UMinho | 2011

Nuno Vasco Moreira Lopes A QoS-aware Architecture for Mobile Internet

Universidade do Minho
Escola de Engenharia

Nuno Vasco Moreira Lopes

A QoS-aware Architecture for Mobile Internet

Junho de 2011

Universidade do Minho
Escola de Engenharia

Nuno Vasco Moreira Lopes

A QoS-aware Architecture for Mobile Internet

Tese de Doutoramento
Informática

Trabalho efectuado sob a orientação de
Professor Doutor Alexandre Santos
Professora Doutora Maria João Nicolau

Junho de 2011

Acknowledgements

First of all I would like to express my deep gratitude to my supervisors, Professor Alexandre Santos and Professor Maria João Nicolau for their unconditional support during this period of my life. Without them this thesis would never been compiled. They were more than supervisors, they were like my extended family, providing me unconditional support on a scientific level as well as on a personal level in times of more anxiety.

Secondly I would to thank my parents for all the love they gave me and for allowing me to be who I am. My father is my strengh. He has always believed in me regardless of the challenge I choose. My mother is my guide and my consciousness. Her goodness and wisdom gave me valuable life lessons which I try to remember and follow throughout life. My older brother who has always been my partner and protector.

Thirdly I would like thank my wife - my life long companion, for supporting me at home and for being tolerant with all inconveniences caused by my long periods of absense, and most important of all my children whom I am truly proud of, Mariana, Gonçalo e Bárbara for their unconditional love, joy and affection.

I would also like to thank to Professor António Costa for his technical support which has been crucial in many instances especially those related with technical aspects. Another thank you goes out to Professor Joaquim Macedo for his valuable scientific advices.
My sincere thanks to Silvy Ferreira for the thoroughly English revision of my thesis.

Last but not least I would like to thank Fundação para a Ciência e a Tecnologia for the PhD financial support (Grant SFRH/BD/35245/2007).

Dedicated to

My parents, my wife, my children and my brother

Resumo

Hoje em dia, as pessoas pretendem ter simultaneamente mobilidade, qualidade de serviço e estar sempre connectados à Internet. No intuito, de satisfazer estes clientes muito exigentes, os mercados das telecomunicações estão a impor novos e dificeis desafios às redes móveis, através da demanda, de heterogeneidade em termos de tecnologias de acesso rádio, novos serviços, niveis de qualidade de serviço adequados aos requisitos das aplicações de tempo real, elevada taxa de utilização do recursos disponiveis e melhor capacidade de desempenho.

A Internet foi concebida para fornecer serviços sem qualquer tipo de garantias de qualidade às aplicações, apenas se comprometendo em oferecer o melhor serviço possível. No entanto, nos útlimos anos diversos esforços foram levados a cabo no sentido de dotar a Internet com o suporte à qualidade de serviço. Dos esforços desenvolvidos resultaram dois paradigmas para o suporte da qualidade de serviço: o modelo de Serviços Integrados (Integrated Services - IntServ) e o modelo de Serviços Diferenciados (Differentiated Services - DiffServ). Todavia, estes modelos de qualidade de serviço (QoS) foram concebido antes da existência da Internet móvel, portanto o desenvolvimento destes modelos não teve em consideração a questão da mobilidade.

Por outro lado, o protocolo padrão actual para a Internet móvel, o MIPv6, revela algumas limitações nos cenários onde os utilizadores estão constantemente a moverem-se para outros pontos de acesso. Neste tipo de cenários, o MIPv6 introduz tempos de latência que não são sustentáveis para aplicações com requisitos de QoS mais restritos. Os factos revelados, demonstram que existe uma emergente necessidade de adaptar o actual protocolo de mobilidade, e também de adaptar os modelos de QoS, ou então criar modelos alternativos de QoS, para satisfazer às exigências do utilizador de hoje de redes móveis.

Para alcançar este objectivo o presente trabalho propõe melhorias no sistema de gestão da mobilidade do protocolo MIPv6 e na gestão de recursos do modelo DiffServ. O MIPv6 foi melhorado para os cenários de micro-mobilidade com a abordagem para micro-mobilidade

do F-HMIPv6. Enquanto que, o modelo DiffServ foi melhorado para os ambientes móveis com funcionalidades dinâmicas e adaptativas através da utilização de sinalização de QoS e da gestão distribuida dos recursos.

A gestão da mobilidade e dos recursos foi também acoplada na solução proposta com o propósito de optimizar a utilização dos recursos num meio onde os recursos são tipicamente escassos.

O modelo proposto é simples, é de fácil implementação, tem em consideração os requisitos da Internet móvel, e provou ser eficiente e capaz de fornecer serviços com QoS de elevada fiabilidade às aplicações.

Abstract

Over the last few years, several network communication challenges have arisen as a result of the growing number of users demanding Quality of Service (QoS) and mobility simultaneously.

In order to satisfy these very demanding customers, the markets are imposing new challenges to wireless networks by demanding heterogeneity in terms of wireless access technologies, new services, suited QoS levels to real-time applications, high usability and improved performance.

However, the Internet has been designed for providing application services without quality guarantees. That explains why, in the last years several efforts have been made to endow Internet with QoS support. From the developed efforts have resulted two QoS paradigms: Integrated Services (IntServ) which offers the guaranteed service model and the Differentiated Services (DiffServ) which offers the predictive service model.

Although these QoS models have been designed before the existence of mobile Internet, so they do not consider the mobility issue. For instance, the guaranteed service model requires that whenever a Mobile Node (MN) wants to move to a new location, the allocated resources in the old path must be released and a new resource reservation in a new path must be made, resulting in extra signaling overhead, heavy processing and state load. Therefore, if handovers are frequent, large mobility and QoS signaling messages will be created in the access networks. Consequently, significant scalability problems may arise with this type of service model.

The predicted service model, on the other hand, requires an additional features such as dynamic and adaptive resource management in order to be efficient in a very dynamic network such as a mobile network.

A QoS solution for mobile environments must provide the capacity to adapt its resource utilization to a changeable nature of wireless networks because they have a more dynamic

behavior due to incoming or outgoing handovers. For this reason, a QoS signalization for dynamic resource provisioning is necessary in order to supply adequate QoS levels to mobile users.

On the other hand, the current standard protocol for mobile Internet, Mobile IPv6 (MIPv6), reveals limitations in scenarios where users are constantly moving to another point of attachment. In these situations, MIPv6 introduces latency times that are not sustainable for applications with strict QoS requirements.

All things considered, reveal the emerging need to adapt the current standard mobility protocol and QoS models to satisfy today's mobile user's requirements.

To accomplish this goal, the present work proposes enhancements in terms of the MIPv6 protocol mobility management scheme as well as in DiffServ QoS model resource management. The former was enhanced for micro-mobility scenarios with a specific combination of FMIPv6 (Fast Mobile IPv6) and HMIPv6 (Hierarchical Mobile IPv6) protocols. Whereas, the latter was enhanced for mobile environments with dynamic and adaptive features by using QoS signalization as well as distributed resource management.

The mobility and resource management has also been coupled in the proposed solution with the objective of optimizing the resource utilization in a environment where resources are typically scarce.

In order to assess model performance as well as its parametrization, a simulation model has been designed and implemented in the Network Simulator version two (NS-2).

The model´s performance evaluation has been conducted based on the respective data acquired from statistical analysis in order to validate and consolidate the conclusions. Simulation results indicate that the solution avoids network congestion and starvation of less priority DiffServ classes.

Moreover, the results also indicate that bandwidth utilization for priority classes increases and the QoS offered to MN's applications, in each DiffServ class, remains unchangeable with MN mobility.

The proposed model is simple and easy to implement. It considers mobile Internet requirements and has proven to be effective and capable of providing services with highly reliable QoS to mobile applications.

Contents

List of Acronyms

3GPP Third Generation Partnership Project

3GPP2 Third Generation Partnership Project 2

AAAC Authentication, Authorization, Accounting and Charging

AC Admission Control

AG Access Gateway

AN Access Network

AODV Ad hoc On-Demand Distance Vector

API Access Point Identifiers

API Application Programming Interface

ARROWS Advanced Radio Resource Management for Wireless Services

BB Bandwidth Broker

BCMP BRAIN Candidate Micro-mobility Protocol

BGP Border Gateway Protocol

BRAIN Broadband Radio Access for IP-based Networks

BSC Base Station Controllers

BU Binding Update

CIP Cellular IP

CLI Command Line Interfaces

CN Correspondent Network

CoA Care of Address

COPS Common Open Policy Service

DAD Duplication Address Detection

DAIDALOS Designing Advanced Interfaces for the Delivery and Administration of Location independent Optimised personal Services

DiffServ Differentiated Services

DPS Dynamic Packet State

DS Differentiated Service

DSCP DiffServ CodePoint

DSDV Destination-Sequenced Distance-Vector

DSR Dynamic Source Routing

EAC Endpoint AC

EMA Exponential Moving Average

ER Edge Router

EuQoS End-to-End Quality of Service over Heterogeneuos Networks

EURANE Enhanced UMTS Radio Access Network Extensions for NSv2

EV-DO/DV EVolution, Data Only/Data Voice

F-BAck Fast Binding Update Acknowledgment

F-BU Fast Binding Update

F-HMIPv6 Fast Handovers over Hierarchical Mobile Internet Protocol version 6

FA Foreign Agent

FNA Fast Neighbour Advertisement

FTP File Transfer Protocol

GIST General Internet Signaling Transport

GW Gateway

HA Home Agent

HAck Handover Acknowledge

HI Handover Initiate

HMRSVP Hierarchical Mobile RSVP

HTTP Hyper Text Transfer Protocol

IEEE Institute of Electrical and Electronics Engineers

IETF Internet Engineering Task Force

IntServ Integrated Services

ISP Internet Service Provider

ITU International Telecommunication Union

LDAP Lightweight Directory Access Protocol

MAP Mobility Anchor Point

MBAC Measurement-Based AC

MIH Media Independent Handover

MIPv6 Mobile Internet Protocol Version 6

MN Mobile Node

MNF Multicast Non-Forwarding

MRI Message Routing Information

MRSVP Mobile Resource Reservation Protocol

MSC Mobile Switching Centers

MSF Multiple Stream Forwarding

MT Mobile Terminal

nAR new Access Router

ND Neighbor Discovery

NGN Next Generation of Networks

NGWN Next Generation of Wireless Networks

NIST National Institute of Standards and Technology

NMS Network Management System

NS-2 Network Simulator version 2

NSLP NSIS Signaling Layer Protocol

NTLP NSIS Transport Layer Protocol

OTcl Object Oriented Extension of Tcl

pAR previous Access Router

PBMS Policy-Based Management System

PDP Packet Data Protocol

PDP Policy Decision Point

PDR Per Domain Reservation

PEP Policy Enforcement Point

PHB Per-Hop Behaviors

PHR Per-Hop Reservation

PRI Strict Priority

PrRtAdv Proxy Router Advertisement

QNE QoS NSIS Element

QNI QoS NSLP initiator

QNR QoS NSLP receiver

QoE Quality of Experience

QoS Quality of Service

QoSM QoS model

RA Resource Allocator

RA Router Advertisement

RCoA Regional CoA

RED Random Early Detection

RFC Request for Comments

RFM Resource Management Function

RM Resource Management

RMD Resource Management in DiffServ

RSVP Resource Reservation Protocol

RtSolPr Router Solicitation for Proxy

SID Session IDentifier

SLA Service Level Agreement

SLS Service Level Specification

SNMP Simple Network Management Protocol

SSF Single Stream Forwarding

TCA Traffic Conditioning Agreement

TCP Transmission Control Protocol

TDBAC Traffic Descriptor-Based AC

TORA Temporally-Ordered Routing Algorithm

TSW Time Sliding Window

UCBT Bluetooh Extension for NSv2 at the University of Cincinnati

UDP User Datagram Protocol

UMTS Universal Mobile Telecommunications Systems

UNF Unicast Non-Forwarding

VoIP Voice Over Internet Protocol

WFQ Weight Fair Queueing

WG Working Group

WLAN Wireless Local Area Network

WRR Weighted Round Robin

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

This thesis addresses the problem of designing a suitable Quality of Service (QoS) solution for mobile environments. The solution proposed in this thesis deploys a dynamic QoS provisioning scheme capable of supplying predicted services for local and global mobility. The dynamic QoS provisioning encompasses a QoS architecture that uses explicit and implicit setup mechanisms to request resources from the network for the purpose of supporting control plane functions as well as optimizing resource allocation.

In order to optimize resource allocation, the resource and mobility managements have been coupled resulting in a QoS/Mobility aware network architecture able to react proactively to mobility events. Both managements have been optimized to work together in order to support seamless handovers for mobile users running real-time applications.

In general, the proposed solution introduces enhancements in global IP mobility efficiency as well as in the standard DiffServ resource management.

This Chapter is structured in the following manner: section 1.1 presents the motivation for the deployment of a dynamic QoS provisioning solution for mobile environments, section 1.2 presents the objectives and contributions of this thesis, and finally section 1.3 describes the thesis structure.

1.1 Motivation

Today, users want mobility, quality of service and a permanent connnection to the Internet simultaneously. In order to satisfy these very demanding customers, the markets are imposing new challenges on wireless networks by demanding heterogeneity in terms of wireless access technologies, new services, suited Quality of Service (QoS) levels to real-time applications, high usability and improved performance. Heterogeneity is an important issue because of complementary characteristics between different access technologies. For instance, the advantage of 3G cellular networks, such as the Universal Mobile Telecommunication System (UMTS) and the EVolution-Data Only/Data Voice (EV-DO/DV), stems from their global coverage while their disadvantages lies in low bandwidth capacity and high operational costs. In contrast, with Third-Generation (3G) cellular networks, Wireless Local Area Network (WLAN) technology such as IEEE 802.11 exhibit higher bandwidth with lower operational costs and reduced coverage area.

Mobile devices have evolved undoubtedly for a new mobility paradigm whose purpose is to support different radio access technologies. These new mobility paradigms provide the opportunity for new and emerging multimedia services; due to higher usability and better connectivity conditions which has been supplied by actual mobile networks. However, some of these new multimedia services will require quality of service support thus causing a necessity for QoS provisioning in wireless networks.

In an attempt to achieve this goal, the scientific community is making all efforts to provide end-to-end quality of service in the Third Generation Partnership Project (3GPP[1]/ 3GPP2[2]) as well as in the Internet Engineering Task Force (IETF [3]) standards, towards their convergence into the Next Generation or Fourth Generation of Wireless Networks (NGWN/4G).

The principle of the incoming 4G wireless networks is to embrace all wireless network technologies and all interoperability mechanisms thus enabling the mobile user to have

[1] The 3rd Generation Partnership Project (3GPP) scope was originally to produce Technical Specifications and Technical Reports for a 3G Mobile System based on evolved GSM core networks and the radio access technologies that they support (i.e., Universal Terrestrial Radio Access (UTRA) both Frequency Division Duplex (FDD) and Time Division Duplex (TDD) modes). The scope was subsequently amended to include the maintenance and development of the Global System for Mobile communication (GSM) Technical Specifications and Technical Reports including evolved radio access technologies (e.g. General Packet Radio Service (GPRS) and Enhanced Data rates for GSM Evolution (EDGE)).

[2] The Third Generation Partnership Project 2 (3GPP2) is a collaborative third generation (3G) telecommunications specifications-setting project comprising North American and Asian interests developing global specifications for ANSI/TIA/EIA-41 Cellular Radio-telecommunication Inter-system Operations network evolution to 3G and global specifications for the radio transmission technologies (RTTs) supported by ANSI/TIA/EIA-41.

