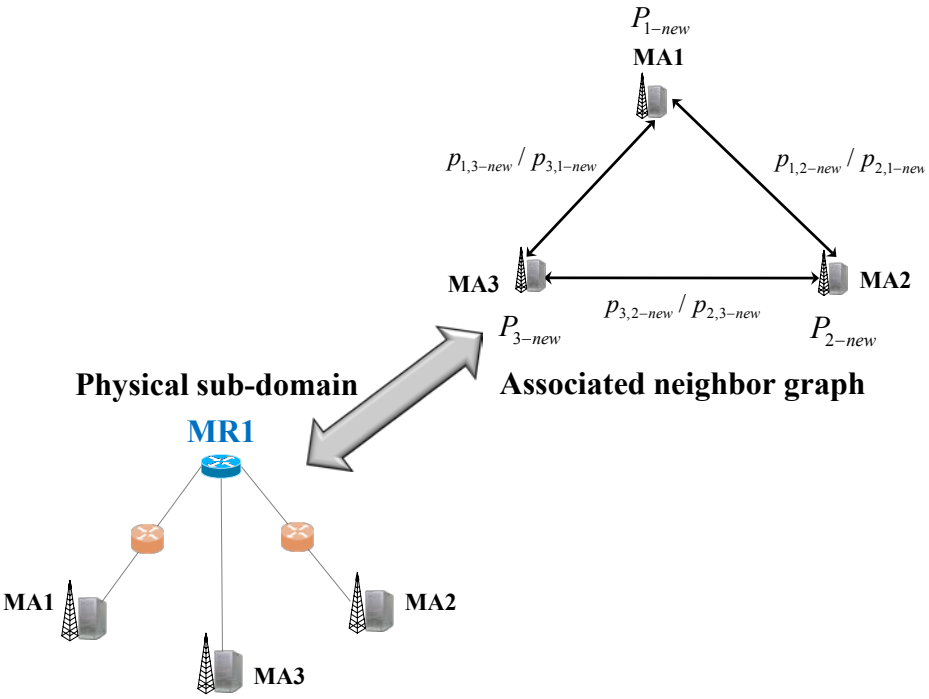


Figure 7.4: An example sub-domain consisting of a MR (MR1 in Figure 7.3) and the MAs residing beneath it in addition to the associated neighbor graph

$$\overline{D}_{MA,MR_n} = \sum_i \sum_j P_{i-new} * P_{i,j-new} * D_{MA_j,MR_n} \; where \quad i,j \in I(MR_n) \tag{7}$$

P_{i-new} and $P_{i,j-new}$ can be derived from equations (8) and (9).

$$P_{i-new} = \frac{P_i * 100}{\sum_k P_k} where \quad k \in I(MR_n) \tag{8}$$

$$P_{i,j-new} = \frac{P_{i,j} * 100}{\sum_k P_{i,k}} where \quad k \in I(MR_n) \text{ and } i \neq k \tag{9}$$

Following the calculation of the average distance between each MR located in the domain and the MAs residing beneath it, the average distance between MAs and MRs of the domain can be written as follows:

$$\overline{D}_{MA,MR} = \sum_l^{\vartheta} P(MR_l) * \overline{D}_{MA,MR_l} \tag{10}$$

The basic idea behind the calculation of the average distance between MAs and MRs in the way we have written above is as follows: each MR and all MAs located beneath it are considered as a single sub-domain containing a MN that moves only inside this sub-domain. Based on this idea, the probability that a MN will reside in the coverage area of each MA belonging to the sub-domain in addition to the transition probabilities between the MAs contained in $I(MR_n)$ must be updated accordingly. After the calculation of the average distance between

each MR located in the domain and the MAs located beneath it, the analysis must consider all MRs in the domain and, thus, calculate the average distance between the MAs and these MRs.

7.1.4 Signaling Cost Estimation

As we mentioned earlier, signaling cost mainly includes the location update and packet delivery cost. The total cost is then the sum of both these costs using an adquate weighting factor. These costs will be modeled in the next following.

7.1.4.1 Location Update Cost

The location update cost resulting from updating the location of the MN at the $BUnode$ can be calculated using equation (11).

$$luc_{BUnode} = k_1 * Tc^S_{MN,MA} + k_2 * Tc^S_{currentMA,BUnode} + k_1' * a'_{MA} + k_2' * a'_{BUnode} + ni * k_3' * a'_{InNode} + \gamma'' \quad (11)$$

where k_1 represents the number of control messages transmitted on the wireless link during a handoff. k_2 expresses the number of control messages to be exchanged between the current MA and the $BUnode$ within a handoff. k_1',k_2' and k_3' represent the number of times the **update message** has been processed in the current MA, $BUnode$ and $InNode$, respectively. ni represents the number of intermediate nodes ($InNode$). γ'' is used to capture any extra cost appearing during the movement, e.g. movement tracking, notifications of neighbors, etc. Naturally, $\gamma''$is protocol-specific and may be calculated via protocol-specific equations.

Based on [Dia10], the term $Tc^S_{x,y}$ on a wired link is proportional to the distance $D_{x,y}$ with a proportional constant δ_s, while this term on a wireless link is ρ times more than on a wired link. Thus, we can write

$$Tc^S_{x,y} = \delta_S * D_{x,y} \quad (12)$$

$$Tc^S_{MN,MA} = \rho * \delta_S \quad (13)$$

Filling in the terms written above, equation (14) results.

$$luc_{BUnode} = \delta_S * (\rho * k_1 + k_2 * D_{currentMA,BUnode}) + k_1' * a'_{MA} + k_2' * a'_{BUnode} + ni * k_3' * a'_{InNode} + \gamma'' \quad (14)$$

To capture the location update cost for all possible $BUnode$, a location update cost vector is defined as $LUC = [\, luc_{MR} \;\; luc_{GW} \;\; luc_{HA} \;\; luc_{MA-MR} \;\; luc_{MA-GW} \;\; luc_{ANP}]$. luc_{MR} , luc_{GW} , luc_{HA} and luc_{ANP} capture the location update cost when the $BUnode$ is the MR, GW, HA and the ANP, respectively. luc_{MA-MR} represent the location update cost when the old MA or new MA if the crossover router is a MR while luc_{MA-GW} is the location update cost when old MA or new MA if the crossover router is the GW. The vector LUC is protocol-specific, e.g. $LUC = [0 \;\; luc_{GW} \;\; 0 \;\; 0 \;\; 0 \;\; 0]$ means that the mobility is controlled only by the GW inside the domain.

Based on the equations defined above, the average location update cost per time unit can be calculated from equation (15).

$$luc_{TimeUnit} = \frac{B * LUC^{-1}}{T_r} \tag{15}$$

7.1.4.2 Packet Delivery Cost

Packet delivery cost consists of two terms, namely the transmission cost of data packets and the processing cost incurred in participating nodes. Notice that processing cost is defined here as the cost resulting from movements of MNs, e.g. the cost required to encapsulate and de-capsulate data packets. The packet delivery cost per time unit $pdc_{TimeUnit}$ can be written as in equation (16).

$$pdc_{TimeUnit} = pdc_{CN,MN} + Fc_{handoff} \tag{16}$$

$pdc_{CN,MN}$ is the packet delivery cost per time unit that data packets incur on the path from the CN to the MN, while $Fc_{handoff}$ is the cost produced when data packets are forwarded due to handoffs, e.g. forwarding of data packets from the old to the new MA to ensure seamless handoffs.

To capture the processing cost at the nodes that may reside on the path between the CN and the MN, a protocol-specific processing cost vector $d = [\, d_{MR} \;\; d_{GW} \;\; d_{HA} \;\; d_{oMA} \;\; d_{nMA} \;\; d_{InNode}]$ is defined. This vector contains the cost resulting from forwarding data packets in the MR, GW, HA, old MA, new MA and all $InNodes$, respectively. Again, this cost results from the mobility of MNs and not from the standard routing of data packets.

Based on the discussion introduced in this section, $pdc_{CN,MN}$ and $Fc_{handoff}$ are calculated from equations (17) and (18), respectively.

$$pdc_{CN,MN} = \sum_{i=0}^{5} d[i] + Tc^{D}_{CN,MA} + Tc^{D}_{MA,MN} \tag{17}$$

$$Fc_{handoff} = \frac{\Delta * (\sum_{i=0}^{5} d[i] + k_9 * Tc^{D}_{RNode,newMA})}{T_r} \tag{18}$$

where k_9 is the number of MAs to which data packets are forwarded due to handoffs, since some protocols multicast data packets during handoffs to a set of candidate MAs. Δ is the average time duration, during which data packets are transmitted to k_9 MAs. The forwarding of data packets due to handoffs is triggered when handoffs begin, or even earlier, and stopped after handoffs are completed. The node responsible for controlling the mentioned data forwarding is referred to as a Control Node ($ContNode$). The node that forwards data packets to k_9 MAs during the handoff is termed as Routing Node ($RNode$). It is worth mentioning that the vector d used to calculate $pdc_{CN,MN}$ and $Fc_{handoff}$ may not be the same. Moreover, the elements of the vector may be calculated using protocol-specific equations.

Again, $Tc^{D}_{x,y}$ on a wired link is proportional to the distance $D_{x,y}$ with a proportional constant δ_D, while $Tc^{D}_{x,y}$ on a wireless link is ρ times more than on a wired link. Based on this, equations (19) and (20) are defined as follows.

$$Tc^{D}_{x,y} = \lambda * \delta_{D} * D_{x,y} \tag{19}$$

$$Tc^{D}_{MN,MA} = \lambda * \rho * \delta_{D} \tag{20}$$

Filling in the terms defined above, equations (21) and (22) result.

$$pdc_{CN,MN} = \sum_{i=0}^{5} d[i] + \lambda * \delta_{D} * (D_{CN,MA} + \rho) \tag{21}$$

$$Fc_{handoff} = \frac{\Delta * (\sum_{i=0}^{5} d[i] + k_{9} * \lambda * \delta_{D} * D_{RNode,newMA})}{T_{r}} \tag{22}$$

7.1.4.3 Total Cost

As mentioned in section 7.1, the total cost is the sum of both location update and packet delivery costs using an adequate weighting factor. Thus, the total cost per time unit can be written as in equation (23). φ is a weighting factor that expresses the importance of the location update cost against the packet delivery cost.

$$C_{Total} = \varphi * luc_{TimeUnit} + (1-\varphi) * pdc_{TimeUnit} \tag{23}$$

7.1.5 Validation of the Mathematical Model

The mathematical model we use in our study was validated in [Dia10] compared to simulations as well as a real testbed. The model was applied to certain topologies under certain circumstances and parameters, which are further used for the simulations as well as the testbed. Validation results showed that the generic mathematical model presented provides sound evaluation of mobility management protocols in low-loaded networks. The accuracy of the generic mathematical model lies in a range of ±23 % when comparing to simulation results under various loads and ±30 % when comparing to real testbed results.

So, because the scenario applied in our signaling cost investigation presents a low-load scenario, the results the generic mathematical model delivers are accurate based on the validation results shown above.

7.2 Application of the Generic Mathematical Model

This section discusses the application of the generic mathematical model introduced in section 7.1 to both studied protocols, namely QoMIFA and Simple QoS. The section first introduces the basic assumptions, on which our analysis is based. Following that, the deployed network topology and the applied movement models are presented. Finally, the parameterization of QoMIFA and Simple QoS is introduced.

7.2.1 Basic Assumptions

As often mentioned in this chapter, the goal of our study is to investigate the location update and packet delivery cost. For this purpose, our study makes the following assumptions:

1. Regarding the location update cost

a. The location update cost comprises both the signaling used to register with the $BUnode$ and the signaling issued to reserve resources.
b. After the MN detects a movement to a new cell, it always exchanges a solicitation and an advertisement with the new detected MA directly after the completion of the layer 2 handoff (the establishment of a wireless link).
c. Because our analysis focuses on the signaling cost resulting from handoffs, the analysis considers neither the cost resulting from the initial registration, nor the cost required to refresh the reserved resources, nor the cost produced when resources are released. One must also consider that the cost necessary for the refresh and release of resources is comparable in the two studied protocols, since this cost relates to the path between the MN and the CN, which is almost identical in both protocols. Note that although QoMIFA reserves resources between the old and new FA to accelerate the handoff, these resources are temporal and will be released with the resources reserved on the old path once the resources on the new path are reserved. The extra cost resulting from the release is negligible and not considered. The refresh of resources on the new path between the MN and the CN is the same in both protocols and is, therefore, not considered in our study as mentioned.
d. For simplicity, we assume that all reservations are bidirectional.
e. Similar assumptions to those followed by the studies implemented in [JAk02] and [Dia10] are followed. The cost required to transmit signaling messages is assumed to be available (i.e. $Tc_{x,y}^{S}$ is available). The cost required to process messages in MRs, the GW, the HA, MAs and $InNodes$ is also assumed to be available. The cost can be expressed as the delay necessary to process and transmit control messages. Other assumptions are also allowed, e.g. considering criteria such as the available bandwidth, expenses necessary to operate a particular node, etc.

2. Regarding the packet delivery cost
   a. Our study investigates the packet delivery cost incurred between the CN and the MN.
   b. Similar to the location update cost, we assume that the cost required to transmit and process data packets is available (i.e. $Tc_{x,y}^{D}$ is available).
   c. Although all reservations are bidirectional, only the packet delivery cost for the downlink will be calculated and analyzed. This is because the packet delivery cost for the uplink traffic is the same as produced by the downlink traffic based on our assumptions discussed in section 7.1.1.
   d. The packet arrival rate (λ) for both downlink and uplink CBR UDP streams is equal to 50 packets per second.

7.2.2 Applied Network Topology

The network topology used in our analysis is shown in Figure 7.5. As the figure illustrates, the topology used in the analysis that we aim to achieve is the same as that used in the simulation.

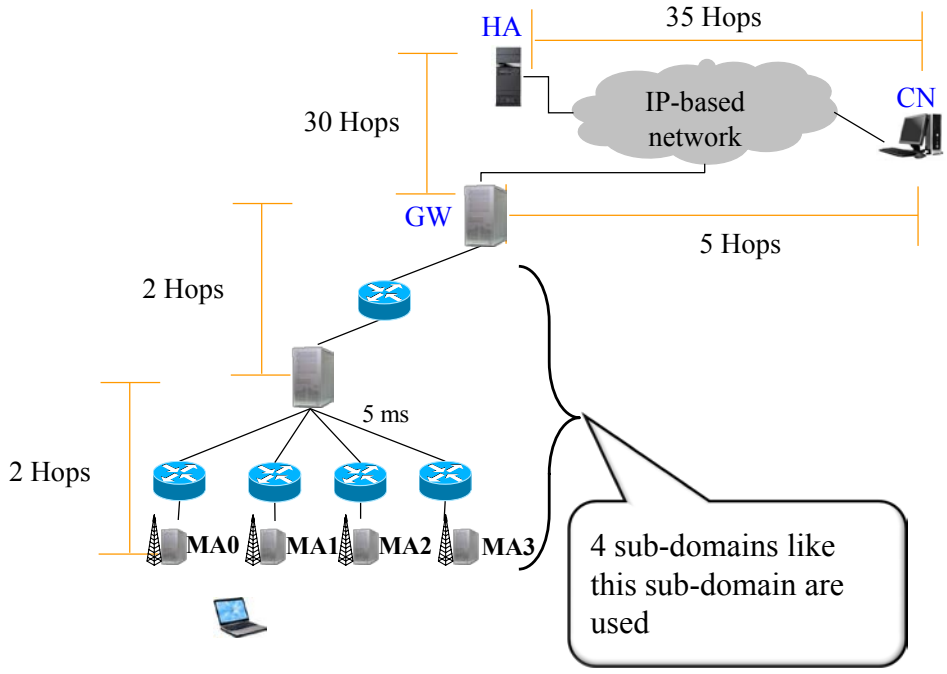


Figure 7.5: Network topology applied

The parameters of the assumed network topology are provided in Table 7-3. The values of the proportional constants δ_s and δ_D as well as ρ are taken from [Dia10].

δ_s	δ_D	ρ	$D_{MA,MR}$	$D_{MA,GW}$	$D_{GW,HA}$	$D_{GW,CN}$	$D_{CN,HA}$
0.5	0.05	10	2 hops	2 hops	30 hops	5 hops	35 hops

Table 7-3: The parameters of the network topology deployed

Notice that the number of hops between the GW and the HA is 30. This corresponds to a delay of 150 msec in the topology used for the simulation (i.e. the delay for each hop is 5 msec), see section 6.2. Other distances are calculated in a similar manner. Notice that other parameters such as the delay on wireless links, the delay on each wired link, etc. are not shown in the table, since they are not required in our cost analysis.

7.2.3 Applied Movement Models

To achieve a comprehensive analysis of signaling cost, the neighbor graph of the MAs located in the domain is first assumed to be as displayed in Figure 7.6. Notice that the figure only shows the transitions from MA0 to its neighbors. All other transitions from any given MA to its neighbors are similar. We assume that all MAs located inside a circle with a radius of 198 m and centered at any given MA are members of the L3-FHR of that particular MA.