[3] The IETF (Internet Engineering Task Force) is an international community that defines and promotes Internet standards.

seamless movement over different access network technologies while maintaining Internet connectivity with the desired service quality for multimedia applications.

The manner in which different access networks must be inter-connected towards embracing heterogeneity in future networks must be defined in order to select the most appropriate mechanisms for resource and mobility management. There seems to be a general consensus affirming that the interconnectivity protocol will be based on Mobile IP due largely to the fact that the IP is widely deployed as the main Internet protocol [1].

Nevertheless, the current Mobile Internet Protocol version 6 (MIPv6) [2] standard lacks in QoS provisions, robustness and scalability. The MIPv6 is considered a macro-mobility solution and is generally ineffective for handling micro-mobility scenarios; where cell size is small and high frequency handovers are common. In order to overcome this inefficiency, there are a few micro-mobility proposals such as Hierarchical Mobile IPv6 (HMIPv6) [3], Fast Handovers for Mobile IPv6 (FMIPv6) [4], Cellular IP [5] and Handoff-Aware Wireless Access Internet Infrastructure (HAWAII) [6]. It is important to note that neither micro-mobility protocols nor mobile IPv6 support QoS, therefore they are not able to provide any type of QoS guarantees for real-time applications.

The IETF community has been working for some years now towards defining Internet QoS models that are able to meet this need but the task continues to challenge the scientific community. Integrated Services (IntServ) [7] and Differentiated Services (DiffServ) [8] are the primary QoS models developed within IETF. The latter QoS model has been designed to overcome the well known scalability and complexity problems of the former by shifting the complexity and processing load to border routers and keeping core routers as simple as possible. Nonetheless, either IntServ or DiffServ models were developed aiming to provide QoS guarantees in wired networks where user mobility and wireless bandwidth constraints were not a problem [9]. Due to the fact that wireless networks present a more dynamic behavior where cell resource availability is constantly changing due to incoming or outgoing handovers, the mobile user needs some sort of QoS signaling for resource management in order to improve network utilization. Therefore, wireless networks demand a dynamic QoS solution suited to its environment constraints. On the other hand, the existing dynamic QoS models have been designed for fixed networks thus, when enforced to wireless networks, they may become ineffective because they do not consider mobility.

Furthermore, the existing dynamic QoS models work regardless of the mobility management model. This uncouplement produces suboptimal solutions because the resource management is unaware of mobile events. Unlike fixed network environments, wireless networks allow users to change their point of attachment to the network many times during a session. This results in the requirement of QoS renegotiation in the new point of attachment in order to maintain its previous QoS commitments [10]. This renegotiation will introduce extra signaling overhead as well as processing load which in turn may lead to scalability problems in wireless networks .

Therefore, the goal of this work is to design a dynamic QoS Model for Fast Handovers over the Hierarchical MIPv6 (F-HMIPv6) protocol that is capable of providing highly reliable predicted services. In order to accomplish this, a novel resource management scheme for the DiffServ model was coupled to the micro-mobility management of F-HMIPv6. In this approach all routers within the F-HMIPv6 domain are DiffServ aware [11]. This symbiotic combination of components was optimized to work together in order to support seamless handovers with suitable QoS levels for mobile users running real-time applications.

In summary, the proposed model enhances MIPv6 with F-HMIPv6 and enables the F-HMIPv6 with QoS support for different levels of QoS requirements.

1.2 Objectives and Contributions

In general, the objective of this research work is to design and build a QoS-Aware architecture for mobile environments. The adopted strategy has been to develop a QoS architecture for mobile environments based on the DiffServ QoS model as the first approach.

This first approach resulted in a preliminary study where a method for implementing the DiffServ QoS model in the IP core of a UMTS network [12] has been proposed. The performance improvement achieved with this method was evaluated using a simulation model.

In the simulation model, the VoIP, Video, FTP and HTTP traffic models were evaluated in a mobile environment with the Universal Mobile Telecommunications System (UMTS) as radio access network technology. In addition to this, a revenue function for estimating the profits that an Internet Service Provider (ISP) could expect was proposed.

The next step was to advance towards a dynamic and technologically independent QoS-Aware architecture for mobile networks. In this context, the micro-mobility protocols appear as a useful tool in reducing handover latency and losses of the MIPv6.

However, micro-mobility protocols alone do not support QoS requirements for real-time applications. Therefore, the focus was to propose a dynamic QoS architecture for Fast Handovers over Hierarchical MIPv6 (F-HMIPv6) micro-mobility protocol.

In order to achieve this goal, a novel resource management scheme for the DiffServ QoS model was integrated as an add-on into F-HMIPv6 micro-mobility protocol.

Furthermore, in order to provide suitable QoS requirements for real-time applications in wireless networks, the new resource management scheme inter-operates with the F-HMIPv6 micro-mobility protocol to accomplish network congestion control, and dynamic bandwidth reallocations to improve network utilization.

When speaking in terms of the contributions of the model itself, it implements a solution that comprises enhancements in the mobility management of the Mobile IPv6 (MIPv6) as well as in the resource management of the Differentiated Services (DiffServ) QoS model where the mobility management of MIPv6 has been extended with fast and local handovers to improve its efficiency in micro-mobility scenarios with frequent handovers.

In relation to DiffServ QoS Model, its resource management has been extended with an adaptive and dynamic QoS provisioning scheme for improving the use of resources in mobile networks.

Furthermore, mobility and QoS messages were coupled together in an attempt to optimize resource use. This coupling provided a resource management able to react proactively to mobile events by adjusting its resource configurations. The evaluation of the performance improvement of the proposed solution as well as its parametrization was based on a simulation model which has been designed and implemented.

In addition to the evaluation of the model, a study regarding rate estimators was also conducted in order to decide which is most suitable for the global architecture.

In order to delimitate and determine the work field as well as its main objectives, some considerations have been assumed:

- The solution should be designed at layer 3 to be independent from access network technology;
- The solution should support multiple access technologies;
- The solution should consider mobile environment requirements;
- The IP network should be based on the DiffServ QoS model;
- The proposed model should support QoS capabilities suitable for mobile environments;
- The model should provide mobile users with QoS guarantees even during the handover period;
- The QoS architecture should be able to self-adapt its configuration to network conditions;
- A global inter-operation with the existing IP QoS management architectures should be supported;
- The solution should couple the management of macro-mobility, micro-mobility and QoS in order to possess an optimized solution;
- The QoS architecture should be aligned with the current IEFT, International Telecommunication Union[4] (ITU) and 3GPP standards;

Having named the considerations of this research, the first challenges faced in pursuit of the objectives mentioned above have been the investigation of the IP mobility and QoS paradigms as well as the mechanisms used to control bandwidth access particularly, admission control mechanisms and QoS signaling protocols used in dynamic resources reservations. Then, the elaboration of a list numbering the general grounding requirements that Mobile IP places on QoS mechanisms. Next, make a survey of the existing dynamic

[4] The International Telecommunications Union is an organization of the United Nations responsible for information and telecommunication issues. It comprises four sectors - Radio-communication (ITU-R), Standardization (ITU-T), Development (ITU-D) and ITU-TELECOM. The ITU-T sector is responsible for making the telecommunication standards, which are published as ITU-T Recommendations.
http://www.itu.int

QoS provisioning solutions and the evaluation of the existing dynamic QoS provisioning solutions in comparison to the list containing general grounding requirements.

After the accomplishment of this previous work and taking them into consideration, the next challenge was the definition of a Mobility/QoS-aware network architecture solution able to:

- support roaming with the same QoS without having to tear down the active connections and re-establish a new session, this should comprise horizontal and vertical handovers;
- provide dynamic QoS funcionalities for resource management;
- provide adaptive resource management features;
- achieve fast and seamless handovers;
- resolve scalability problems that may arise with the MIPv6 protocol and dynamic QoS architectures;
- fulfill QoS requirements in the new access router for incoming and existing traffic.
- and at long last, implement the simulation model and evaluate the performance improvement of the Mobility/QoS-aware network architecture solution.

Finally, after presenting the research objectives and describing the steps taken to achieve them, the contributions accomplished with this research can be summarized into the following itemized list:

- An evaluation of the most suitable rate estimator for global architecture was achieved in [13].
- A distributed resource management scheme for the DiffServ model that is able to provide a QoS framework solution with the following characteristics: simple; scalable; effective; on-demand admission control and resource management; fast and seamless handover capability; lower state information; and adaptive resource allocation [14].
- A reallocation mechanism based on a hysteresis control method for provide a self-adaptive behavior to access routers was designed and implemented [15].

- A simulation model of the proposed model was implemented in the Network Simulator 2 (NS-2) [16,17].

- A parametrization method for the proposed model was defined and assessed [18].

- The design of a distributed resource management scheme for the Diffserv QoS model. This resource management scheme was added to the FHMIPv6 enabling the micro-mobility protocol with QoS support. This research work also proposes a QoS micro-mobility solution able to provide QoS support for global mobility [19].

- A study that intends to reveal the importance of a QoS architecture based on Diffserv Model has in a mobile environment when the network is facing traffic stressing conditions. The QoS architecture was evaluated in this study in two separate contexts: the purely technical context and the economic context.

1.3 Thesis Structure

This thesis is organized into seven Chapters:

Chapter 1: Introduction

This Chapter provides the reason for the design of a distributed resource management scheme for the FHMIPv6 micro-mobility protocol followed by the objectives and the contributions that have been made.

Chapter 2: Mobility and Quality of Service

This Chapter provides background information related to the research area of the present work. It explores the evolution of mobility protocols, QoS paradigms, QoS mechanisms and the main problems related with supporting QoS in mobile environments.

Chapter 3: Dynamic QoS Provisioning

This chapter presents a survey of the state-of-the-art in dynamic QoS provisioning solutions for fixed and mobile environments.

Chapter 4: Proposed Model

This chapter presents the proposed solution for supporting QoS in mobile environments. It describes the proposed model in detail including its signalization process, resource management function and dynamic allocator behavior.

Chapter 5: Model Implementation in NS-2 and Data Processing

This chapter presents an introduction to wireless models in NS-2 and briefly describes the implementation of the simulation model in this simulator. The Chapter ends with a description of how the simulation data results have been conducted and processed.

Chapter 6: Model Results

This Chapter shows the obtained simulation results and discusses the result analysis. The suitability of rate estimators as well as the performance improvement of the proposed solution as well as its parametrization have been analyzed and evaluated with respect to its applicability to mobile environments in a detailed manner.

Chapter 7: Conclusions

Finally, this thesis concludes with a summary of the research work´s main objectives as well as the contributions to the research in the area of QoS in mobile environments. The Chapter ends by outlining several new directions for future research.

Chapter 2

Mobility and Quality of Service

The fourth-generation (4G) of wireless networks want to embrace all wireless network technologies and all interoperability mechanisms enabling the mobile user to have seamless movement between different access network technologies while maintaining Internet connectivity with the desired quality of service for multimedia applications (voice and video).

The manner in which the different access networks need to be inter-connected in order to embrace heterogeneity in future networks, must be defined in order to select the most suitable mechanisms for resource and mobility management. There seems to be a general consensus in the scientific community that the inter-connectivity protocol will be based on Mobile IP. This is due mainly to the fact that IP is being widely deployed as the main protocol for the Internet [1]. The IP protocol can be also viewed as a connectivity framework where different protocols can work "all over IP and all below IP". The particular feature of IP protocol that allows the inter-operation of its routing mechanisms over different lower-layer protocols and under different upper-layer protocols permits a facilitated extensibility of the Internet network. Consequently, the selection of Mobile IP as the mobility management protocol in order to provide mobility in the Internet is very natural because it has been designed based on the assumption that it is an extension of the basic IP.

During the last years, several network communications challenges have arisen with a growing number of users demanding QoS and mobility simultaneously. The first challenge which aimed to answer those demands consisted in enabling the support of Quality of Service (QoS) to fixed Internet. The second challenge was to endow Internet with mobility support. Nowdays, the challenge is to adapt the existing QoS models to mobile Internet making them suitable for incoming 4G networks although mobile Internet has also introduced its own challenge of becoming effective for micro-mobility scenarios with frequent handovers. Therefore, the current challenge is much more than adapting the existing QoS models to mobile Internet in general. The challenge is to enhance the mobility management of Mobile IP for micro-mobility scenarios and simultaneously design a QoS solution suitable for all types of mobile environments (macro and micro). Unfortunately the existing QoS models were designed for fixed Internet thus, they do not take into account mobility when they perform resource management, resulting in unsuitable solutions for mobile Internet which are unable to deal with the changeable nature of wireless networks [20].

On the other hand, most of the existing solutions for QoS provisioning within mobile environments are grounded on the guaranteed service model. The use of a guaranteed service model for QoS provision in high dynamic networks where mobile nodes are continuously changing their point of attachment has proved to have elevated costs in terms of state information maintenance, signaling, processing and consistency [21–25].

This Chapter briefly describes the Mobile IPv6 protocol as well as the main proposals for its enhancement in micro-mobility scenarios in Section 2.1. Sections 2.2 and 2.3 present the current Internet QoS paradigms as well as their QoS mechanisms for the resource management of the network. Section 2.3 also includes a review of the signaling protocols as well as a classification of different admission control methods. In Sub-section 2.3.3 a more detailed description of the measurement-based admission control scheme components is elaborated. This scheme also exemplifies how the measurement-based admission control scheme components are interconnected. The Chapter ends by presenting the challenges, problems and requirements encountered by the design of a QoS solution in order to function properly in mobile environments.

2.1 Mobility

This section begins with a simple overview of Mobile IP protocol mobility management and continues by describing a few micro-mobility approaches that enhance MIP mobility management.

The mobility management in IP networks encompasses several important issues such as handover management (moving from one access router to another), the passive connectivity and paging (power capacity), as well as the traffic routing (data transmission).

In an IP network the mobile node's connectivity is mainly ensured using MIPv4 [26] or MIPv6 [2] protocols. Accordingly to MIP protocols, when a mobile node wants to move from one access router to another while maintaining its connectivity to Internet, it must undergo a process known as handover. MIP protocols allow an MN to maintain its connectivity to the Internet during handover by proceeding movement detection, IP configuration and location update operations. These three basic MIP protocol operations provide the correct routing of MN's packets to its new point of attachment, regardless of its current location. In order to accomplish these assignments, MIP must maintain a location database which maps the mobile identifier (Care-of-Address - CoA) to a given location information (access router).

By definition, Mobile IP version 4 comprises three operational entities: the Home Agent (HA), the Foreign Agent (FA) and the Mobile Node (MN). The MN entity contains two addresses. The first identifies the MN's home network and the second identifies the foreign or visit network. Whenever an MN reaches a foreign network it acquires a new CoA which must be registered within its HA agent. Once the registration is complete, the HA agent begins to tunnel MN's packets to its new CoA (see Fig. 2.1).