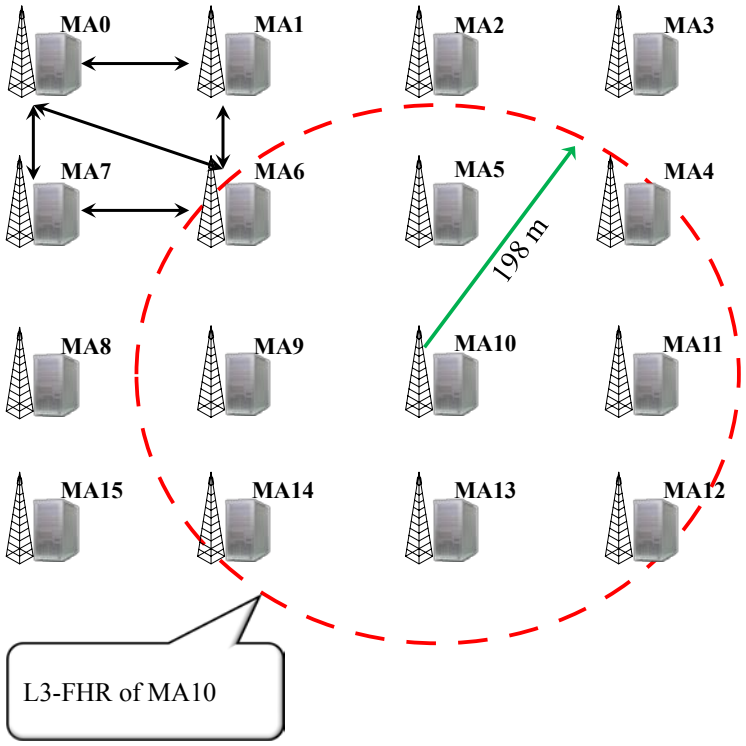


Figure 7.6: Movement model applied (mesh scenario)

For simplicity, we assume that the MN is turned on with an equal probability inside the coverage area of each MA, i.e. $q_1 = q_2 = \cdots = q_{15} = 0.0625$. Furthermore, the MN is capable of moving from any given MA to $N-1$ others with an equal probability ($\frac{1}{N-1}$), where $N-1$ is the number of the given L3-FHR members of that MA, as Figure 7.6 shows. Based on the neighbor graph presented in the figure, the matrix P can be derived as follows.

$$P = \begin{bmatrix}
0 & 0.33 & 0 & 0 & 0 & 0 & 0.33 & 0.33 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.2 & 0 & 0.2 & 0 & 0 & 0.2 & 0.2 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.2 & 0 & 0.2 & 0.2 & 0.2 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.33 & 0 & 0.33 & 0.33 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.2 & 0.2 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0.2 & 0.2 & 0 & 0 & 0 & 0 \\
0 & 0.125 & 0.125 & 0.125 & 0.125 & 0 & 0.125 & 0 & 0 & 0.125 & 0.125 & 0.125 & 0 & 0 & 0 & 0 \\
0.125 & 0.125 & 0.125 & 0 & 0 & 0.125 & 0 & 0.125 & 0.125 & 0.125 & 0.125 & 0 & 0 & 0 & 0 & 0 \\
0.2 & 0.2 & 0 & 0 & 0 & 0 & 0.2 & 0 & 0.2 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.2 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0.2 & 0.2 \\
0 & 0 & 0 & 0 & 0 & 0.125 & 0.125 & 0.125 & 0.125 & 0 & 0.125 & 0 & 0 & 0.125 & 0.125 & 0.125 \\
0 & 0 & 0 & 0 & 0.125 & 0.125 & 0.125 & 0 & 0 & 0.125 & 0 & 0.125 & 0.125 & 0.125 & 0.125 & 0 \\
0 & 0 & 0 & 0 & 0.2 & 0.2 & 0 & 0 & 0 & 0 & 0.2 & 0 & 0.2 & 0.2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.33 & 0.33 & 0 & 0.33 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.2 & 0.2 & 0.2 & 0 & 0.2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.2 & 0.2 & 0 & 0 & 0.2 & 0 & 0.2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.33 & 0.33 & 0 & 0 & 0 & 0 & 0.33 & 0
\end{bmatrix}$$

From this matrix one drives that R is 0.3 and G is 0.7.

The second mobility model under study is the same movement model applied in the simulation, see Figure 7.1. The MN turns on in the range of MA0 with a probability equal to 1, i.e. $q_1 = q_2 = \cdots = q_{15} = 0$. The transitions between MAs are shown in Figure 7.7.

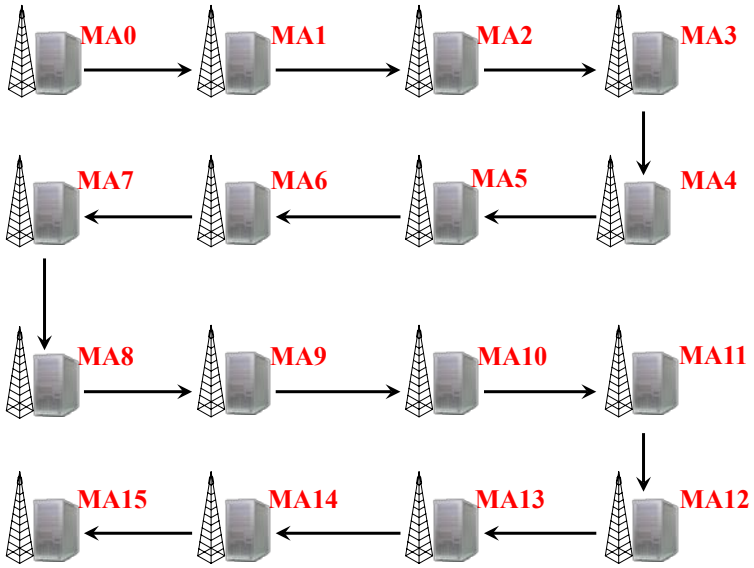


Figure 7.7: Movement model applied (linear scenario)

Based on the neighbor graph provided in the figures above, the matrix P results.

$$P = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

Based on this matrix, we can derive that R will be 0.8 and G will be 0.2.

7.2.4 Application of the Generic Mathematical Model to QoMIFA

The MN that employs QoMIFA contacts the old MA to quickly resume its communication while the registration with the HA is in progress. Thus, there are two $BUnodes$ from the cost point of view, namely the old MA and the HA. The vector B is written as follows: $B = [0 \;\; 0 \;\; R+G \;\; R \;\; G \;\; 0]$. This vector indicates that the MN updates its mobility bindings at the HA after all movements. The bindings are actualized at the old MA in R % via the MRs and in G % via the GW. In a similar way, the location update cost vector (LUC) is defined

as $[0 \;\; 0 \;\; luc_{HA} \;\; luc_{MA-MR} \;\; luc_{MA-GW} \;\; 0]$. luc_{HA}, luc_{MA-MR} and luc_{MA-GW} are calculated using equation (11). The parameters required to calculate luc_{MA-MR} and luc_{MA-GW} are provided in Table 7-4. γ'' represents the cost experienced when exchanging a solicitation and an advertisement with the new MA.

k_1	k_2	k'_1	k'_2	k'_3	a'_{MA}	a'_{BUnode}	a'_{InNode}	ni	γ''	$D_{currentMA,BUnode}$
4	4	4	2	0	10	25	0	0	$2 * \rho * \delta_s$	4 hops (crossover router is a MR) 8 hops (crossover router is the GW)

Table 7-4: Parameters required to calculate luc_{MA-MR} and luc_{MA-GW} when employing QoMIFA

The parameters necessary to compute luc_{HA} are provided in Table 7-5. The value of k_1 is set to 0 when calculating luc_{HA} because the HA is notified by the new MA and not by the MN.

k_1	k_2	k'_1	k'_2	k'_3	a'_{MA}	a'_{BUnode}	a'_{InNode}	ni	γ''	$D_{currentMA,BUnod}$
0	4	2	2	0	10	25	0	0	a'_{MA} $+ 2 * N_{av} * \delta_s$ $+\overline{D}_{currentMA,nei}$	34 hops

Table 7-5: Parameters required to calculate luc_{HA} when employing QoMIFA

γ'' represents the cost resulting from the notification of neighbor MAs of the incoming MN. N_{av} indicates the average number of MAs in the L3-FHRs. This term can be written as in equation (24).

$$N_{av} = \frac{\sum_{i=0}^{Z-1} N_i}{Z} \tag{24}$$

where N_i is the number of MAs present in the L3-FHR of MA_i. $\overline{D}_{currentMA,neiMA}$ is the average distance between the current MA and its neighbor. The distance $D_{currentMA,neiMA}$ is equal to 4 hops when the crossover router is a MR and 8 hops when the crossover router is the GW. Thus, $\overline{D}_{currentMA,neiMA} = 4 * R + 8 * G$.

Let us now investigate the packet delivery cost. Because data packets are forwarded from the CN to the MN via a triangular route (CN → HA → current MA → MN), the processing cost vector required to calculate $pdc_{CN,MN}$ is equal to $[0 \;\; 0 \;\; d_{HA} \;\; 0 \;\; d_{nMA} \;\; 0]$, where d_{HA} and d_{nMA} are computed using equations (25) and (26), respectively.

$$d_{HA} = \eta_1 * \lambda \tag{25}$$

$$d_{MA} = \eta_2 * \lambda \tag{26}$$

η_1 and η_2 represent the packet delivery processing cost constants in the HA and a MA, respectively. We assume in our study that both η_1 and η_2 are equal to 1.