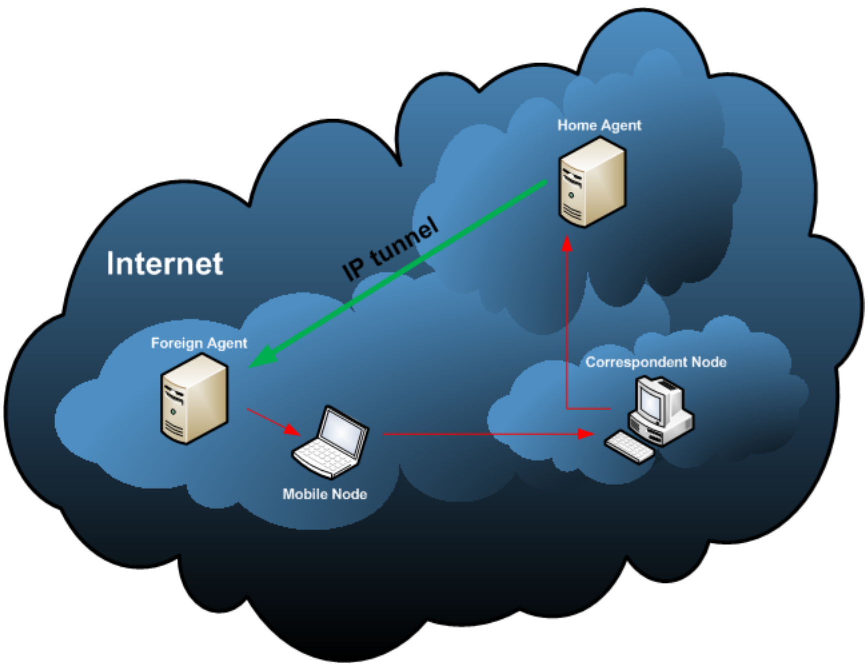


Figure 2.1: Mobility IP Architecture

The way MIPv4 operates results in a triangular routing between the MN and any Correspondent Node (CN). The triangular routing results from the fact that only the HA possesses the actual MN's CoA address thus, the CN must send the packets to HA and then by means of MIP encapsulation, the HA sends the packets to FA. The MIP encapsulation, known as IP within IP (or tunnel), consists of the insertion of a new packet header containing the CoA address as the destination address [27].

Hence, to overcome the triangular routing problem, the route optimization functionality has been proposed. The problem has been resolved by enabling MN with the capability of sending binding updates containing its current foreign network IP address directly to the correspondent node. The CN, is than able to send packets directly to MN without requiring support from HA after it has received the binding updates. By default, routing optimization functionality enhancement is incorporated into the main MIPv6 protocol however, their support remains optional [28]. Generally, handover delay in Mobile IPv6 route optimization is more of a one round-trip time between the mobile node and the CN which corresponds to the MN registration at the CN. This additive handover delay caused by route optimization can significantly affect the quality of interactive and real-time applications which can worsen when CNs are more distant.

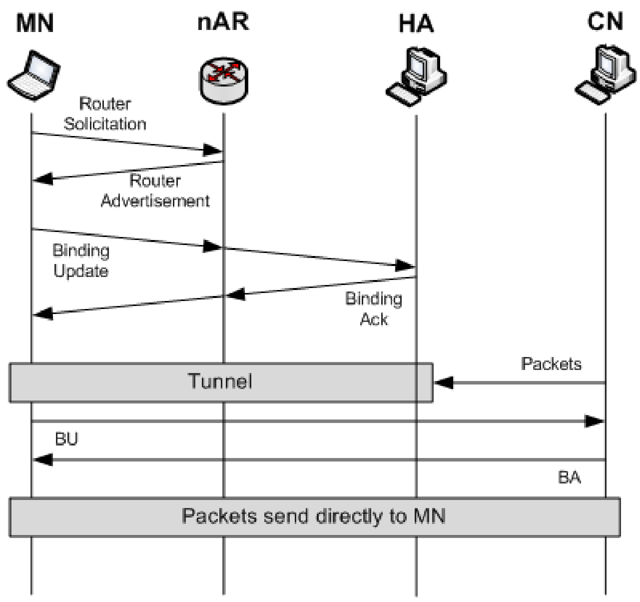


Figure 2.2: MIPv6 Signaling Procedure

Figure 2.2 shows the MIPv6 handover signaling procedure with the routing optimization functionality on. So, when an MN detects a new sub-net through a Router Advertisement (RA) message sent by an access router, it generates a new care of address derived from the information carried in the RA message. An MN may also initiate a handover by sending a router solicitation message to an access router. Due to the fact that the generated CoA must be unique, the MN must execute the Duplication Address Detection (DAD) procedure and only then does it execute the auto-configuration procedure to form the new CoA. After this has been done, the MN registers its new CoA in its HA by sending a Binding Update (BU) message. At this moment, the handover process is completed and the MN may send a BU message to CNs in order to receive the packets directly from them [2].

During this process, the link switching delay and IP protocol operations prevents the MN from sending or receiving packets - this period of time is denominated as handover latency. Handover latency corresponds to the time needed for link switching plus the MIP's operations of movement detection, new Care of Address configuration and Binding Update. Generally, MIP handover latency is unacceptable to real-time traffic such as Voice over IP or video, its reduction is very important to real-time applications and it is also very beneficial to non-real- time traffic.

Over the last few years, a number of micro-mobility protocols have been proposed in an attempt to enhance the main Mobile IP protocol with fast, seamless and local handover control. Micro-mobility protocols are focused on efficient and local mobility support in limited scenarios. These were designed for scenarios where handovers are frequent. Their purpose is to reduce delay, packet loss and overhead signaling. They are particularly important for real-time and interactive wireless applications in order to avoid the noticeable degradation of quality of service in MIP when handovers are frequent. Due to the interest of industries in MIP mobility improvement, several standardization efforts were made within Internet Engineering Task Force (IETF) Mobile IP [29] and Seamoby (concluded WG) [30] Working Groups on latency handover and IP paging.

The following sections describe the most relevant micro-mobility proposals such as the Cellular IP, HAWAII, Hierarchical MIP and Fast MIP. [31–35]

2.1.1 Cellular IP

The main idea behind Cellular IP (CIP) [5] is to integrate the principles of cellular telephony in IP networks. A Cellular IP network is connected to global Internet via a gateway router. The mobility between gateways (i.e., Cellular IP access networks) is managed by a Mobile IP protocol while local mobility is handled by Cellular IP within access networks. The

Gateway (GW) node contains routing entries for all MNs within the network. It is also the unique point of attachment that connects it to the outside of Cellular IP network.

A Cellular IP node (i.e. a base station) constitutes the main component of a Cellular IP network because it serves as a wireless access point and routes IP packets simultaneously. A Cellular IP node integrates cellular control functionalities that were traditionally found in Mobile Switching Centers (MSC) as well as Base Station Controllers (BSC).

Cellular IP nodes are modified IP nodes where standard IP routing is replaced by Cellular IP's own routing and by management functions for mobile node location. The packets are first addressed to the HA which then tunnels the packets to the gateway. At the gateway, the packets are decapsulated and forwarded to the base stations. The MN is identified by its home address and no address conversions are made within the Inside Cellular IP network. The packets sent by a MN are routed first to the gateway and from there towards the CN. These packets are used to establish MN location information in a routing cache. As the packets are sent on a hop-by-hop basis, they pass each base station on route towards the gateway, in this manner the MN location information is recorded in the routing cache of base stations.

The routing cache stores the MN's IP address as well as the interface over which the packets have been received.

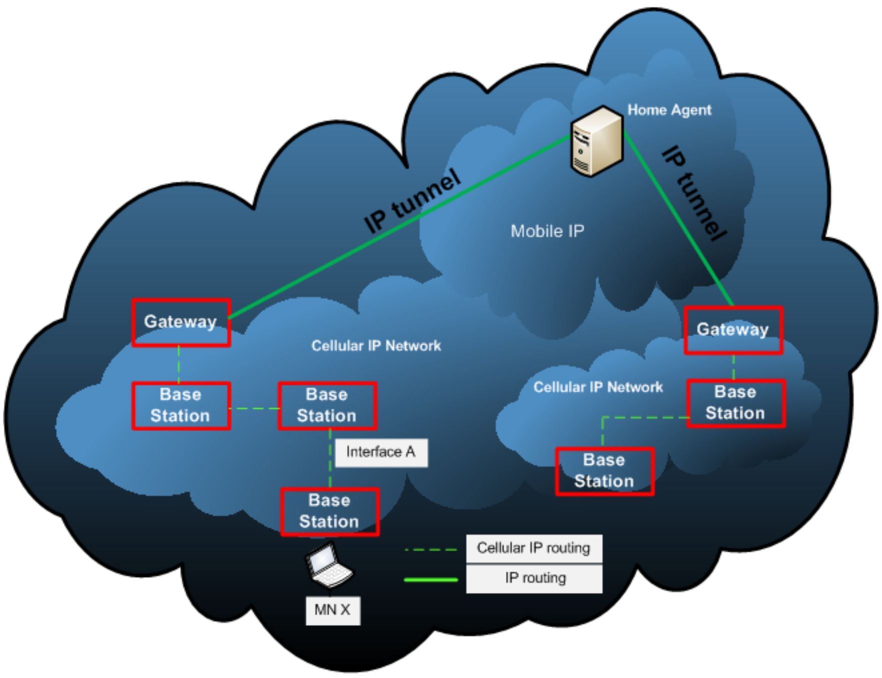


Figure 2.3: Cellular IP Architecture

Figure 2.3 indicates how the packets sent by an MN with an IP address X enters the second base station on the path through the interface A. The entry (X,A) is added to the routing cache of this base station. In doing so, the routing cache receives an entry that contains a reference to the next base station that is closer to the MN. The packets sent from the CN to the MN use the entries created in routing cache with the packets sent by the MN to reverse the route toward the MN location. The packets are forwarded to MN's location inside the Cellular IP network via this chain of base stations.

Cellular IP supports two types of handover schemes: hard and semi-soft handover. Hard handover, assumes support of fast and simple handovers are achieved at the cost of some potential packet losses. In this scheme, when an MN moves to a new base station, it sends a route-update packet to create an entry in the routing cache for mapping the route towards the gateway and in this manner, an MN configures the downlink route to the new base station. Semi-soft handover, prepares a handover proactively by sending a semi-soft update packet to the new base station while it is still connected to the old base station. The semi-soft packet configures the routing cache entry associated with the new base station ensuring that the MN continues receiving packets immediately after moving to the new base station.

An innovation introduced by Cellular IP apart from supporting IP paging is the capability of distinguishing active and idle mobile nodes. Cellular IP maintains a paging cache with the localization of idle hosts, allowing the accommodation of a large number of users attached to the network, without overloading the MN location management in routing cache. Paging systems also help minimize signaling messages and reduce the power consumption of mobile nodes. As a consequence, the battery life is extended and the traffic in the air interface is reduced. When packets need to be sent to an idle mobile node, the MN is paged using a paging packet with a limited broadcast scope. A mobile node becomes active upon the reception of a paging packet and begins updating its location until it moves to an idle state again.

In summary, in Cellular IP access networks the MN location management and handover support are integrated with the routing process. Furthermore, in order to minimize control messages, the packets transmitted by MNs are used to refresh MN location information in the routing cache. The reverse path route (i.e. downlink route) is used to deliver the packets from CN to the MN.

2.1.2 HAWAII

The Handoff-Aware Wireless Access Internet Infrastructure (HAWAII) protocol from Lucent Technologies proposes a separate routing protocol to handle intra-domain mobility [6].

HAWAII contrasts with Cellular IP in the sense that it does not substitute IP protocol. It merely extends the IP funcionalities. Even so, they have some functional similarities such as access routers which maintain a routing cache to handle the mobility as well as the use of a hop-by-hop option to trigger the routing cache updating in the access routers.

In order to provide wide-area inter-domain mobility, HAWAII makes use of Mobile IP protocol. Therefore, when an MN is entering in a new FA domain, it is assigned a new unique care-of-address in order to redirect the MIP tunnel to this location. However, the mobile node retains its care-of-address unchanged while it is moving within the foreign domain. In this context, the HA does not need to be involved.

The routers within the HAWAII network execute the traditional IP routing protocol while maintaining mobility specific routing information which must be added to legacy routing tables. For this reason, HAWAII nodes can be considered enhanced IP routers where the existing packet forwarding function is reused.

Contrary to Cellular IP which uses the data packets sent by MNs to create, update, and modify the MN location information within the routing cache, the HAWAII uses explicit signaling messages sent by MNs. When a mobile node moves within a domain (performing a intra-domain handover), it sends an explicit HAWAII packet so that all the involved routing tables can be modified in order to redirect packets to the mobile's new location.

In HAWAII protocol, the update of the routing entries which corresponds to the MN's new location information can be done resorting to four different schemes denominated as: Multiple Stream Forwarding (MSF), Single Stream Forwarding (SSF), Unicast Non-Forwarding (UNF), and Multicast Non-Forwarding (MNF). Basically, these schemes determine when, how, and which routers will be updated.

The most appropriated scheme is selected based on the operator's requirements such as eliminating packet loss, minimizing handover latency, and maintaining packet ordering.

Just as in Cellular IP, the HAWAII has the inconvenient of only holding a limited number of entries in the routing cache which may result in severe scalability problems [6].

2.1.3 HMIPv6

The Hierarchical Mobile IPv6 protocol uses a hierarchical tree of access routers (ARs) to handle Mobile IP registration locally [3]. HMIPv6 has been designed to reduce the amount of signaling exchanged between MN, its CN, and its HA. The HMIPv6 approach pretends to be a simple extension of the Mobile IPv6 protocol.

A HMIPv6-aware mobile node, with an implementation of Mobile IPv6, should be able to select the usage of the Mobility Anchor Point (MAP) for a local registration when the visited network contains this functionality. It is important to note that in some cases the mobile node may prefer to use the standard Mobile IPv6 implementation. Normally, only one level of MAP hierarchy is considered where all ARs are connected to these MAP. HMIPv6 was designed as a natural extension to Mobile IP to provide efficient micro-mobility support.

The key idea beyond HMIPv6 is to hide MN's movements within a domain to mobility agent at its home network. When an MN moves into a foreign network, it performs a regional registration with a MAP which will be aware of all the local mobility, without needing to inform mobile agents outside the domain. A MAP is essentially a local Home Agent. The main purpose of a hierarchical mobility management model in relation to Mobile IPv6 has been to improve Mobile IPv6 performance while attenuating the impact of the modifications on Mobile IPv6 or other IPv6 protocols. Similar to Mobile IPv6, this solution is independent of the underlying access technology, allowing mobility within or between different types of access networks.

Figure 2.4 shows the HMIPv6's architecture. In HMIPv6, an MN has two addresses associated to him, one is a Care Of Address (CoA) (access router address) and the other is a regional address in the HA. When an MN moves inside a domain, it only has to notify the new AR (nAR) and the local MAP, in order to redirect the local tunnel to the nAR, although HA-MAP tunnel remains exactly the same. This optimization avoids HA registration by hiding local movements from HAs thus resulting in a decrease of registration latency as well as less mobility signaling messages to external networks. The HMIPv6 architecture increases substantially the scalability of the regular MIPv6 handover process. Paging extensions for Hierarchical Mobile IPs are also possible for allowing idle mobile nodes to operate in a power saving mode while located within a paging area. Hierarchical MIPv6 also supports Fast Mobile IPv6 Handovers to help Mobile Nodes achieve seamless mobility.