QoMIFA forwards data packets during handoffs from the old to the new MA. Thus, the processing cost vector applied to compute $Fc_{handoff}$ is equal to $[0\ \ 0\ \ 0\ \ d_{oMA}\ \ d_{nMA}\ \ 0]$. Moreover, the old MA forwards data packets during the handoff only to the new MA. This means that k_9 is equal to 1. The $ContNode$ that stops the forwarding of data packets is the HA, whereas $RNode$ is the old MA. The average value of Δ is assumed to be 305 msec for the mesh scenario and 315 msec for the linear one. These values are calculated as follows: the new MA simultaneously sends two PATH messages after the MN hands off to its range, one towards the old MA with a PFA_Not message encapsulated in a mobility object and another to the HA carrying a HA_Not message. The average time the old MA requires to get notified of the new CoA is 36 msec when applying the first mobility scenario and 26 msec when applying the second mobility scenario. The time required to inform the HA (i.e. the time at which the HA receives the PATH message) is equal to 171 msec. This means that the old MA will forward data packets for 135 msec and 145 msec until the HA gets notified when applying the mesh and linear scenario, respectively. Let us assume that the HA has just sent a data packet before it received the PATH message from the new MA. This data packet requires 170 msec to reach the old MA. This means that the old MA continues to forward data packets to the new MA for a duration of 305 msec and 315 msec when applying the mesh and linear scenario, respectively.

7.2.5 Application of the Generic Mathematical Model to Simple QoS

Simple QoS first employs MIPv4 and then establishes an RSVP tunnel between the new MN and the HA. Therefore, the signaling cost resulting from MIPv4 will be discussed first before discussing that resulting from the RSVP tunnel.

In the case of MIPv4, the MN that employs MIPv4 updates its bindings always at the HA. In other words, the $BUnode$ is the HA. As a result, the vector B is equal to $[0\ \ 0\ \ R+G\ \ 0\ \ 0\ \ 0]$. The location update cost vector (LUC) is equal to $[0\ \ 0\ \ luc_{HA}\ \ 0\ \ 0\ \ 0]$, where luc_{HA} is calculated using equation (11). The parameters required for the computation of luc_{HA} are provided in Table 7-6. Again, γ'' results from the exchange of a solicitation and an advertisement message with the new MA.

k_1	k_2	k'_1	k'_2	k'_3	a'_{MA}	a'_{BUnode}	a'_{InNode}	ni	γ''	$D_{currentMA,BUnode}$
2	2	2	1	0	10	25	0	0	$2*\rho*\delta_s$	34 hops

Table 7-6: Parameters required to calculate luc_{HA} when employing MIPv4

As mentioned above, Simple QoS establishes a bidirectional RSVP tunnel between the new MA and the HA following the handoff. From the cost point of view, the $BUnode$ is the HA. The vector B is thus equal to $[0\ \ 0\ \ R+G\ \ 0\ \ 0\ \ 0]$, while the location update cost vector (LUC) is equal to $[0\ \ 0\ \ luc_{HA}\ \ 0\ \ 0\ \ 0]$. Again, luc_{HA} is computed using equation (11). Other parameters necessary for the calculation of luc_{HA} are provided in Table 7-7. Notice that γ'' is set to 0 because no extra messages are exchanged other than the PATH and RESV messages exchanged between the new MA and the HA.

k_1	k_2	k'_1	k'_2	k'_3	a'_{MA}	a'_{BUnode}	a'_{InNode}	ni	γ''	$D_{currentMA,BUnode}$
0	4	2	2	0	10	25	0	0	0	34 hops

Table 7-7: Parameters required to calculate luc_{HA} when establishing the bidirectional RSVP tunnel

Let us now discuss the packet delivery cost. Similar to QoMIFA, the processing cost vector required to calculate $pdc_{CN,MN}$ is equal to [0 0 d_{HA} 0 d_{nMA} 0] because data packets are forwarded via a triangular route (CN → HA → current MA → MN). d_{HA} and d_{nMA} can be calculated using equations (25) and (26), respectively. Again, η_1 and η_2 are assumed to be 1. Furthermore, there is no forwarding of data packets during the handoff when Simple QoS is used. This implies that the term $Fc_{handoff}$ is equal to 0.

7.3 Analytical Results

This section estimates the cost resulting from both studied protocols. It first goes through the location update cost employing both assumed mobility models and then discusses the packet delivery cost.

7.3.1 Location Update Cost

Figure 7.8 shows the location update cost experienced when applying both studied protocols as a function of the residence time (T_r) when the mesh scenario is applied. As can be seen in Figure 7.8, the location update cost has a negative exponential distribution as a function of the residence time for both studied protocols, i.e. the location update cost decreases exponentially with increasing residence time.

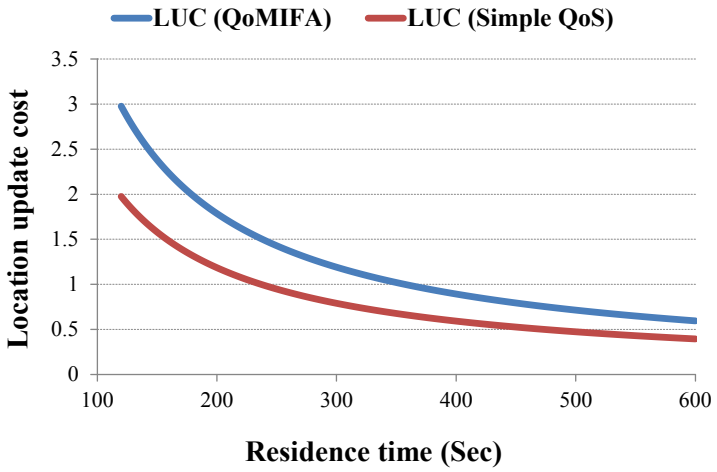


Figure 7.8: Location update cost experienced when employing both studied protocols as a function of the residence time (T_r) when applying the mesh mobility scenario

Although QoMIFA only transmits RSVP messages that carry MIFA control messages, whereas Simple QoS employs MIP first followed by RSVP, QoMIFA results in 34 % greater location update cost than Simple QoS according to our results. The main reason behind this is that QoMIFA notifies the old MA in addition to the HA, while Simple QoS only depends on informing the HA. Furthermore, QoMIFA in-advance informs adjacent MAs of incoming MNs, which produces extra cost.

Similar results are observed when applying the linear scenario, see Figure 7.9. Our results indicate that Simple QoS generates 28 % smaller location update cost than QoMIFA. The reasons for this were discussed above.

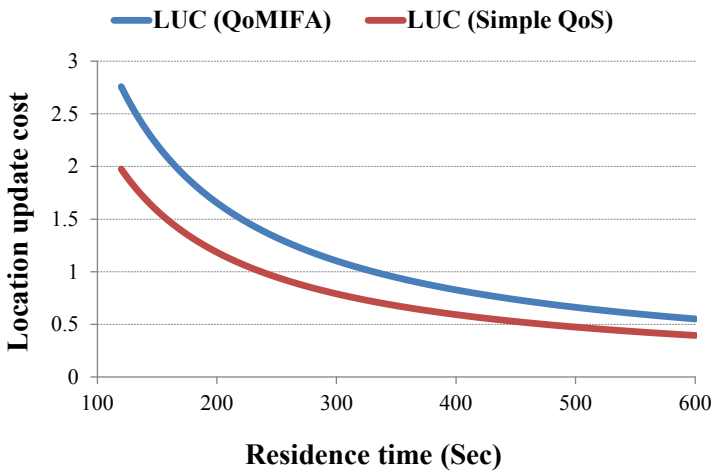


Figure 7.9: Location update cost experienced when employing both studied protocols as a function of the residence time (T_r) when applying the linear mobility scenario

One also notices from both figures that the applied mobility scenario does not have a strong impact on the location update cost. Sure, more mobility scenarios should be considered to strengthen this result.

7.3.2 Packet Delivery Cost

Figure 7.10 displays the packet delivery cost resulting from both studied protocols as a function of the packet arrival rate (λ) when the mesh scenario is applied. The residence time in the figure is 120 sec. Notice that the packet delivery cost depends mainly on the path between the MN and the CN. This cost is not strongly related to the applied movement model since the number of hops between the MN and the CN does not change when the MN changes the point of attachment in the assumed network topology. Therefore, we focus only on the mesh scenario.

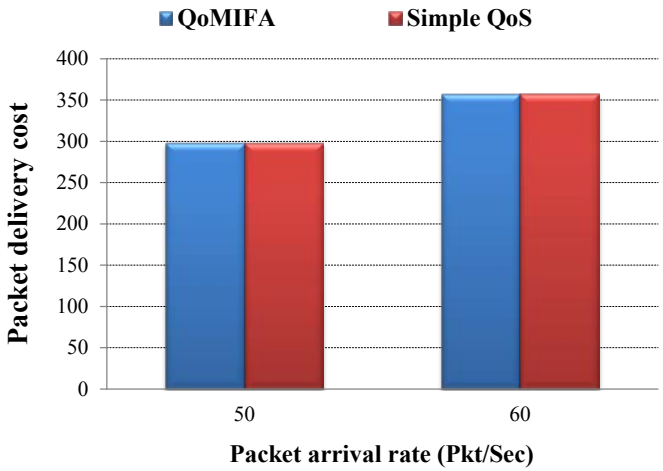


Figure 7.10: Packet delivery cost experienced when employing both studied protocols as a function of the packet arrival rate (λ) when applying the mesh scenario

As shown in the figure, the packet delivery cost resulting from both protocols is approximately equal. According to our results, QoMIFA generates only 0.09 % greater packet delivery cost than Simple QoS. The reason for this is the forwarding of data packets from the old to the new MA during handoffs.

7.4 Conclusion

This chapter provided a comprehensive investigation of the signaling cost resulting from the employment of QoMIFA as compared to Simple QoS. Evaluation metrics comprised the location update and packet delivery cost. The main results of our studies in this chapter can be summarized as follows:

- The location update cost has a negative exponential distribution as a function of the residence time.
- QoMIFA results in a greater location update cost than Simple QoS.
- QoMIFA generates a slightly higher packet delivery cost than Simple QoS.

Chapter 8: Conclusions and Outlook

The chapter is structured as follows: section 8.1 summarizes the main results and section 8.2 discusses issues that can be considered topics for future research built on the results obtained from this dissertation.

8.1. Conclusions

The dissertation has highlighted that the ambitious goals of all-IP networks cannot be reached without overcoming many challenges. An important challenge and also the challenge of interest for this dissertation, is "how can we provide QoS guarantees in such highly dynamic mobile environments." In other words, how can QoS be guaranteed while considering mobility of users? The dissertation showed that the problem addressed necessitates solutions to quickly achieve handoffs, reserve resources after handoffs and release resources not required anymore.

As known, the mobility problem lies in the dichotomy of IP addresses since theses IP addresses indicate the geographical location of MNs. When a MN changes the point of attachment, it will be assigned a new topology-correct IP address. Altering the IP address during ongoing sessions normally results in a communication disruption. Moreover, if services with QoS demands are operated, the availability of required resources should be checked during or even before the handoff.

The dissertation stated that mobility management and QoS problems are handled separately, although they relate to each other. Therefore, the development of new solutions capable of supporting seamless mobility while simultaneously providing QoS guarantees is the main goal of the dissertation.

To highlight the state of art, Chapter 2 briefly reviewed mobility management solutions in IP-based networks. The review concludes: "**network layer mobility management is the most suitable for future all-IP mobile communication networks**". The review presented in Chapter 2 showed that the support of fast and seamless handoffs is crucial for QoS provision, however, not enough. Issues such as the check of the availability of required resources in the new subnet, reservation of these resources, etc. should be considered, as well.

To investigate the QoS issue, a thorough overview of QoS provision mechanisms in IP-based networks in addition to their pros and cons were provided in Chapter 3 . The chapter showed that ***QoS solutions aim at improving the overall performance of the system with the goal of "user satisfaction"***. The analysis of existing QoS mechanisms and architectures showed that the focus is only on QoS guarantees provision with no/minimal consideration of mobility.

After discussing both mobility management and QoS issues separately, the coupling between them was studied in Chapter 4. In brief, there are three basic strategies to couple between mobility management and QoS approaches, namely hard- coupling, loose-coupling and hybrid strategies. While hard-coupled solutions aim at integrating the solutions of both QoS and mobility in a single protocol, loose-coupled techniques keep mobility management approaches separate from QoS mechanisms. Any change in one of them, however, may produce actions in the other. Hybrid approaches attempt to keep the solutions of mobility management and QoS separate from the implementation point of view, yet allow them to work together, so that they look like one protocol. Following the review, a qualitative comparison of the described approaches was provided. The main results of the comparison can be summarized as follows: there have been several prior efforts to couple between mobility management and QoS tech-

niques in a way capable of achieving seamless handoffs simultaneously with QoS guarantees. The main principles of these approaches are either the work proactively, semi-proactively or the localization of mobility and resource reservations inside access networks. These approaches employ, however, mobility management protocols that are stated to be unable to achieve seamless handoffs. The main outcome of the analysis done in the chapter says: the hybrid strategy is promising because it inherits properties of both hard- and loose-coupling strategies. However, further developments of it to perform as well and be as efficient as hard-coupled solutions is challenging and will be of major interest. Facing this challenge was the main motivation behind the development of our new proposal named QoMIFA.

QoMIFA is detailed in Chapter 5. Our proposal integrates MIFA as a mobility management protocol with RSVP as a QoS reservation protocol. MIFA was selected due to its capability of the provision of fast, secure and robust handoffs, while RSVP is chosen because it presents the standard solution used to support QoS in current IP-based networks. QoMIFA is a hybrid protocol. The hybrid architecture is retained by introducing a new object called "mobility object" to encapsulate MIFA control messages within RSVP messages. QoMIFA works proactively, since it further uses the principle of L3-FHRs employed in MIFA. Based on this principle, the current subnet in advance notifies the members of its L3-FHR of incoming MNs, so that these members can in advance check whether they have resources available and, if so, creates RSVP states. Of course, this significantly accelerates the actual allocation of resources after handoffs. ***QoMIFA is capable of achieving fast handoffs and simultaneously reserving resources without wasting network resources, constraining the network topology or introducing new intermediate nodes to the network more than those known from the standard protocol, MIP***. Furthermore, QoMIFA is robust since it provides mechanisms to recover from most failures that may happen such as the loss of QoMIFA support, control message dropping, etc.

A detailed evaluation of QoMIFA compared to Simple QoS was achieved in Chapter 6 by means of simulation studies modeled in NS2. The evaluation focused on studying the impact of network load and MN speed on the performance of both protocols. With respect to the impact of network load, increasing the network load results in increasing the resource reservation latency per handoff, number of dropped packets per handoff, number of packets sent as best-effort per handoff and probability of dropping sessions for both studied protocols. ***QoMIFA performs, however, significantly better than Simple QoS under all studied loads.*** Furthermore, QoMIFA does not send packets as best-effort on uplink. Regarding the impact of MN speed, the impact of ping-pong effects is observed with both protocols and causes more resource reservation latency, dropped packets and best-effort packets per handoff at low speeds than at higher ones. The worst impact of ping-pong effects is seen at a speed of 3 km/h. Our simulation, however, results say: ***QoMIFA clearly outperforms Simple QoS under low as well as high speeds***.

Chapter 7 evaluated QoMIFA from another point of view, namely from the signaling cost point of view. Our results indicated that ***QoMIFA results in greater location update cost and slightly higher packet delivery cost than Simple QoS***.

8.2. Outlook

Based on the topics handled in the dissertation, many suggestions and ideas to improve this work appear. These suggestions are recommendations for future work on the field of mobility management as well as QoS and summarized in the following:

- **Blocking of resources**: as known, RSVP suffers from resource blocking due to the limitation of resources and double resource reservation on the same path for the same session. This problem is also observed when employing QoMIFA due to the use of

RSVP. Although resource blocking problem could be simply avoided by updating all nodes in the access domain, we have decided not to do it so, because the aim is the minimization of RSVP updates to simplify the employment of QoMIFA (i.e. only FAs and the HA should actualize RSVP). Thus, it is interesting to handle the question of: "how the resource blocking problem could be avoided without updating all nodes of the access domain".

- **Further simulation studies**: in our scenarios, the analysis has focused on a unicast hierarchical scenario and studied the impact of network load and MN speed. Further simulation studies considering multicast scenarios are of a great interest. In addition, other performance metrics are also interesting such as the probability of blocking sessions for instance. Furthermore, comparisons to other protocols in addition to Simple QoS are also of interest.
- **Further investigation of signaling cost**: both QoMIFA and Simple QoS cost were analyzed in terms of location update and packet delivery cost. Comparing QoMIFA with a wide range of protocols coupling between mobility management and QoS techniques is interesting. Moreover, other studies such as the impact of the applied mobility model and network topology on the cost are of a great interest.
- **Utilizing the NSIS framework**: NSIS framework is a generalized signaling framework that enables a simple integration of QoS mechanisms. Thus, it is interesting to study the use of NSIS and, more specifically, the use of QoS-NSLP instead of RSVP. This will enable sender- as well as receiver-based reservations in addition to the capability of somewhat supporting mobility in terms of not reserving resources on the whole path after the handoff. Currently, a proposal to couple between MIFA and QoS-NSLP was introduced [AMD08]. This proposal is termed the Mobility management aware next step In Signaling for All-IP Mobile communication networks (MaISAM). This proposal should be specified in detail. Thereafter, comprehensive simulation studies are necessary.
- **The development of QoMIFA in predictive mode**: QoMIFA works until now in reactive mode. The predictive mode will further enhance the performance of QoMIFA. Thus, the analysis of this mode will be a great contribution.
- **The integration of route optimization mode**: QoMIFA forwards data via a triangular route. Extending QoMIFA to operate route optimization mode will further enhance the performance. Therefore, the investigation of this issue will also be a great contribution.
- **The development of QoMIFA for IPv6 networks**: the proposal is developed for IPv4 networks, since IPv4 is a dominant routing protocol in the current Internet and will be a major part of future networks even when widely integrating IPv6. However, the development of QoMIFA to work with IPv6 is also a major contribution.