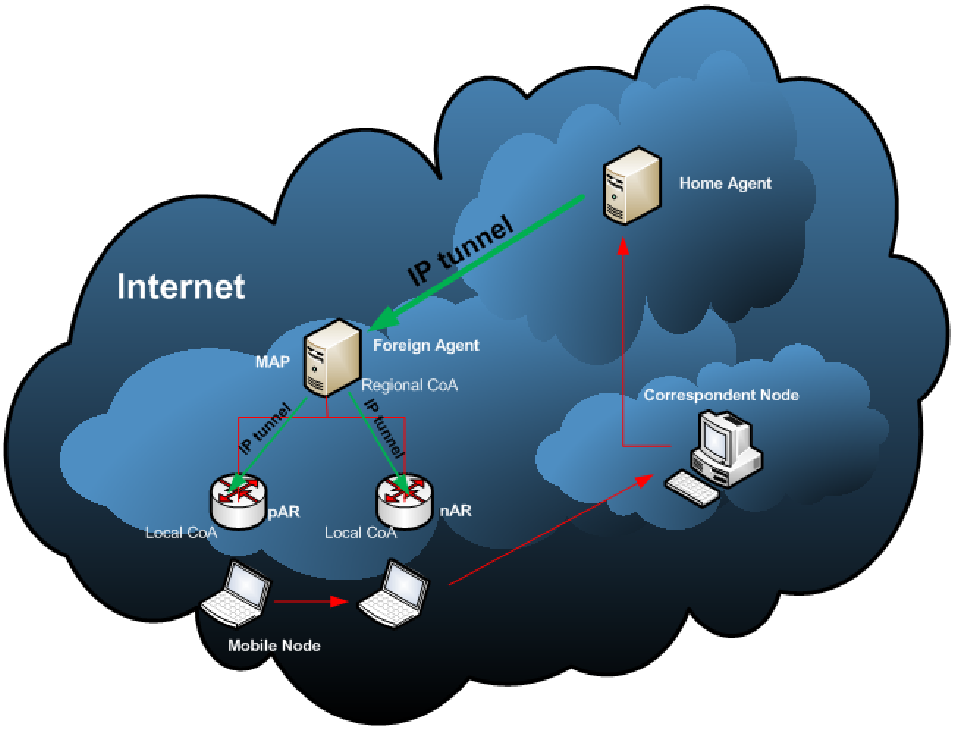


Figure 2.4: HMIPv6 Architecture

2.1.4 FMIPv6

The Fast Mobile Internet Protocol version 6 (FMIPv6) aims to improve handover latency only owing to Mobile IPv6 procedures without addressing link switching latency improvement [36]. The protocol addresses the problem of enabling a mobile node to send packets as soon as it detects a new sub-net link and how to deliver packets to a mobile node as soon as its attachment is detected by a new access router.

Sending packets as soon as possible to a new sub-net link depends on IP connectivity latency, which in turn depends on the movement detection latency and on the configuration latency of the new CoA. The FMIPv6 protocol improves IP connectivity latency by providing a new CoA and its associated sub-net prefix information when a MN is still connected to its current sub-net (see Fig. 2.5).

This mechanism provides the anticipation of layer 3 handover allowing MN's data traffic to be redirected to mobile node's new location before it moves there. The involved access routers and the MN must have the same type of security association established between them for security reasons however, this subject is beyond the scope of the present investigation. The nAR can also rely on its trust relationship with the previous Access Router (pAR) before providing forwarding support for the MN. Which means that the forwarding

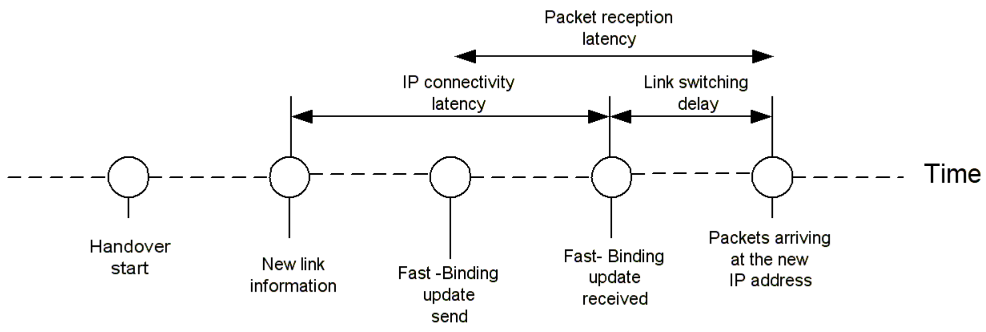


Figure 2.5: IP Connectivity Latency

entry for the new CoA is subject to approval from pAR which it can trust.

Furthermore, the access routers can also transfer network-resident contexts between them such as the access control, QoS, and header compression, in conjunction with handover signaling messages. The transfer of this context information is done by means of Handover Initiate (HI) and Handover Acknowledge (HAck) messages.

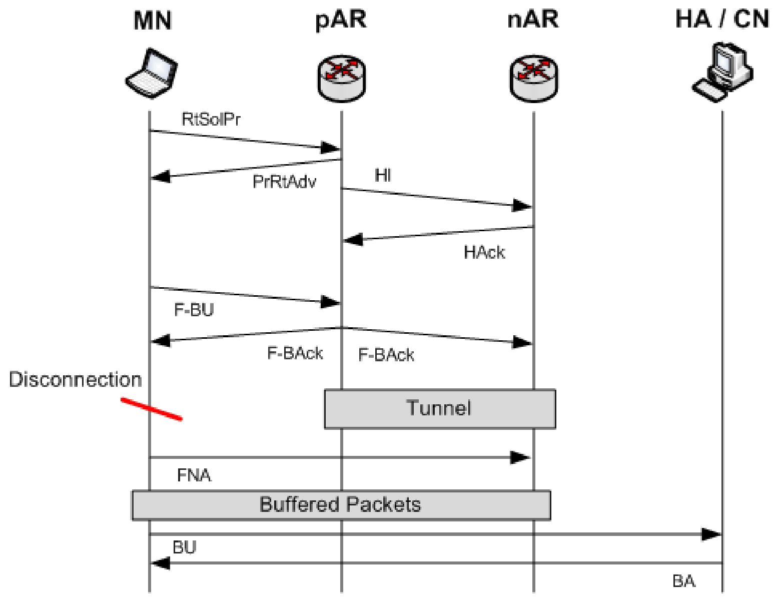


Figure 2.6: FMIPv6 Signaling

Figure 2.6 shows the FMIP signaling process. The handover is triggered when a MN sends a Router Solicitation for Proxy (RtSolPr) message to pAR to resolve one or more Access Point Identifiers (API) to a sub-net specific information. In response, the pAR sends a Proxy Router Advertisement (PrRtAdv) message containing one or more API and a corresponding access router information tuple to MN. The MN formulates a proactive new CoA address with the information provided in the received message. Subsequently, it sends a Fast Binding Update (F-BU) message to pAR for binding previous CoA to new CoA.

After this has been done, packets can be tunnelled to the MN's new location. The Fast Binding Update Acknowledgment (F-BAck) message is sent back to the MN after the pAR validates the new CoA address through the HI and HAck messages. After L2 handover, the MN sends a Fast Neighbour Advertisement (FNA) message to receive the buffered packets in the nAR and performs the registration procedures with HA.

In relation to the correspondent nodes, the MN may send a Binding Update containing its Local CoA instead of its Regional CoA (RCoA), in case they are in the same MAP link. After the correspondent nodes have received a BU, they can deliver the packets directly to the MN.

2.1.5 F-HMIPv6

Despite the fact that HMIPv6 allows the reduction of signaling overhead, it does not provide seamless handovers whereas the integration of both FMIPv6 and HMIPv6 provides seamless handovers with layer 3 (L3) handover anticipation as well as local registration [37,38]. Appendix A of RFC[1] 4140[2] explains how FMIPv6 and HMIPv6 can be integrated to originate the Fast Handovers over HMIPv6 (F-HMIPv6). The integration of FMIPv6 and HMIPv6 has also been proposed in [39].

The integration of both orthogonal protocols allows the reduction of packet losses and registration time. There are two ways in which the Fast Handover Hierarchical MIPv6 mechanisms can be integrated. The first involves placing a MAP in each AR making the ARs function like a MAP entity however, in this case, the existence of the MAP agent seems to be worthless because it degenerates in a process similar to that of the MIP.

The second involves placing the MAP on an aggregation router above all ARs where ARs in the domain could communicate with the MAP. The former case could be inefficient in terms of delay and bandwidth because packets will traverse the MAP-pAR link twice and packets will arrive out of order at the mobile node.

The latter using the MAP on the aggregation router could improve the efficiency of Fast Handovers because MAP could be used to redirect traffic to nAR, thus saving delay and bandwidth between the aggregation router and the pAR.

[1] It is a document describing each Internet protocol prior to being considered a standard
[2] The experimental Request For Comments 4140 (RFC 4140) has been updated by RFC 5380

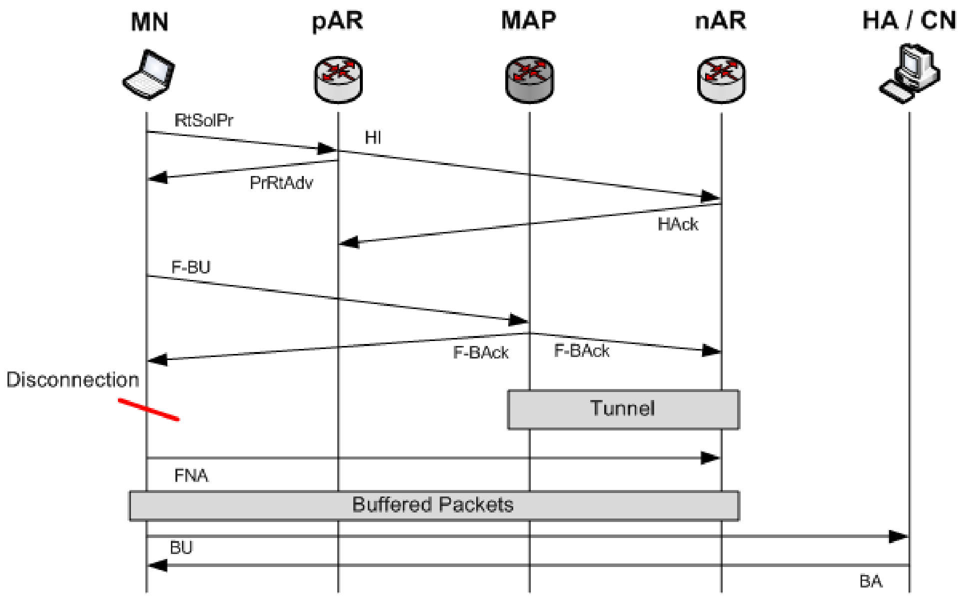


Figure 2.7: Fast Handover over HMIPv6 Signaling Procedure

Figure 2.7 shows the signaling procedure of the association of FMIPv6 and HMIPv6. This association implies the addition of a MAP agent in the communication process of the FMIPv6 protocol. The MAP will then be responsible for MN's registrations, packet tunnels and forwarding signaling messages to nAR on behalf of pAR. In this mobility management scheme, the MAP should only update its binding cache in order to prepare the tunnel after receiving the anticipated F-BU, because it is not aware of the exact moment that mobile node will perform handover to nAR.

Once in the new sub-net link, (on nAR) MN sends an FNA message to receive the buffered packets in the nAR and registers its new CoA with HA and CNs by sending a binding update message.

	Cellular	***HAWAII***	***HMIPv6***	***F-HMIPv6***
Nodes Affected	All CIP	All	AR/MAP	AR/MAP
Address	HA	MN	HA/MN	HA/MN
Operation Layer	L3/2	L3	L3	L3/2
Binding Update	Data packets	Signaling	Signaling	Signaling
Paging	Implicit	Explicit	Explicit	Explicit
Tunneling	no	no	yes (initially)	yes (initially)
L2 Handover Trigger	optional	optional	no	yes
MIP Messages	no	yes	yes	yes

Table 2.1: Comparison of Micro-Mobility Protocols

Table 2.1 summarizes the main characteristics of micro-mobility protocols that have been examined.

Among the micro-mobility protocols described above, only HMIPv6 and FMIPv6 become standard protocols through MIPSHOP Working Group[3], all the others expired while drafts. This can be easily explained because HMIPv6 and FMIPv6 maintain the legacy left by MIPv6 (e.g. IP tunnels), and it was for that reason that these protocols have been selected to enhance MIP mobility management in the scope of this work.

2.2 Quality of Service

The term Quality of Service (QoS) does not have a consensual definition and is often applied incorrectly within the telecommunications universe. However, a well accepted and very general definition is given by ITU in Recommendation E.800 [40] which defines QoS as the degree of satisfaction of a user in relation to the received service. Where the characteristics of service provided to a user must satisfy the stated and implied user needs. In IETF the term QoS is generally used for guaranteed services or when one of RSVP, IntServ, DiffServ or MPLS standards are applied to provide the support of different QoS levels for applications [41, 42].

In contrast to the term QoS, the term network performance can be well defined and measured with mathematical parameters such as packet loss, transmission delay, delay variation (jitter) or throughput.

Another concept which is related to and associated with QoS is the Quality of Experience (QoE). QoE attempts to qualitatively measure or classify the user's degree of statisfation with the service rendered by the Internet Service Provider (ISP) [43]. For instance, to better understand the difference between QoS and QoE, let us consider that a given ISP is delivering a service to a user with the contracted QoS, but the user is still not satisfied with the perceptive service quality, this will result in a low QoE. When a user is using mobile services, the QoE can be a very useful concept because it allows one to measure the effectiveness of the handovers in the network performance from the user's perspective.

[3] Mobility for IP: Performance, Signaling and Handover Optimization (mipshop). The MIPSHOP WG will continue to work on HMIPv6 and FMIPv6, and the necessary extensions to improve these protocols. The MIPSHOP WG will also identify missing components that are required for deploying these protocols and standardize the necessary extensions.
http://www.ietf.org/dyn/wg/charter/mipshop-charter.html

For example, it can be used to measure the influence of the dropped sessions caused by handovers, in the QoE.

Being aware of confusion that prevails with QoS definition, enabling IP networks with QoS support means the employment of QoS control mechanisms in order to provide levels of relative or absolute QoS (also known as, QoS-assured or QoS-guaranteed) to a set of applications.

Addressing the support of QoS for real-time and interactive applications in the IP networks as a global architecture implies a broad design process at different functional planes such as management, control and data planes. Data and control planes are responsible for providing the essential QoS mechanisms to control real-time and interactive traffic based on requested QoS made by applications whereas the management plane is responsible for ensuring that QoS commitments assumed by the network are assured by configuring network resources accordingly to those QoS commitments.

Accordingly to [44], the design of a generalized QoS framework should be based on a set of principles among them transparency, integration, separation, multiple time scales and performance. These principles should dictate the behaviour of a QoS architecture within functional planes. For instance, the distinction between data plane and control plane is mainly characterized by the different time scales under which they operate. These principles serve as guides that will assist the network designers in developing their QoS frameworks. The elements of a generalized QoS framework should also include the QoS specification as well as define whether the QoS management is static or dynamic.

According to literature, there are several QoS architectures proposals for Mobile IP networks however, the research community has not yet decided which is the best solution. Some aspects of which are not consensual are: if the management plane is centralized or distributed, if the dissemination of control messages are in-band or out-of-band, the scalability in large scale scenarios, and the suitability to mobile networks among others.