Bibliography

[3GPP] 3rd Generation Partnership Project, official website: http://www.3gpp.org/, accessed on 16.01.2012.

[3GPP-L] 3rd Generation Partnership Project Long Term Evolution (LTE), official website: http://www.3gpp.org/Highlights/LTE, accessed on 16.01.2012.

[3GPP-TR] 3GPP;Technical Specification Group Services and Systems Aspects: All-IP Network (AIPN) feasibility study (Release 8) ARIB TR-T12-22.978 V8.0.0

[AAg97] D. Awduche, E. Agu, "Mobile Extensions to RSVP", in the proceeding of the 6th International Conference on Computer Communications and Networks (ICCCN '97), pp. 132–36, September 1997.

[ACH96] C. Aurrecoechea, A. Campbell, L. Hauw, "A Survey of Quality of Service Architectures", Multimedia Systems Journal, 1996.

[AMA99] D. Awduche, J. Malcolm, J. Agogbua, M. O'Dell, J. McManus, "Requirements for Traffic Engineering Over MPLS", IETF RFC 2702, September 1999.

[AMB06] E. Alnasouri, A. Mitschele-Thiel, R. Boeringer, A. Diab, "QoMIFA: A QoS Enabled Mobility Management Framework in ALL-IP Networks", in the proceeding of the 17th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'06), Finland, September 2006.

[AMD07] E. Alnasouri, A. Mitschele-Thiel, A. Diab, "Comparative Analysis of T-QoMIFA and HMRSVP", in the proceeding of the IEEE International Symposium on Wireless Communication Systems (ISWCS'07), Norway, October 2007.

[AMD08] E. Alnasouri, A. Mitschele-Thiel, A. Diab, "MaISAM: A New Fast QoS Aware Mobility Management Protocol for ALL-IP Networks", in the proceeding of the 4th International Conference on Wireless and Mobile Communications (ICWMC'08), Greece, July 2008.

[AMD10] E. Alnasouri, A. Mitschele-Thiel, A. Diab, "Handling Mobility Management and QoS Aspects in All-IP Networks", The Journal of Special Issues on Mobility of Systems, Users, Data and Computing, ACM/Springer Mobile Networks and Applications (MONET), January 2010.

[ABG01] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", IETF RFC 3209, December 2001.

[Ami07] M. Amirijoo, "QoS Control of Real-Time Data Services under uncertain workload", Dissertation, Linköpings universitet, Sweden 2007.

[ARe05] M. Atiquzzaman, A. S. Reaz, "Survey and Classification of Transport Layer Mobility Management Schemes", in the proceeding of the 16th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'05), Germany, September 2005.

[ATM96] ATM Forum Technical Committee, "traffic management Specification Version 4.0", ATM Forum Contribution, AF-TM 96-0056.00, April 1996.

[BBa95] A. Bakre, B. R. Badrinath, "I-TCP: Indirect TCP for Mobile Hosts", in the proceeding of the 15th IEEE International Conference on Distributed Computing Systems (ICDCS'95), Canada, May 1995.

[BBC98] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, "An Architecture for Differentiated Services", IETF RFC 2475, December 1998.

[BCS94] R. Braden, D. Clark, S. Shenker, "Integrated Services in the Internet Architecture: an Overview", IETF RFC 1633, June 1994.

[BEB95] R. Braden., D. Estrin, S. Berson, S. Herzog, D. Zappala, "The Design of the RSVP Protocol", USC/Information Sciences Institute, Final Report, June1995

[BFY98] Y. Bernet, P. Ford, R. Yavatkar, F. Baker, et al., "A Framework forIntegrated. Services Operation over Diffserv Networks", IETF RFC 2998, November 2000

[BWK10] A. Bader, L. Westberg, G. Karagiannis, C. Kappler, T. Phelan, "RMD-QOSM: The NSIS Quality-of-Service Model for Resource Management in Diffserv", IETF RFC 5977, October 2010.

[BZB97] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, "Resource ReSerVation Protocol (RSVP) Version1 Functional Specification", IETF RFC 2205, September 1997.

[CCH94] A. Campbell, G. Coulson, D. Hutchison, "A Quality of Service Architecture", ACM SIGCOMM Computer Communication Review, v.24 n.2, pp.6-27, April 1994.

[CCH95] A. Campbell, G. Coulson,D. Hutchison, "Supporting Adaptive Flows in a Quality of Service Architecture", Multimedia Systems Journal, November 1995.

[CHu00] W.-T. Chen, L.-C. Huang, "RSVP mobility support: A signaling protocol for integrated services Internet with mobile hosts", in the proceeding of 15th Conference of IEEE INFOCOM 2000, vol. 3, Tel Aviv, Israel, March 2000.

[Cisco92] Cisco System, "Advanced Topics in MPLS-TE Deployment", 1992, available at: http:// www.cisco.com/warp/public/cc/pd/iosw/prodlit/mwglp_wp.pdf, accessed on 16.03.2012.

[Cla88] D. Clark, "The Design Philosophy of the DARPA Internet Protocols", in the proceeding of ACM SIGCOMM'88, August 1988.

[CTS02] A. Császár, A. Takács, R. Szabó, V. Rexhepi, G. Karagiannis, "Severe Congestion Handling with Resource Management in Diffserv on Demand", in the proceeding of the 2nd International IFIP-TC6 Networking Conference (Networking), Italy, pp.443-454, 2002

[DCB02] B. Davie, A. Charny, J. C. R. Bennet, K. Benson, J. Y. Le Boudec, W. Courtney, S.Davari, V. Firoiu, D. Stiliadis, "An Expedited Forwarding PHB (Per-Hop Behavior)", IETF RFC 3246, March 2002.

[DHi98] S. Deering, R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.

[Dia10] A. Diab, "Mobility Management in IP-Based Networks, Analysis, Design, Programming and Computer- Based Learning Modules", Dissertation, Ilmenau University of Technology, 2010.

[DKS90] Demers, A., S. Keshavt, S. Shenker, "Analysis and Simulation of a Fair Queueing Algorithm", Journal of Internetworking: Research and Experience, pp.3-26, October 1990.

[DMA04] A. Diab, A. Mitschele-Thiel, E. Alnasouri, R. Böringer, J. Xu, "Mobile IP Fast Authentication Protocol" the third Deutsche-Syrische Workshop (DSW'04), Aleppo, Syria, October 2004.

[DMB05] A. Diab, A. Mitschele-Thiel, R. Boeringer, "Evaluation of Mobile IP Fast Authentication Protocol compared to Hierarchical Mobile IP", in the proceeding of 1st Wireless and Mobile Computing, Networking and Communications (WiMob'05), Montreal, August 2005.

[DMK08] A. Diab, A. Mitschele-Thiel, C. Kellner, "A Comparative Analysis of MIFAv6, HAWAII and MIP", in the proceeding of 19th annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'08), France, September 2008.

[EDC92] J., Escobar, D. Deutsch, C. Partridge, "Flow Synchronisation Protocol", in the proceeding of IEEE GLOBECOM'92, Orlando, Fl., December 1992.

[ESTI] European Telecommunications Standards Institute, official website: http://www.etsi.org/, accessed on 16.01.2012.

[ETSI-ETR003] ETSI. Network Aspects (NA); General Aspects of Quality of Service (QoS) and Network Performance (NP). ETSI Technical Report, ETR 003, Second Edition, October 1994.

[ETSI-TR102] ETSI. Satellite Earth Stations and Systems (SES). Broadband Satellite Multimedia IP, IP Interworking over Satellite; Performance, Availability and Quality of Service. ETSI Technical Report, TR 102 157 V1.0.0, March 2003.

[ETSI-TS123 107] ETSI Universal Mobile Telecommunications System (UMTS): Quality of Service (QoS) concept and architecture (3GPP TS 23.107 version 6.3.0 Release 6).

[Fer90] D. Ferrari, "Client Requirement for real-time Communication Services", IEEE Communication Magazine vol.28, no.11, November1990.

[FHu98] P. Ferguson, G. Huston, "Quality of Service: delivering QoS on the Internet and in corporate Networks", ISBN 0-471-24358-2, John Wiley & Sons, Inc, 1998.

[FJP07] E. Fogelstroem, A. Jonsson, C. E. Perkins, "Mobile IPv4 Regional Registration", RFC 4857, June 2007.

[FKK02] X. Fu, H. Karl, C. Kappler, "QoS-Conditionalized Handoff for Mobile IPv6", in the proceeding of the 2nd International IFIP-TC6 Networking Conference on Networking Technologies, Services, and Protocols (Networking'02), Italy, May 2002, available at: http://www.tkn.tu-berlin.de/research/SeQoMo/, accessed on 16.01.2012.

[FMA05] S. Fu, L. Ma, M. Atiquzzaman, Y. Lee, "Architecture and Performance of SIGMA: A Seamless Handover Scheme for Data Networks", in the proceeding of the 40th IEEE International Conference on Communications (ICC'05), South Korea, May 2005.

[FPC05] H. Fathi, R. Prasad, S. Chakraborty, "Mobility Management for VoIP in 3G Systems: Evaluation of Low-Latency Handoff Schemes", IEEE WirelessCommunications, vol. 12, April 2005.

[Fre09] T. Frenzel, "Evaluierung von Mobilitätsprotokollen unter Berücksichtigung von QoS", Diploma thesis, Technical university Ilmenau, 2009.

[FSV97] M. Fry, A. Seneviratne, A. Vogel, and V. Witana, "QoS Management in a World Wide Web Environment Which Supports Continuous Media", Journal of Distributed Systems Engineering, vol.4, no.1, pp. 38-47, 1997.

[FYT97] D. Funato, K. Yasuda, H. Tokuda, "TCP-R: TCP Mobility Support for Continuous Operation", in the proceeding of the 5th IEEE International Conference on Network Protocols (ICNP'97), USA, October 1997.

[GEN01] A. Grilo, P. Estrela, M. Nunes, "Terminal Independent Mobility for IP (TIMIP)", in the proceeding of IEEE Communications'01, vol. 39, no.12, December 2001.

[Gre99] M. Greis, “RSVP ns: An Implementation of RSVP for the Network Simulator ns-2”, 1999

[Gro02] D. Grossman, “New Terminology and Clarifications for Diffserv”, IETF RFC 3260, April 2002.

[GRu05] L. Galindo-Snchez, P. Ruiz-Martnez, “QoS and Micro mobility Coupling”, European Journal for the Informatics Professional (Upgrade), vol. VI, no. 2, 2005.

[HAg97] Z. J. Haas, P. Agrawal, “Mobile-TCP: An Asymmetric Transport Protocol Design for Mobile Systems”, in the proceeding of the 32th IEEE International Conference on Communications (ICC’97), Canada, June 1997.

[Hal96] F. Halsall, “Data Communications, Computer Networks and Open Systems”, printed by Addison-Wesley, 4th Edition, ISBN: 0-201-42293-X, 1996.

[Har01] W. C. Hardy, “QoS Measurement and Evaluation of Telecommunications Quality of Service”, John Wiley & Sons, Chichester, England, 2001.

[HBW99] J. Heinanen, F. Baker, W. Weiss, J. Wroclawski, “Assured Forwarding PHB Group”, IETF RFC 2597, June 1999.

[HKL05] R. Hancock, G. Karagiannis, J. Loughney, S. Van den Bosch, “Next Steps in Signaling (NSIS): Framework”, IETF RFC 4080, June 2005.

[HKZ03] H. Hsieh, K. Kim, Y. Zhu, R. Sivakumar, “A Receiver-Centric Transport Protocol for Mobile Hosts with Heterogeneous Wireless Interfaces”, in the proceeding of the 9th ACM International Conference on Mobile Computing and Networking (MobiCom’03), USA, September 2003.

[HLa97] J.-F. Huard, A. A. Lazar, “On QoS Mapping in Multimedia Networks”, in the proceeding the 21st International Computer Software and Applications Conference (COMPSAC’79), August 1997.

[Hoa03] D.H. Hoang, "Quality of Service Control in the Mobile Wireless Environments", PETER LANG Publisher, Frankfurt lM-Berlin-Bern Bruxelles- New York-Oxford-Wien, ISBN 3-531-50578-7, US-ISBN 08204- 6402-3, 2003.

[HSS99] M. Handley, H. Schulzrinne, E. Schooler, J. Rosenberg, “SIP: Session Initiation Protocol”, IETF RFC 2543, March 1999.

[IEstd] IEEE Computer Society, “Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications”, Standard IEEE 802.11, ISBN: 0-7381-5656-6 SS95708, June 2007.

[IETF] Internet Engineering Task Force (IETF), http://www.ietf.org/

[IETFDiff] IETF DiffServ Working Group, http://datatracker.ietf.org/wg/diffserv/

[IETFNSIS] IETF NSIS working group: http://datatracker.ietf.org/wg/nsis/

[ISO 8402] ISO 8402, “Quality management and quality assurance – Vocabulary”, 1994.

[ISO 9000] ISO 9000, “Quality management systems -- Fundamentals and vocabulary”, 2005

[ITU-TE.800] ITU-T Recommendation. Terms and Definitions Related to Quality of Service and Network Performance Including Dependability. ITU-T Recommendation E.800, August 1994.

[ITU-T-Y.1540] ITU-T Recommendation, "IP Packet Transfer and Availability Performance Parameters", ITU-T Recommendation Y.1540, November 2002.

[ITU-T-Y.1541] ITU-T Recommendation, "Network Performance Objectives for IP-Based Services", ITU-T Recommendation Y.1541, February 2003.

[Jac91] V. Jacobson, "Private Communication", 1991.

[Jac93] V. Jacobson, "VAT: Visual Audio Tool", vat manual pages, Feb 1993.

[Jai95] R. Jain, "Congestion Control and Traffic Management in ATM Networks: Recebt Advances and a Survey", Computer Networks and ISDN Systems, vol. 28, no 13, pp. 1723-1738, February 1995.

[JAk02] X. Jiang, I. F. Akyildiz, "A Novel Distributed Dynamic Location Management Scheme for-Minimizing Signaling Costs in Mobile IP", IEEE Transactions on Mobile Computing Journal, July 2002.

[JPA04] D. Johnson, C. Perkins, J. Arkko, "Mobility Support in IPv6", IETF RFC 3775, June 2004.

[Kes91] S.Keshav, "On the Efficient Implementation of Fair Queuing", Internetworking: Research and Experiences, vol. 2, pp 157-173, 1991.

[KFS12] C. Kappler, X. Fu, B. Schloer, "A QoS Model for Signaling IntServ Controlled-Load Service with NSIS", IETF draft-kappler-nsis-qosmodel-controlledload-13, March 2012.

[LGh90] T.D.C Little, , A. Ghafoor, "Synchronisation Properties and Storage Models for Multimedia Objects", IEEE Journal on Selected Areas on Communications, vol. 8, no. 3, pp. 229-238, April 1990.

[LKL08] S. Lee, M. Kim, K. Lee, S. Seol, G. Lee, "Seamless QoS Guarantees in Mobile Internet using NSIS with Advance Resource Reservation", in the proceeding of 22nd International Conference on Advanced Information Networking and Applications (AINA), 2008.

[LPN07] Z. Lin, L. Peizhen, S. A. Noor, "A QoS extension for Next Step in Signaling in Mobile IPv6", in the proceeding of the 4th international conference on mobile technology, applications, and systems and the 1st international symposium on Computer human interaction in mobile technology, 2007.

[Lucent97] Lucent Technologies, "WavePOINT™-IIAccess Point Wireless LAN Products", 1997.

[LVM01] A. López, H. Velayos, J. Manner, N. Villaseñor, "Reservation Based QoS Provision for Mobile Environments", in the proceedings of the IEEE 1st International Workshop on Services and Applications in the Wireless Public Infrastructure, July 2001.

[Mal07] K. El Malki et al., "Low Latency Handoffs in Mobile IPv4", IETF RFC 4881, June 2007.

[Man03] J. Manner, "Provision of Quality of Service in IP-based Mobile Access Networks", Academic Dissertation, University of Helsinki, Department of Computer Science, Faculty of Science. December 2003.

[Mar07] M. Marchese, "QoS over Heterogeneous Networks", Department of Communications, Computer and System Science University of Genoa, Italy © 2007 John Wiley & Sons, Ltd.

[MIND02] IST-2000-28584 MIND, D1,2, "Top-level architecture for providing seamless QoS, security, accounting and mobility to applications and services", November 2002, available at: http://folk.uio.no/paalee/referencing_publications/ref-auth-istmind-2002.pdf.

[MKM10] J. Manner, G. Karagiannis, A. McDonald, "NSLP for Quality-of-Service signaling", IETF RFC 5974, October 2010.

[MLM02] J. Manner, A. Lopez, A. Mihailovic, et al, "Evaluation of Mobility and Quality of Service Interaction", Computer Networks journal, vol.38, no.2, February 2002.

[Mon01] G. Montenegro, "Reverse Tunneling for Mobile IP, revised", IETF RFC 3024, January 2001.