On the other hand, the micro-mobility protocols have been proposed to overcome the unreasonably high signaling load and latency problems associated with Mobile IP in scenarios where the MN moves frequently within a single administrative domain nevertheless, they do not provide QoS for real-time and interactive applications. Therefore, there is an emerging demand for designing a QoS architecture suitable for those wireless environments. In spite of this emerging demand, the research community has been somehow neglected in finding an appropriate standard for this purpose. Therefore, in order to meet this demand, a suitable QoS architecture for wireless environments needs to be designed. The design of a suitable QoS architecture requires a thorough understanding of the existing QoS paradigms as well as the challenges, problems and requirements that QoS paradigms face in the mobile environment. In order to accomplish these requirements, the remaining part of this Chapter describes the existing QoS paradigms as well as its QoS mechanisms, and presents the problems associated with QoS provisions in wireless IP networks for real-time and interactive applications

2.2.1 Internet QoS Paradigms

In the last few years the Internet Engineering Task Force (IETF) has been working in order to define QoS models for the Internet. However, as stated previously, the defined models have been designed to work well with fixed Internet. Two service class models have been suggested within the defined models. The first is based on reservations and the other on priorities. The IntServ [7] model is based on flow reservation basis, whereas DiffServ [8] is based on packet priority basis.

	Service Class
Integrated Service (IntServ)	- Guaranteed - Controlled Load - Best-effort
Differentiated Service (DiffServ)	- Premium (Expedite Forward) - Assured (Assured Forward) - Best-Effort

Table 2.2: Service Classes defined for IntServ and DiffServ Models by IETF

Table 2.2 shows the classes of services defined by IETF for each model [45–49].

The guaranteed[4] service model provided by IntServ differs from the predictive[5] service model provided by DiffServ mainly in the manner in which throughput and delay bounds are calculated and considered. In the case of guaranteed services, the worst case delay and throughput bound are used, whereas for predictive service the network measures delay and

[4]In the context of quality of service in the Internet the terms guaranteed service, reserved service, absolute service and deterministic service are used with the same meaning.

[5]In the context of quality of service in the Internet the terms better-than-best-effort service, assured service, relative service, probabilistic service and predictive service are used with the same meaning.

throughput to calculate the bounds. A service model consists of a set of service commitments in relation to a request service invoked to the network, for delivering service. These service commitments for real-time and interactive applications in practice correspond to the choice between absolute or probabilistic service. Thus for inherence, an inevitable consequence of this choice is also the choice between IntServ and DiffServ model. In the IntServ model, the service commitments are made to individual flows, these service commitments are mainly focused on delay requirements whereas in the DiffServ model, the service commitments are made to a class of traffic by policing the aggregated bandwidth distributed among the classes, according to some set of specified thresholds shares.

2.2.1.1 Integrated Services Model

The Integrated Service is based on the assumption that resources (e.g., bandwidth) must be explicitly managed in order to achieve the application's QoS requirements and this should be accomplished with resource reservation and admission control as the main building blocks. Contrary to static, QoS provisioning where no explicit setup mechanisms are used, the IntServ provides a dynamic QoS provisioning able to manage the resources´ on-demand basis. In order to make dynamic resource reservations, the router must store the flow-specific state, which represents an important increase in processing and storage load to the Internet model. Accordingly to the IntServ standard [7], the IntServ framework contains four components that should be included in the Internet: the packet scheduler, the admission control policies, the classifier and the signaling protocol (see Fig. 2.8).

The first three components are implemented in the router function. The first, the packet scheduler, manages the packet forwarding of flows using a queuing discipline such as strict priority (PRI), Weighted Round Robin (WRR) or Weight Fair Queuing (WFQ), among others. The second, classifier, is responsible for mapping the incoming packets into a class in order to enable the scheduler to provide the same treatment to packets that belong to the same class. The third, admission control, implements an algorithm that determines whether a new flow QoS request is granted or rejected. Thus, it concerns the enforcement of administrative policies on resource reservations. Additionally, it can also provide accounting and administrative reporting. The last component is implemented resorting to the Resource Reservation Protocol (RSVP) [50]. The RSVP plays an important role in the request for the desired QoS to the network as well as in the creation and maintenance of flow-specific state

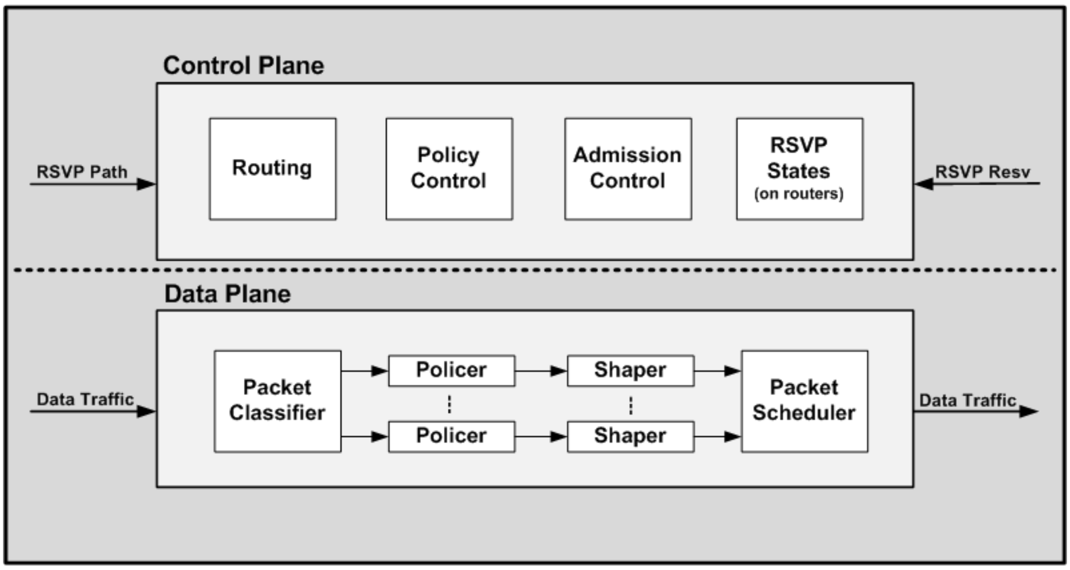


Figure 2.8: IntServ Reference Model

in the endpoints hosts and in all the routers along the data path. The RSVP is based in resource reservations, a RSVP´ reservation specifies the amount of resources to be reserved for all packets in a particular session. A reservation setup consists in the configuration of a scheduler in a way that packets belonging to a particular session should be served within the requested quality of service. Therefore, a resource reservation is the resource allocation along the data path for a particular session to have the same per-hop scheduling behavior in the end-to-end path. This resource allocation quantity is defined by a flow specification (flowspec), while the packets that will receive those resources are defined by a filter specification (filterspec).

Accordingly to the IntServ model, whenever an application starts it must specify the desired QoS using a flowspec message pertaining to the RSVP protocol containing the application QoS parameters. This flowspec is carried out by the RSVP signaling and passed to admission control for a decision test. Considering that admission control has been succeeded, the flowspec will be used to configure the packet scheduler and the filter spec will be used to map the packets into a class.

The RSVP uses receiver-initiation reservation to maintain its consistency in multicast sessions. The receiver generates the desired flowspec towards one or more senders building reservations in all the routers along the path. Each router stores as a cache information all flow-specific states which need to be periodically refreshed by the end hosts in order to remains valid. The unused state is timeouted by the routers or explicitly released. This approach is known as soft-state reservation. The soft-state reservation approach has the disadvantage that if the packet route changes, it needs to install all the states along the new route. Consequently, this results in a strong dependence between the resource reservations and the routing process.

Indeed, an IntServ reservation setup is likely to add a new virtual-circuit Internet layer which somehow is contrary to the fundamental connectionless nature of the current Internet layer and, even without replacing the Internet protocol only extending it with RSVP, the reservation protocol must still coexist with the legacy of the routing protocols in use today in the Internet [7]. These are the main reasons why the IntServ was not widely deployed in the Internet. The RSVP is even more difficult to implement in mobile environments on account of the constant routing changes caused by MN's handovers.

2.2.1.2 Differentiated Services Model

In order to keep track with the continuous increase of Internet users, designing a scalable QoS architecture able to pursue this same increase in terms of the number of users is necessary. The DiffServ QoS model has been defined as the QoS model able to avoid the scalability and complexity problems of the IntServ model, which has been criticized by the research community [51, 52]. Differentiated Services architecture achieves scalability by aggregating traffic with same classification, giving it a particular per-hop forwarding behavior and by pushing up complexity and processing load to border routers, while keeping core routers as simple as possible (see Fig. 2.9).

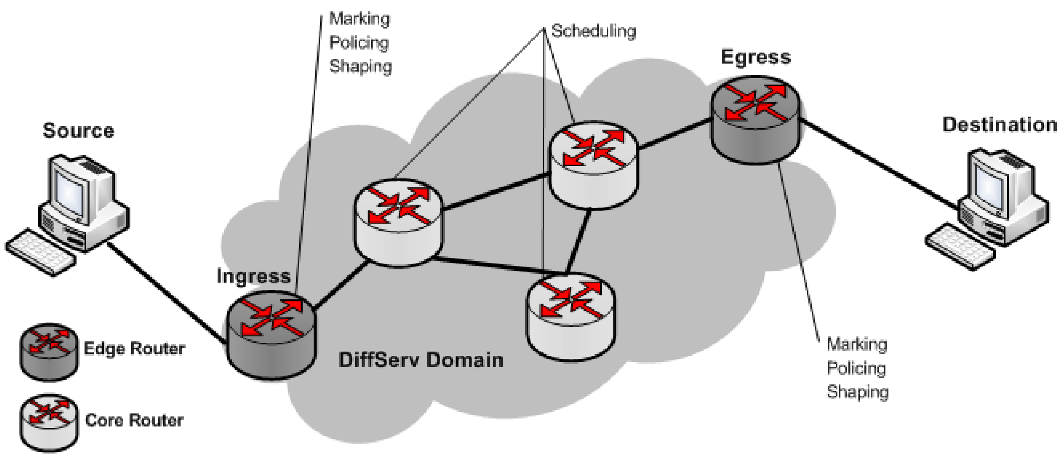


Figure 2.9: DiffServ Router in a Domain

The Differentiated services´ aim is to control bandwidth sharing in a domain based on labeled packets that will receive different priority policies. DiffServ services are implemented by means of classification, marking, policing and shaping operations in edge routers and with per-hop forwarding behaviors in core routers [53]. The resources are allocated to a flow by means of provisioning policies in edge routers which manage and condition the traffic that enters the network and by means of the employment of specific Per-Hop Behaviors (PHB) to traffic aggregates in the interior nodes.

The traffic conditioners, located in ingress and egress boundary nodes, are compounded by meter, marker, shaper, and dropper components. The meter is the content that is measured in the network and is used to supervise if the traffic stream is in compliance with the traffic profile. The measured meter is then used to perform actions such as marking, dropping or shaping to enforce traffic in order to be in conformance with traffic profile contracted (see Fig. 2.10).

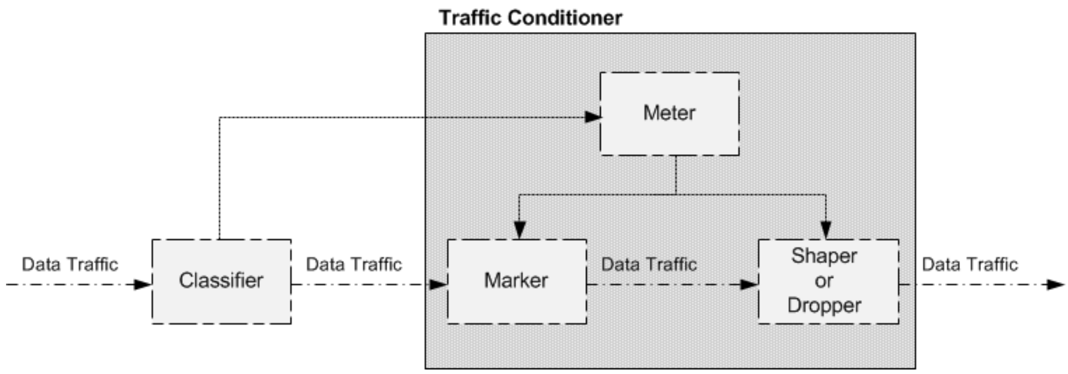


Figure 2.10: Packet Classifier and Traffic Conditioner Block Diagram

PHBs are implemented with buffer management and packet scheduling mechanisms. The selection of per-hop and forwarding behavior is conducted by a classifier which marks the DiffServ CodePoint (DSCP) in a Differentiated Service (DS) field of the IPv4 or IPv6 packets header. The classifier selects the packets from a traffic stream based on one or more header fields, such as source address, destination address, DS field, flow ID, protocol ID, among others.

A DiffServ domain is, by definition, a continuous set of DiffServ-compliant nodes which perform a common service provisioning policy as well as a set of PHB behaviors. Generally a domain consists of one or more networks under the same administration which is responsible for ensuring that the adequate resources for a set of Service Level Agreements (SLA's) is offered in the domain. An SLA is a service contract between a customer and a service provider where the service performance is formally expressed.

Another important concept in DiffServ architecture is a differentiated service region which consists in one or more contiguous DS domains [54]. The Contiguous DS domains allow a service to move across several domains. However, when a service is spanned across several domains, each one must have an SLA describing the Traffic Conditioning Agreement (TCA) which determines how the service should be conditioned from one domain to another. The SLAs contain all the specifications regarding packet classification, re-marking rules and traffic profiles. The traffic profile defines the rules to decide whether a packet is an in or out profile and it can be defined based on a token bucket. For instance, if we assume that a codepoint X uses a token-bucket at a rate of r and a burst size of b, this indicates that all packets marked X should be measured against the rate meter r and burst size meter b.

In RFC 2638 [54] the authors propose the addition of a new entity to the DiffServ model, denominated as Bandwidth Broker (BB), to provide end-to-end dynamic reservations to users in order to deliver end-to-end services with QoS across various domains. This new entity is associated with a domain or an administrative structure that allows the dynamic allocation of flows between two adjacent domains. For this purpose, the BB establishes an association with its peer domain in order to configure the rate and service class of flow across the peer's domain. For instance, whenever a flow allocation is wanted in an adjacent domain, the BB sends a request to the peer BB containing the service type, target rate, a maximum burst and the time the service was required. The BB then verifies if there is sufficient bandwidth to meet the request. If so, the flow is accepted and the flow specification is stored as soft-state information and must be periodically refreshed. Next, the BB as responsible for managing the DiffServ router configuration in the DiffServ domain, needs to configure the edge routers according to the flow specification of the accepted flow.

The control plane of the DiffServ architecture presented in RFC 2475 [8] is responsible for managing, coordinating, monitoring and performing tasks related to the data plane, although it lacks in the provision of the following control functions:

1. Admission Control: verify whether network contains sufficient resources to accept or reject the request of a new flow in order to not affect the QoS of existing ones;

2. Network Resources Management: manage network utilization in order to maximize it while providing the required QoS and;

3. Network Resources Provisioning: responsible for setting up information associated with the required QoS on the edge routers.

	IntServ	DiffServ
Characteristics	- Stateful; - Per-flow management; - Per-flow signaling; - Fair Queueing Congestion Control; - Admission control.	- Stateless; - Per-aggregate management; - RED Congestion Control; - Edge and core Routers; - Per-Hop Behaviors (PHB).
Disadvantages	- Less Scalable; - Less Robust; - More Complex; - Lower resource utilization.	- Less flexible; - Higher resource utilization; - Lower assurance level; - More Scalable.
Advantages	- More flexible; - Higher assurance level	- More Robust; - More simple;

Table 2.3: Comparing both Internet Paradigms

Table 2.3 summarizes and highlights the main characteristics of the described Internet paradigms thus allowing a simple comparison between them.

2.3 QoS Mechanisms

Traffic Control mechanisms allow the construction of efficient traffic management schemes in order to prevent and recover from network congestion as well as provide QoS support to a wide range of service requests. Traffic management comprises a broad range of mechanisms from packet scheduling, buffer management, packet classification, admission control, flow control and signaling protocols, all of which contribute in some manner to the control of resource-sharing or resource-allocation. Both IntServ and DiffServ resort to some or all of these mechanisms to provide QoS support for applications. They basically use the same type of mechanisms however, they are combined, implemented and configured in different ways. In general, what these mechanisms do is to change the packet forwarding in a router, by taking a set of actions, in order to enforce a desired behavior.

The first congestion avoidance mechanism to appear was the Transmission Control Protocol (TCP) which was developed by Van Jacobson in 1986. TCP permits better service under heavy load circumstances. TCP was a necessary and powerful mechanism for the Internet but it was clearly unable to provide quality service in all circumstances. As a result of this, other mechanisms have been developed to fulfill this requirement.

Congestion control algorithms can be divided into two algorithm classes: queue management, which manages queue length by dropping packets when necessary; and scheduling, which characterizes the way, or discipline by which, packets are ordered to be served in routers. These two router mechanisms are tightly interlinked and have complementary features. In the RFC 2309 [55] some suggestions are made to enhance these mechanisms. When speaking of queue management, an active queue management algorithm like Random Early Detection (RED) is recommended to improve Internet performance, instead of traditional algorithms which drop packets only when the buffer is full, (instead RED drops the packets probabilistically), when speaking of packet scheduler, it has been suggested that packets should be classified before being processed, based on one or more packet header fields into several classes to be forwarded accordingly to its class. The simplest approach is the priority scheme, which orders packets by priority where the highest priority always leaves first. An alternative approach is the round-robin scheme and its variants such as WRR and WFQ, which provide a specific minimum of bandwidth to all traffic classes.

An important traffic control mechanism for real-time and multimedia services is the admission control which ensures the commitments assumed with the requested services by controlling the admission of new traffic into the network. An explicit request with a signaling protocol is necessary to inform the application with regards to resource availability as well as the network about the arrival of new flows. Therefore, as a broad range of real-time and multimedia services are envisage for NGWN, these mechanisms will be crucial to provide support to dynamic QoS provisioning architectures in order to provide suitable QoS levels to these multimedia services.

2.3.1 QoS Signaling Protocols

Enabling the Internet with several QoS levels for multimedia services requires explicit admission control procedures consequently, adequate admission control mechanisms as well as signaling protocols for wireless networks must be chosen.

Recently, a significant number of works have addressed how to enhance or create a signaling protocol for resource control. However, the design of the fundamental characteristics for signaling protocols for the guaranteed service model is a challenging task. Characteristics such as state management, signaling message exchanges, sender/receiver-based resource requests and separation of QoS signaling from routing, require a very careful conception and evaluation.

Currently, the main concerns within signaling protocols are the complexity of the protocol, which affects implementations and processing overhead, and the security of the signaling. In the early years, protocols such as the Internet Stream Protocol Version 2 (ST2) [56] and Resource ReSerVation Protocol (RSVP) [50] were designed to be multicast-oriented, introducing increased complexity in its processing. This complexity and scalability has been criticized by the scientific community.

Currently, IETF is specifying a framework that actually describes how two existing QoS architectures, e.g. DiffServ and Universal Mobile Telecommunication System (UMTS) QoS architectures, could inter-operate. This framework has been developed by the Next Steps in Signalling (NSIS) working group [57]. The main goal of NSIS is to provide a general model capable of supporting several signaling applications. Its QoS signaling protocol, denoted as QoS-NSLP (QoS- NSIS Signalling Layer Protocol) [58] is conceptually similar to RSVP but contains additional requirements [59] which include the support of sender/receiver initiated requests, bi-directional requests and support of requests between arbitrary nodes nevertheless, it does not support multicast. The design of QoS NSLP can be compared to the decoupling of RSVP from IntServ architecture. This explains why a distinction is made between the operation of the signaling protocol and the operation of the Resource Management Function (RMF).

2.3.2 Admission Control

The traffic generated by real-time applications demands a certain level of quality of service to perform correctly. These traffic requirements are normally expressed in terms of metrics such as minimum bandwidth, end-to-end delay or maximum packet loss. Thus, when a particular network provides support to such applications, it must be able to admit new applications without affecting the QoS ensured by the network to the existing applications. For this reason, the enhancement of network resources with admission control mechanisms and QoS provision is important. The QoS provision can be deterministic or statistical, depending on the application´s tolerance to violations. Applications that do not tolerate violations require a guaranteed service model and applications which may eventually accept some degree of violations require a predictive service model.

In a wireless network, the admission control can play an important role in determining whether the resources available in the network are able to support the MN QoS requirements during the handover period. Admission Control (AC) schemes across heterogeneous wireless networks based on resource availability, user mobility and user QoS constrains still remains an open issue.

In literature, admission control approaches can be classified according to their localization, either centralized (e.g. Bandwidth Brokers) or distributed, and according to the method used to decide if there are enough resources to accommodate new requests. According to the method used, they can be divided into three categories: Endpoint AC (EAC), Traffic Descriptor-Based AC (TDBAC) and Measurement-Based AC (MBAC).

In EAC the hosts (the end points) send an end-to-end probe packet to the network in order to detect the most congested link in the data path. After having received the probe packet, the flow will be admitted if the congestion level is below its application QoS requirements. The design of such schemes implies higher setup delays while simultaneous probing packets by many sources cause more traffic leading to a situation known as thrashing [60]. This results in a very low bandwidth utilization.

TDBAC is based on the assumption that a traffic descriptor, either deterministic or stochastic, is provided by the application for each requested service, prior to its establishment. This approach uses a priory characterization for incoming traffic and also determines existing traffic behavior. Therefore, an admission control decision is based on history of the past metric, or metrics, which have already been used to request the service. This admission control approach is typically used to provide the guaranteed service model. Such admission control algorithms normally result in low network utilization because the worst-case scenario rarely occurs simultaneously in real traffic [61].

In MBAC, the prior source characterization is only used for new flows while for the existing flows it uses measurements to characterize them. As in wireless networks, the services are typically predictive or soft real-time services an MBAC approach for AC is enough to achieve this kind of service. An MBAC scheme has the advantage of not requiring flow state maintenance and if the measurements are done in aggregate traffic, the processing load and state overhead are not critical [62].

MBAC algorithms assume that the traffic information carried by the traffic descriptor will be directly used for admission tests without further complex calculation. The misrepresentation problem of traffic descriptors in incoming traffic may be insignificant because the algorithm relies on active on-line measurements in order to characterize the established traffic, which helps limit the impact of a misrepresentation. However, a precise individual flow characterization may help enhance the efficiency and robustness of MBAC algorithms [63]. The MBAC mainly differs from the traditional traffic descriptor-based AC in two aspects [64]:

1. the service model is less reliable and;
2. the behavior of existing traffic is calculated by measurements instead of priority traffic characterizations.

On the other hand, measurement-based admission control schemes take advantage of statistical multiplexing. The advantage of statistical multiplexing is easily explained using a simple buffer which is receiving ON-OFF sources as an example. Considering $\frac{1}{\alpha}$ the mean time on the period ON and $\frac{1}{\beta}$ the mean time on the period OFF, then the probability of the source being ON is $p = \frac{\beta}{\alpha+\beta}$. Thus the average rate of each source is pR , where R is the peak-rate of the source. Thus we can conclude that for N sources the total average rate is $\sum p_i R_i$, and the link utilization is,

$$\mu = \frac{\sum p_i R_i}{C} \tag{2.1}$$

Where C is the capacity of the multiplexed link.

Given the equation 2.1, it is now possible to identify the maximum number of sources which can be multiplexed on the link. The equation clearly indicates that the best strategy in terms of bandwidth utilization is the usage of the average bandwidth to meet QoS requests. However, this expression does not consider the statistical fluctuations which may result in packet losses due to the excess of traffic in the link. In order to minimize this problem, the source traffic descriptor should be specified with stricter QoS parameters by assigning a value close to the peak-rate.

An important characteristic of MBAC is the fact that it is in consonance with the DiffServ philosophy as well as with the nature of the wireless networks in terms of the service model they are able to provide: a service model with soft guarantees, that offers a fairly, but not absolutely reliable bound for real-time applications. The absence of absolute QoS guarantees in wireless environments does not result in a problem since most real-time applications in wireless networks have adaptive playback times, thus tolerating occasional packet losses and varying delays. Thereby, MBAC is a suitable admission control scheme that supports soft real-time applications in wireless networks and is also able to provide higher link utilization [10, 65–67].

The robustness of MBAC schemes depends on elements such as estimation error. It is difficult to have a precise estimate due to the stochastic nature of real traffic and system dynamics because the estimation process is based on measures obtained in ongoing flows while the admission decisions are made in flow arrival.The quality of MBAC could be also improved using additional information from the past, though this type of information also results in a less adaptable MBAC.

In summary, the MBAC approach provides a fairly reliable service model that is a viable alternative to guaranteed service in wireless networks. Furthermore, it is able to provide a higher network utilization in an environment where resources are scarce such as wireless networks.

2.3.2.1 Admission Control Policy & Algorithms

According to the following literature, several admission control criteria have been presented and studied [62,68–71]. The admission control scheme may use the rate and/or the delay as input parameters. The parameters may be obtained from a traffic descriptor (parameter-based) or from measurements (measurement-based).

The simple rate sum basically ensures that the sum of request resources, new and admitted, do not exceed the link capacity. This scheme does not provide any additional assumptions regarding the source or traffic behavior process beyond those provided in traffic descriptors. The admission control criteria is simply based on,

$$\nu + r_i \leq \mu \tag{2.2}$$

where ν is the sum of the reserved resources, r_i is the rate requested by flow i and μ is the link capacity. The flow rate r_i parameter is typically obtained by means of the token fill rate of the token bucket. This scheme can provide guarantees to traffic delivery if the flow rate is in constant conformance with the traffic descriptor which is not the case for real traffic. This strict guarantees are achieved at the expense of a lower network utilization due to the fact that flows do not send traffic when they are inactive, however the reserved rate is being considered in the network resource.

In an effort to overcome these drawbacks, the measure rate sum scheme measures the current network load and substitutes the value of variable ν by the estimation of the existing traffic load. When a new flow is admitted, the load estimate is updated to $v + r_i$. However, as the queuing delay typically increases as the system approaches maximum utilization, it is convenient to identify a target utilization below the maximum capacity and keep the admission threshold below this target [62].

In [72, 73] the authors propose an algorithm that uses the equivalent bandwidth of the aggregated traffic for making admission control decisions. The equivalent bandwidth is defined by the authors as bandwidth $C(\varepsilon)$. The bandwidth $C(\varepsilon)$ is an estimate of the arrival rate of a traffic class such that the stationary arrival rate of traffic exceeds $C(\varepsilon)$ with a probability of ε, where ε may represent the loss rate. An admission control decision is based on $C(\varepsilon)$ plus the peak rate p of the new flow which must be less than or equal to the bandwidth allocated (C) for a given class.

Gibbens and Kelly in the works [69, 74] propose a measurement-based admission control algorithm that computes an acceptance region that maximizes the network utilization in detriment of packet loss. The algorithm decides whether or not to admit a new flow based on the current state of the network resources and whether the state lies within the acceptance region or not. The shortcomings of this scheme results from mathematical simplifications of the network model, such as considering that all sources are homogeneous ON-OFF sources. These assumptions limit the applicability scope of the algorithm to real world [75].

Another, measurement-based admission control algorithm is proposed by Jamin et al in the work [62]. This algorithm performs admission control decisions based on the delay and bandwidth constrains. The algorithm provides distinct delay and bandwidth expression conditions for the different kinds of traffic (best-effort, guaranteed and predictive services) in the network. Based on these conditions, the algorithm decides whether or not to admit a new flow into the network.

2.3.2.2 Measurement-Based Admission Control

In this subsection the basic components of an MBAC scheme as well as the relationship between them are described. A measurement-based admission control scheme consists in an admission control unit that controls all components, a measurement process that estimates the current load, a traffic descriptor which characterizes the traffic nature of the flow and finally, an admission policy that delineates the rules by which the admission control unit decides to accept or reject a flow (see Fig. 2.11).

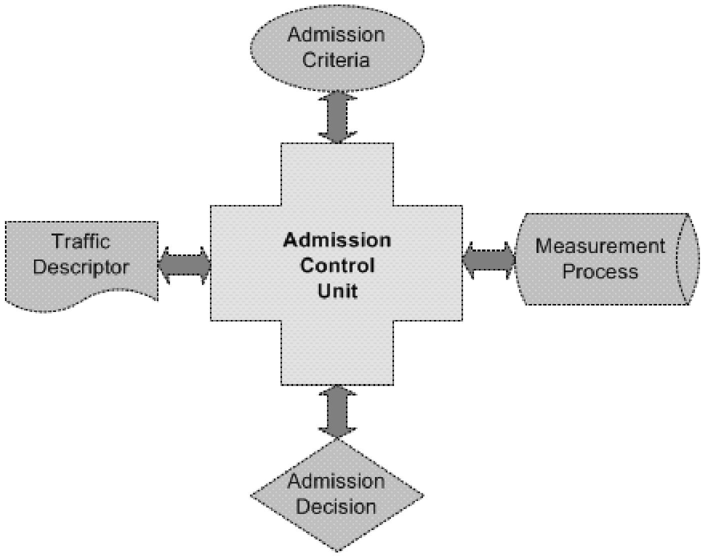


Figure 2.11: MBAC Components

A traffic descriptor consists of an assembly of parameters which specify or characterize the traffic nature of the source. A conventional traffic descriptor is a token bucket [76]. A token bucket consists of a token fill rate "r" and a token bucket size "b". The token bucket size controls the maximum amount of data the flow is able to send at the peak rate. A source specified with a token bucket descriptor cannot send more data than "$r \times t + b$" during any arbitrarily chosen interval t in the life of the source.

An admission control algorithm may possess various parameters as inputs. For instance, if we assume that a traffic descriptor is able to specify its traffic accurately, the admission control unit can only use those parameters to characterize the applications in order to take the admission decision. Nevertheless, real-time applications are extremely difficult to specify due to the multiple time scales´ characteristics in addition to their inherent "chaotic" or bursty nature [72]. For this reason, the admission control unit should monitor the network dynamics´ behavior by conducting real-time measurements of the resource utilization in order to provide the admission control algorithm with the measurement information. Thus, the admission control is able to make a more intelligent decision by knowing the network status[6].

Admission policy implies the conditions by which an admission control scheme rules in order to decide whether to accept or reject flows. As the network resources are shared by all flows in the same class, a decision influences the QoS commitments made to the previously admitted flows of that class. Hence, if the decision is made based on an estimation of the possible effect that the new flow will have on the QoS of ongoing flows, it is possible to admit the new flow and respect the QoS commitments made with the existing ones. The admission control decisions should be signaled to the source node, that if it so wishes, can make the necessary adjustments in the QoS request based on the information received.

2.3.3 Traffic Measurement

In order for a measurement-based scheme to be efficient, it must conduct an accurate measurement of resource utilization in the network. There are many measurement mechanisms described in the literature [62,77]. Each one possesses its own peculiar characteristics which will be reflected in different admission control behaviors.

In the next subsection, two of the most commonly used measurement mechanisms are described: the Time Sliding Window (TSW) and the Exponential Moving Average (EMA), respectively.

2.3.3.1 Time Sliding Window

The time sliding window mechanism measures the current network load during a period of time in order to estimate the current network load so that it can use this estimation as input by an admission control algorithm.