[MPS01] A.Marquetant, O.Pop, R.Szabo, G.Dinnyes, Z.Turanyi, "Novel Enhancements to Load Control - A Soft-State, Lightweight Admission Control Protocol", in the proceeding of the 2nd International Workshop on Quality of Future Internet Services, Portugal, pp.82-96 Septmber 2001.

[MRa03] J. Manner, K. Raatikainen, "Localized QoS Management for Multimedia Applications in Wireless Access Networks", in the proceeding of the IASTED International Conference on Internet and Multimedia Systems and Applications (IMSA'03), August, 2003

[MSi00] I. Mahadevan, k.M. Sivalingam, "Architecture and Experimental Results for Quality of Service in Mobile Networks Using RSVP and CBQ", ACM/Baltzer Wireless Networks Journal, vol.6, no.3, 2000.

[MYO03] H. Matsuoka, T. Yoshimura, T. Ohya, "End-to-End Robust IP Soft Handover", in the proceeding of the 38th IEEE International Conference on Communications (ICC'03), Alaska, May 2003.

[Nam] Network Animator, official webseite: http://www.isi.edu/nsnam/nam/, accessed on 16.01.2011.

[NS2] Network Simulator 2 (NS2), official webseite: http://www.isi.edu/nsnam/ns/, accessed on 16.01.2011.

[Par03] S. Parameswaran, "WLRP: A Resource Reservation Protocol for Quality of Service in Next-Generation Wireless Networks", in the proceeding of the 28th IEEE Conference on Local Computer Networks (LCN'03), 2003.

[Par92] C. Partridge, "A Proposed Flow Specification", Internet Request for Comments, IETF RFC 1363, September 1992.

[Par92a] A.K.J. Parekh, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks", Ph.D. dissertation, LIDS-TH-2089, Massachusetts Institute of Technology, February 1992.

[Per02] C. Perkins, "IP Mobility Support for IPv4", IETF RFC 3344, August 2002.

[PGa93] A.K.J. Parekh, R.G. Gallager, "A Generalized Processor Sharing Approach to Flow Control In Integrated Services Networks: The Single-Node Case", ACM/IEEE Transactions On Networking, vol. 1, no.3, pp. 344-357, Juni 1993.

[Pos80] J. Postel, "User Datagram Protocol", IETF RFC 768, August 1980.

[Pos81] J. Postel, "Internet Protocol", IETF RFC 791, September 1981.

[Pos81a] J. Postel, "Transmission Control Protocol", IETF RFC 793, September 1981.

[PSt95] G. Pacifici, , R. Stadler, "An Architecture for Performance Management of Multimedia Networks", in the Proceeding of IFIP/IEEE International Symposium on Integrated Network Management, Santa Barbara, May 1995.

[RLT00] R. Ramjee, T. La Porta, S. Thuel, K. Varadhan, “IP Micro-Mobility Support Using HA-WAII”, Internet Draft <draft-ietf-mobileip-hawaii-01>, July 2000.

[RRT01] E. Rosen, Y. Rekhter, D.Tappan, , G. Fedorkow, D. Farinacci, A. Conta, “MPLS Label Stack Encoding”, IETF RFC 3032, January 2001.

[RVC01] E. Rosen, A. Viswanathan, R. Callon, “Multiprotocol Label Switching Architecture”, IETF RFC 3031, January 2001.

[SBa00] A. C. Snoeren, H. Balakrishnan, “An End-to-End Approach to Host Mobility”, in the proceeding of the 6th ACM International Conference on Mobile Computing and Networking (MobiCom’00), USA, August 2000.

[Sch03] J. Schiller, “Mobile Communication”, printed by Addison-Wesley, 2rd Edition, ISBN: 0-321-12381-6, 2003.

[SCM05] H. Soliman, C. Castelluccia, K. El-Malki, L. Bellier, “Hierarchical Mobile IPv6 Mobility Management (HMIPv6)”, IETF RFC 4140, August 2005.

[SHa10] H. Schulzrinne, R. Hancock, “GIST: General Internet Signalling Transport”, IETF RFC 5971, October 2010

[SPG97] S. Shenker, C. Partridge, R. Guerin, “Specification of Guaranteed Quality of Service”, IETF RFC 2212, September 1997.

[SRo99] H. Schulzrinne, J. Rosenberg, “Internet Telephony: Architecture and Protocols – an IETF perspective”, Computer Networks and ISDN Systems Journal, February 1999.

[SSL01] Q. C. Shen, W. Seah, A. Lo, H. Zheng, M.Greis, “An Interoperation Framework for Using RSVP in Mobile IPv6 Networks”, Internet draft, July 2001.

[SWe00] H. Schulzrinne, E. Wedlund, “Application-Layer Mobility Using SIP”, ACM SIGMOBILE Mobile Computing and Communications Review, July 2000.

[SZC04] S. Sharma, N. Zhu, T. Chiueh, “Low-latency mobile IP handoff for infrastructure-mode wireless LANs”, IEEE Journal on Selected Areas in Communication, Special issue on All IP Wireless Networks, vol. 22, May 2004.

[Tan03] A. S. Tanenbaum, “Computer Networks”, 4th Edition, Prentice-Hall PTR, 2003.

[TBA99] A. K. Talukdar, B. R. Badrinath, A. Acharya, “Integrated services packet networks with mobile hosts: Architecture and performance”, journal of Wireless Networks, vol. 5, no.2, pp. 111-124, 1999.

[TBB01] A. K. Talukdar, B. R. Badrination, B. Badrinath, A. Acharya, “MRSVP: A Resource Reservation Protocol for an Integrated Services Network with Mobile Hosts”, The Journal of Wireless Networks, vol. 7, no. 1, 2001.

[TLL03] C-C. Tseng, G.-C. Lee, R-S. Liu, T.-P. Wang, “HMRSVP: A Hierarchical Mobile RSVP Protocol”, in the proceeding of Distributed Computing Systems Workshop, 2001 International Conference, vol. 9, April 2003.

[TMT08] F. Tommasi, S. Molendini, A. Tricco, E. Scialpi, “Fast Re-establishment of QoS with NSIS Protocols in Mobile Networks”, Springer, vol. 5200/2008, 2008

[Tru12] Wikipedia webpage, official webseite: http://en.wikipedia.org/wiki/Trunking, accessed on 17.03.2012.

[TSZ99] A.Terzis, M. Srivastava, L. Zhang, “A Simple QoS Signaling Protocol for Mobile Hosts in the Integrated Services Internet”, in the proceeding of 18th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM’99), 1999.

[Tur86] J.S.Turner, “New Directions in Communications (or Which Way to the Information Age?)”, IEEE Communications Magazine, vol. 24, pp.8-15, 1986.

[TYC05] C.-C. Tseng, L.-H. Yen, H.-H. Chang, K.-C. Hsu, “Topology-Aided Cross-Layer Fast Handoff Designs for IEEE 802.11/Mobile IP Environments”, IEEE Communications Magazine, vol. 43, December 2005.

[Val99] A. G. Valko, “Cellular IP - A New Approach to Internet Host Mobility”, ACM Sigcomm Computer Communication Review Magazine, January 1999.

[VPK03] D. VaIi, S. Paskalis, A. Kaloxylos, L. Merakos, “An Efficient Micro-Mobility Solution for SIP Networks”, in the proceeding of the 46th IEEE Global Telecommunications Conference (GLOBECOM’03), USA, December 2003.

[VPK03a] D. Vali, S. Paskalis, A. Kaloxylos, L. Merakos, “A SIP-Based Method for Intra-Domain Handoffs”, in the proceeding of the 58th IEEE Vehicular Technology Conference (VTC’03), USA, October 2003.

[WAb03] Q. Wang, M. A. Abu-Rgheff, “Integrated Mobile IP and SIP Approach for Advanced Location Management”, in the proceeding of the 4th International Conference on 3G Mobile Communication Technologies (3G’03), UK, June 2003.

[WMa03] Q. Wang, M. A. Abu-Rgheff, “A Multi-Layer Mobility Management ArchitectureUsing Cross-Layer Signalling Interactions”, in the proceeding of the IEE 5th European Personal Mobile Communications Conference (EPMCC'03), UK, April 2003.

[Wro97] J. Wroclawski, “The use of RSVP with integrated services”, IETF RFC 2210, September 1997.

[Wro97a] J. Wroclawski, “Specification of the Controlled-Load Network Element Service”, IETF RFC 2211, September 1997.

[WSK02] L. Westberg, A. Császár, G. Karagiannis, et al., “Resource Management in Diffserv (RMD): A Functionality and Performance Behavior Overview”, in the proceeding of the 7th International Workshop on Protocols for High-Speed Networks (PfHSN’02), Germany, pp.17-34, 2002.

[XGra] Trace Graph - network simulator ns trace files analyser, official webseite: http://www.isi.edu/nsnam/ , accessed on 16.02.2012.