As illustrated in Figure 2.12, the time window scheme samples the network load during each sample period (S) which occurs in every *"S"* units of time. After a measured window

[6] A typical measurement information is the average rate of the aggregated traffic in the flow's class.

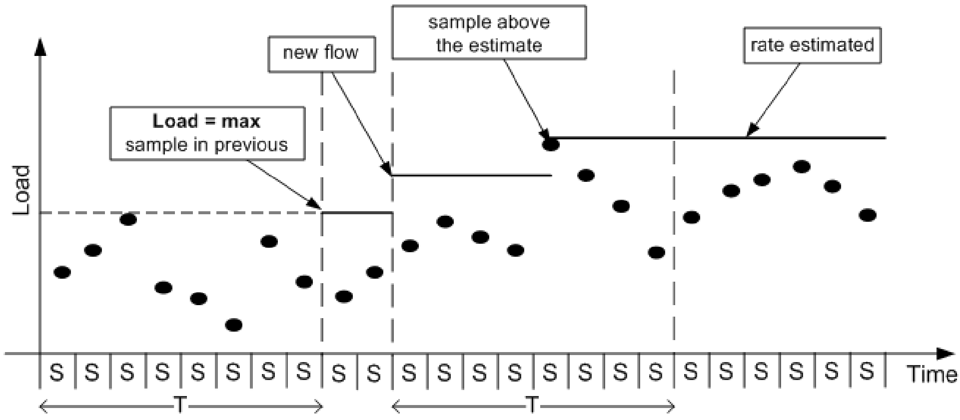


Figure 2.12: Time Window Measurement Scheme

of length *"T"*, it tracks the highest sample load and updates the estimated load with the higher sample load. Whenever a new flow is admitted to the network, the estimation value is increased with the parameter requested by the flow and the window is restarted with the parameter value already reflected in the estimated load [62]. The estimated load is also updated when, after a window restarts, one of the sample values is higher than the current estimate load. The authors of [78, 79] describe how the performance of admission control schemes can be affected by tuning the time windows parameters *"T"* and *"S"*. For instance, smaller values of S represent a higher sampling frequency leading to higher processing overhead thus resulting in a more sensible behavior to burstiness. On the other hand, larger values of "S" leads to a lower sampling frequency resulting in a smoothening of the traffic appearance. Likewise, the size of "T" determines the level of the measurements´adaptability to the variability of traffic. Thus, a small "T" leads to a less conservative admission control and a larger "T" implies a larger memory which permits the reduction of the traffic burstiness effect.

2.3.3.2 Exponential Moving Average

Similarly, to the time sliding window mechanism, the exponential moving average mechanism collects a sample of the network load in every "S" interval to measure the average arrival rate of traffic [62, 72]. The average arrival rate is calculated using an exponential weight function with two terms. The first term represents the weight of the past measurements E_{t-1} and the second term represents the weight of the present measurement O_t in the equation 2.3,

$$E_t = (1 - w) \times E_{t-1} + w \times O_t \tag{2.3}$$

where $0 \leq w \leq 1$ is a constant that determines weight or the importance of historical data and O_t is the measure load at time "t". "w" determines the adaptability of the estimation to the new measurements. A large "w" results in a quick update of the estimation to network dynamics. "t" influences the size of the historical of the past measurements in the measurement process since a little "t" generates frequent measurements. Therefore, the time "t" should be as long as the time interval starting from the point at which the new flow is admitted until the new flow's traffic can be reflected in the measurements. This avoids an overly optimistic estimate of the traffic load because the recently new admitted flows will be taken into account in the future admissions.

2.4 QoS in Mobility

In order to endow the Internet with QoS support, all the layers in the Internet protocol stack must be involved starting by application entities such as Session Initiation Protocol[7] or H.323[8], passing through IP QoS solutions such as IntServ or DiffServ and ending in layer 2 QoS provisioning such as 802.1q [80] or 802.16 [81]. As a QoS based handover management carried out at layer-2 demands a specific strategy suited for each type of wireless access network and one of the objectives of the present research work is to have an independent QoS solution of up/down layers, the QoS handover management will be managed at layer-3. Hence, the present work is only concerned on QoS solutions at IP level.

As stated before the Mobile IP solution was found to be non-optimal for the support of regional mobility within one domain. This explains why several efforts to shorten handover delay have been made by means of micro-mobility protocols. However, handover schemes such as Fast Handover[9], Smooth Handover[10] and Seamless Handover[11], provided by micro-mobility protocols treat different applications in the same manner, without any type of

[7]SIP is a signaling protocol used for controlling multimedia communication sessions such as voice and video streaming over IP

[8]H323 defines the protocols to provide audio-visual communication sessions on any packet network. It is a recommendation from the ITU Telecommunication Standardization Sector (ITU-T)

[9]It is a handover that can comply strict delay bounds

[10]It is a handover that minimizes the lost packets

[11]It is a handover with minimum perceptible degradation of services

network differentiation. Moreover, the existing QoS models do not take into account mobile users. Consequently, QoS support during the handover period still remains unresolved.

Supporting QoS during handovers is challenging due to changing routes between endpoints and varying link characteristics when connecting to different access points. Providing dynamic QoS provisioning during handover in such critical conditions imposes a re-negotiation of QoS parameters in the new access router with an architecture that is aware of the current mobility context and QoS. The QoS context may be transfered to a new access router to be subject to some resource management by means of contexts transfers [82, 83]. But only transfer the QoS context is not enough because mobility and QoS management schemes still remain working independently producing non-optimal solutions in terms of signaling and processing load as well as handover latency.

Context transfers are a very useful functionality that provides support to QoS handovers in IP networks. This functionality is helpful in the support of seamless handovers because it permits QoS re-establishment in the new access router by transfering MN QoS context from one router to another without the need of establishing the QoS in the new router from scratch.

Another important issue in QoS for mobile environments is the type of service model. For instance, in QoS architectures based on guaranteed service models whenever an MN moves to a new location it must release the previously allocated resources in the old path and make new resource reservations in the new path resulting in extra signaling overhead and heavy processing and state load. If the handovers are very frequent large signaling loads of mobility and QoS will be created in the access networks. In consequence, significant scalability problems arise with this service model. Moreover, given the unpredictable nature of wireless links, it is hardly possible to provide absolute guarantees in mobile networks.

On the other hand, if the QoS architecture is based on the predicted service model, additional features such as dynamic QoS funcionalities for resource management and adaptive resource management must be implemented in order to provide efficient resource management for high dynamic mobile networks. Thus, traffic management mechanisms such as admission control that decide whether a router can accept or reject a flow, bandwidth reallocation and signaling protocols are necessary.

When speaking of admission control; in fixed networks, the admission control decision is for new flows exclusively, whereas in wireless networks, the decision is made for new flows as well as handover flows. Since forced call termination due to handover has profound impact on network reliability and user quality perception, the admission control policies should consider the specificities of handover flows [66].

In order to enable QoS handover support to MIPv6 an optimized mobility management scheme with Fast and Smooth handovers is fundamental. The Fast handover scheme provides the anticipation of layer 3 handover allowing data traffic to be efficiently redirected to a new access router before it moves to there. The hierarchical mobility management model allows the performance enhancement of Mobile IPv6 with local bindings while using Fast Handovers allows MNs achieve seamless mobility.

Another important feature for a QoS framework solution in mobile environments is its adaptation capacity to the changeable nature of wireless networks. The wireless networks possess a more dynamic behavior and cell resource availability is constantly changing due to incoming or outgoing handovers. For this reason, supplying adequate QoS levels to MNs on a given cell implies that the user mobility requires a QoS signalization for dynamic resource provisioning. As a consequence this involves the use of two important mechanisms: an admission control mechanism to avoid excess of data and a signaling protocol to request the desired service and also to inform the requesters about the network elements decision/conditions.

2.5 Conclusion

Mobile users want to be permanently connected to Internet and receive services with appropriate quality for real-time and multimedia applications. The standard protocol for mobile Internet (MIPv6) was not found to be optimal in supporting regional mobility within one domain. Firstly, it generates significant signaling traffic in a core network with mobile node movement and secondly, it creates a significant delay in the diffusion of mobile node localization updates. In order to overcome MIPv6 inefficiency in micro mobility scenarios, some efforts to shorten handover delay have been made by a few number of micro mobility protocols. Micro mobility protocols improve the network overall QoS by reducing delays and losses, though their implementations are not enough to ensure different QoS levels to applications with distinct QoS requirements.

The current mobility schemes treat all applications as best-effort traffic, not providing adifferentiated service to real-time and multimedia applications. Therefore, the development of QoS-enabled mobile networks for supporting real-time applications with quality of service is necessary.

However, handover imposes several critical conditions in QoS provisioning. Thus, an important feature of a QoS framework solution in wireless environments is the interaction of QoS management with mobility management whenever a handover occurs in order to have an optimized mobility/QoS-aware solution.

Another important feature for a QoS framework solution in mobile environments is its adaptation capacity to the changeable nature of wireless networks. The wireless networks possess dynamic behavior and cell resource availability is constantly changing due to incoming or outgoing handovers. Hence, in order to supply adequate QoS levels to MNs on a given cell, user mobility requires QoS signalization for dynamic resource provisioning. As a consequence, this involves applying an admission control mechanism for avoiding excessive data charges and a signaling protocol for requesting desired service and informing requesters about the network resources decision/conditions.

In conclusion, the problems and main requirements identified in this Chapter in order to support QoS in mobile Internet will be taken into account in designing the proposed QoS model, which will be later presented in Chapter 4.

Chapter 3

Dynamic QoS provisioning

As the Internet becomes increasingly sophisticated, it is expected that some portions of its resources will be set statistically and others dynamically. The ability to support these choices simultaneously will be important for the purpose of scalability as well as allowing ISPs to offer the user a wide range of offers and prices.

The standardized QoS models, IntServ [7] and DiffServ [8], make use of both dynamic and static managements. In the case of the IntServ model, resources are dynamically provisioned by RSVP [50] while in the DiffServ model, resources may be statically provisioned or dynamically provisioned by Bandwidth Brokers [54]. In dynamic networks, the QoS mechanisms are configured according to flow specification, resource availability and resource management policy.

QoS architectures with dynamic QoS provisioning make use of explicit setup mechanisms to request resources from the network. The network responds to these requests by means of explicit admission control decisions. These explicit setup mechanisms will permit a more efficient and optimized network resource allocation.

The admission control in DiffServ is provided statically in edge routers with policing parameters. The static admission control can be quite ineffective. If we consider, for instance, a DiffServ class provisioned with 30Kbps and 6 sessions, each one requiring a flow rate of 10Kbps, the end result is that it is probable that all 6 sessions will not obtain a satisfactory service when in fact there are sufficient resources available to fully satisfy three sessions. In the case of explicit QoS signalization, the network will signal the rejection of three of them informing the rejected applications with the necessary information to proceed with the adaptive actions if they wish to do so.

This Chapter presents a review of dynamic QoS provisioning solutions for wired and wireless networks. It aims to identify the main characteristics of current dynamic QoS provisioning solutions in the context of their applicability in Mobile IP networks. Then, various projects concerning QoS-aware mobility are briefly mentioned. The Chapter finalizes with a conclusion referent to solutions mentioned in terms of their suitability in wireless environments.

3.1 Dynamic QoS Solutions for Wired Networks

The common process to make traffic commitments in the Internet is the allocation of resources along the data path. This allocation can be strictly preallocated or dynamically allocated and should be capable of encompassing the entire bandwidth in any mix of traffic. The allocation can also be processed in order to provide deterministic or probabilistic services.

The most relevant dynamic QoS provisioning frameworks proposed for fixed networks are discussed within this section.

3.1.1 IntServ over DiffServ

The IntServ over DiffServ framework [84] pretends to take advantage of the best characteristics of both models by integrating the RSVP signaling as well as the models´ traffic control mechanisms. This way, the framework attempts to achieve scalability with the DiffServ aggregation in the network core while maintaining the advantages of end-to-end signaling.

The QoS models in this framework operate in a complementary manner where IntServ provides per-flow resource requirements along the path and responds to the admissibility of requests, while DiffServ provides the necessary scalability for large networks. In order to have seamless inter-operation, the framework treats DiffServ regions as virtual links connected by IntServ routers or end hosts. In the DiffServ regions, the IntServ services are mapped into DiffServ classes.

Figure 3.1 shows an IntServ over DiffServ network representation. In this framework, the IntServ can operate over DiffServ regions in three different ways: the first in which

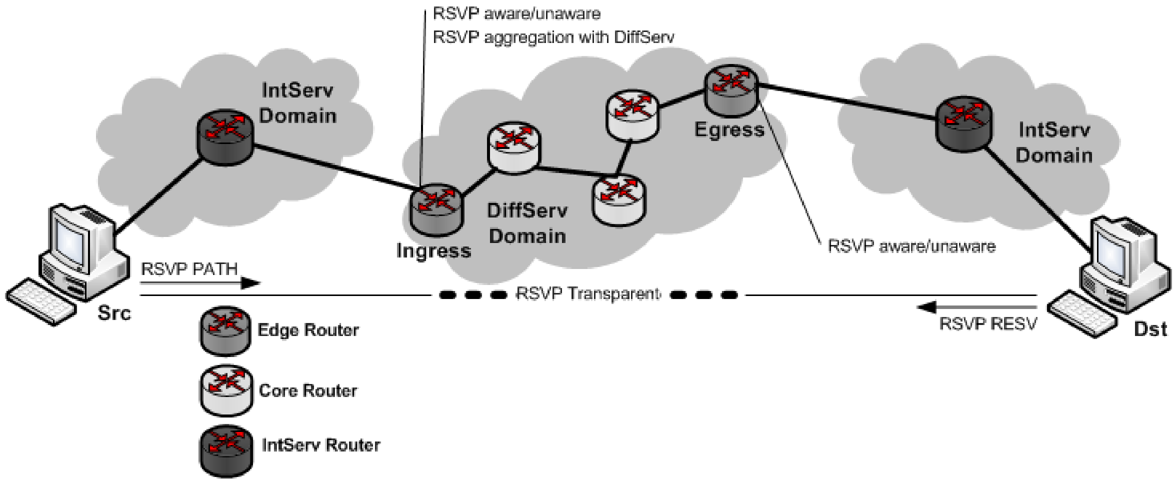


Figure 3.1: IntServ over DiffServ

DiffServ routers are RSVP unaware, the second where the ingress border router is RSVP aware, and the third in which the ingress border router may use any form of signaling in the DiffServ region to interact with an agent possessing all the information regarding the Diffserv region´s resource availability such as a bandwidth broker entity which acts as a centralized admission control.

For example, assuming that a customer and the owner of a DiffServ network have negotiated a static contract (or SLS[1]) for a given DiffServ service level, the service level is defined by factors such as burst size, delay and peak rate, among others. After the contract has been agreed upon, the user has the following end-to-end QoS process sequence to use the specified service:

1. The Src (Source) host sends an RSVP PATH message containing the traffic description.

2. The PATH is carried out towards the Dst (destination) host. The standard RSVP processing and RSVP state is applied within the IntServ region. At the DiffServ edge router, the standard process is also applied if the router is RSVP aware. After this, the PATH message is sent onward to the DiffServ region.

3. When the PATH message reaches the Dst host, it generates an RSVP RESV message.

4. The RSVP RESV message is sent back to the Src host. The message can be rejected at any RSVP capable node along the end-to-end path. In the DiffServ edge node,

[1]SLS - Service Level Specification

the IntServ service is mapped into a DiffServ service and the corresponding policy decision is applied based on the negotiated SLS.

5. At Src, the RESV message is interpreted as an indication that the specified flow service has been admitted. The Src may also mark the DSCP within the packet headers of the admitted traffic flow.

In summary, the framework attempts to achieve per-flow end-to-end QoS provisioning across DiffServ regions.

3.1.2 RSVP Aggregation

The RSVP aggregation solution [85] enhances the standard RSVP protocol with the aggregation of individual RSVP reservations into classes across certain regions denominated as "aggregation regions"[2]. The solution addresses the problems caused by individual reservations in terms of the amount of signaling messages, processing load, and memory which are required for making these individual reservations in all routers along the data path.

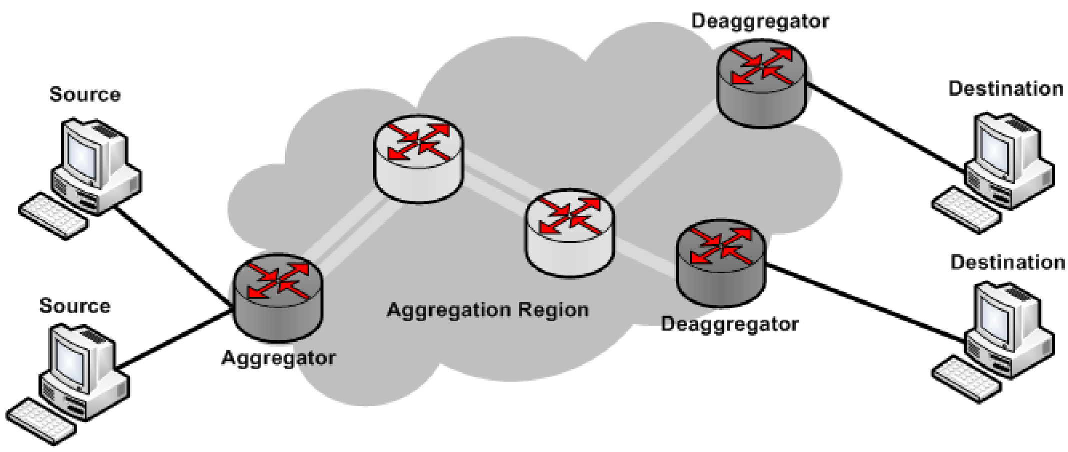


Figure 3.2: Aggregation Region

Figure 3.2 shows the elements of an aggregation region. The solution presents many similarities with the IntServ over DiffServ framework. The main difference resides in the fact that the RSVP aggregation, apart from simply marking traffic with a DSCP for scheduling and classification operations, also installs one or more aggregate reservations in ingress (aggregator) and egress (deaggregator) border routers thus enabling the reservation state to be aggregated by means of DiffServ classification in an "aggregation region".

[2]An aggregation region is a contiguous set of systems capable of performing RSVP aggregation

3.1.3 Scalable Core (SCORE)

The main idea behind the SCORE solution [86,87] is conveying the state information in the data packets within a DiffServ domain. This technique, for conveying state information, was denominated as Dynamic Packet State (DPS). A DSP packet is used for coordinating the actions between edge and core routers. The SCORE solution adds a new QoS scheduling scheme as well as an admission control algorithm to the DiffServ QoS model.

In a SCORE network, the QoS scheduling and the admission control are implemented with two distributed algorithms. The former algorithm is for data plane and is based on a Virtual Clock algorithm [88], and the latter algorithm is for the control plane and is based on a MBAC algorithm.

The state is initialized in the ingress node and is updated along the data path by core routers. The core routers process each incoming packet and update the packet state using the state information carried in packet header and its internal state. The DPS packet is then forwarded to the next hop.

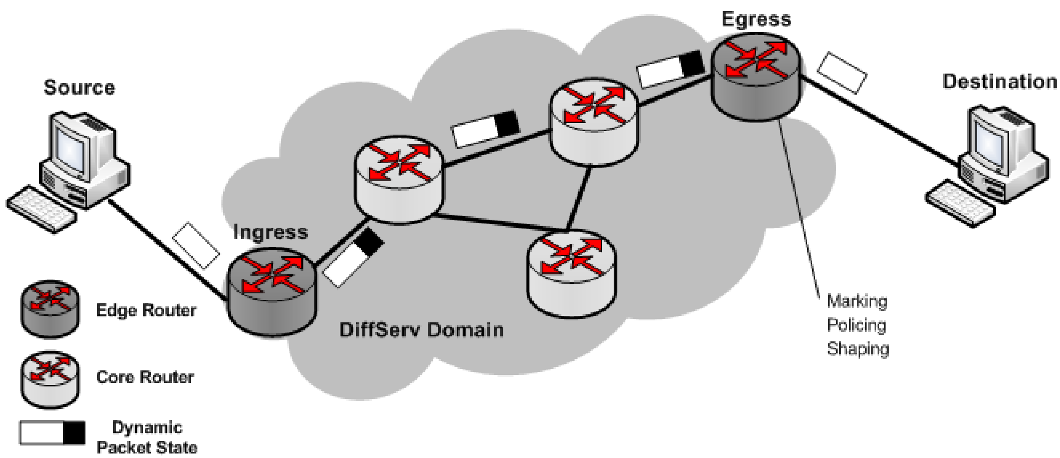


Figure 3.3: SCORE Architecture

Figure 3.3 illustrates the DPS technique used in a SCORE network:

1. The state information is inserted into the packet header once a packet arrives at the ingress node;

2. the core nodes processes the packet based on the packet state information and, if necessary, updates both its internal state and the packet state information before forwarding it;

3. the egress node removes the state information from the packet header and forwards it.

3.1.4 Bandwidth Brokers

A bandwidth broker is an entity responsible for managing resources within a DiffServ domain. The management is made by controlling the network load with the rejection or the acceptance of bandwidth requests from users.

The bandwidth broker entity is defined and described in RFC 2638 [54]. Apart from describing a dynamic QoS framework based on bandwidth brokers, the standard also presents two new types of services - premium and assured services - as well as a set of mechanisms used for the implementation of these services.

The premium service aims to provide a service analogous to the service offered by a virtual leased line without the necessity of building a separate network. For this effect, a small percentage of the total network bandwidth is permanently allocated. In this type of service, customers would be able to expect the desired requirement but it would obviously translate into an elevated cost. Contrary to the best effort traffic which has a bursty nature, requiring a queue management to deal with congestion periods, the premium service is mainly appropriate for regular traffic patterns such as real-time applications thus, it only requires a small or even nonexistent queue. The premium service implements a guaranteed peak bandwidth service with very low queuing delay.

The assured service allows flows using an additional available capacity during congesting periods resulting in a lower penalty than best-effort traffic.

However, premium and assured services are only as good as their admission control schemes, i.e. they are highly dependent on the admission control scheme performance.

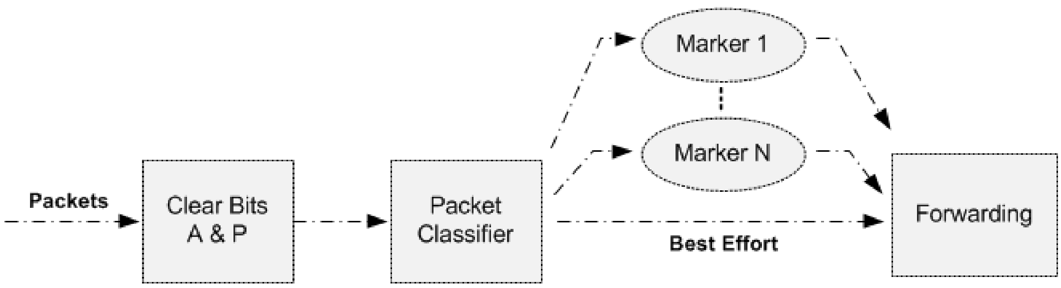


Figure 3.4: Block Diagram of a Boundary Router

Figure 3.4 shows a diagram including the main functions of a boundary router. The type of service being offered to a packet is identified by a classifier that is based on two bit-patterns from the IP header precedent field. These can be classified as a P-bit (premium) or an A-bit (assured) service. After the packet has been classified, they will be subject to the corresponding traffic conditioning mechanisms such as policer, shaper, classifier and priority schedulers.

When speaking about bandwidth brokers, they claim that this new agent should contain organizational policies as well as marked traffic´s track of the current allocation in the same pool. This would permit the interpretation of new bandwidth requests in light of information concerning the local policies and current allocation. For example, whenever a user requests a specific service it must send the request specifying the type of service, target rate, maximum burst and time of request to its BB. The BB then authenticates the requester and verifies whether or not there are enough resources available to meet the request. If the request is accepted, the available bandwidth is reduced by the requested amount of bandwidth and the flow state information is stored. In order to configure the border routers according to the accepted request of QoS requirements, the BB may use for instance, Command Line Interfaces (CLI), Simple Network Management Protocol (SNMP) or Common Open Policy Service (COPS). Inside a DiffServ domain, packets are served according to the way they were marked.

BBs are also responsible for the delimitation of a DiffServ domain as well as the management of the flows sent across domain's boundaries to adjacent BBs.

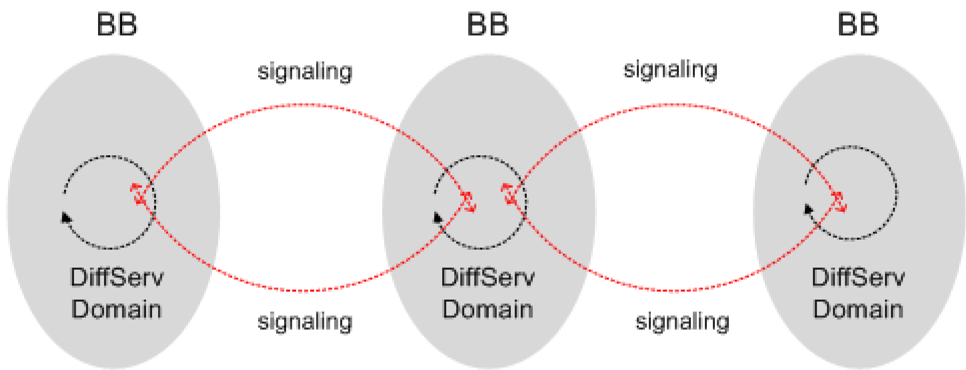


Figure 3.5: Dynamic Allocation with a BB

Figure 3.5 shows how BBs operate at the inter-domain level. At the inter-domain level, the BB establishes a secure connection with its peer BB and negotiates a rate and a service class.

A BB can also easily inter-operate with RSVP in order to convey flows description

(TSpec[3]) to the network. In such case, the RSVP messages sent to ingress router to ask for a service will be intercepted by the local BB which in turn processes the request. If it is approved, it forwards the message to the egress router. Otherwise, it will send a negative RSVP message to the sender with the refusal.

In the particular case of wireless environments, when a MN moves towards a new access router, a BB, as the element responsible for managing the DiffServ router configuration in a DiffServ domain, must be informed about the QoS that will be provided in the new router. Therefore, each MN movement results in QoS sinalization messages between the access router and the BB and consequently, it will not scale.

Thus in this scenario, an intra-domain and inter-domain QoS signaling permitting the exchange of QoS information between the BB and the AR, and also between BBs of adjacent DiffServ domains [89–91] is necessary. After having received the QoS signalization and checking whether sufficient resources are available the BB is able to reconfigure the access routers accordingly.

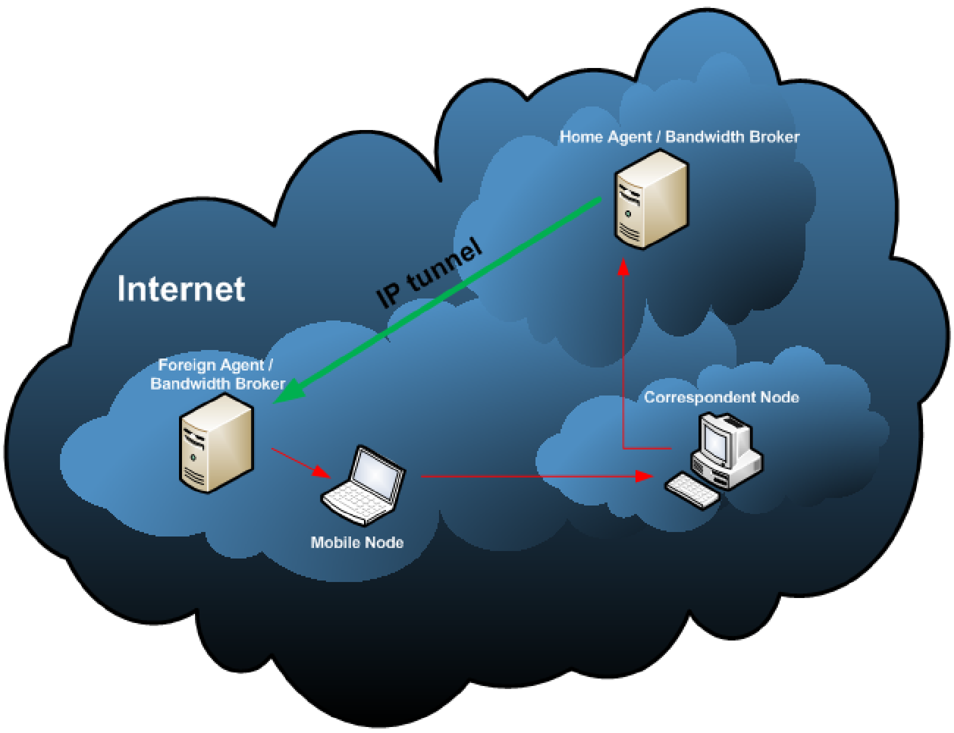


Figure 3.6: Bandwidth Brokers in Mobile Environments

Figure 3.6 illustrates a diagram of a mobile environment with bandwidth brokers. When a MN performs an inter-domain handover by connecting to a foreign domain, its BB trans-

[3]The TSPEC object is defined in RFC 2210 to convey the traffic specification from the sender

mits the MN's SLS information to the foreign BB so that it may configure the access router according to the MN's SLS information.

3.2 Dynamic QoS solutions for Wireless Networks

The dynamic capability is fundamental in order to meet a precommitted service level when a particular source or destination can be anywhere on the Internet. This is the case of wireless networks. QoS solutions enabled with this particular characteristic are also crucial because the static configured levels of services are always constructed as "paid for bits even if you do not use".

3.2.1 Resource Management in DiffServ

The development of Resource Management in DiffServ (RMD) [92] was driven by the new requirements imposed by the wireless networks. The solution aims to provide dynamic resource reservation within a DiffServ domain. In order to accommodate the strict QoS requirements for real-time applications in wireless networks the IETF NSIS working group has decided to [93] build a new QoS solution with the following characteristics: fast dynamic resource reservation, simplicity, scalability, low cost, severe congestion handling and easy implementation. This framework has been developed by the Next Steps in Signaling (NSIS) working group and is described in [57].

The RDM solution provides admission control for flows entering a domain and deals with congestion terminating flows in case of a sudden failure. Admission control is conducted at the edge nodes while the core nodes only apply a resource management based on PHB. The edge nodes generate reservations for each flow QoS request while interior nodes generate aggregated reservations per traffic class. The reservation is quantified in terms of bandwidth and is made in all the nodes in the communication path.

Figure 3.7 shows the RMD signaling protocols. The RMD framework divides signaling protocol into two types: the Per Domain Reservation (PDR) protocol and the Per Hop Reservation (PHR) protocol. The former is used to manage the reservations in the whole DiffServ domain and is implemented only in the edge nodes. This protocol conducts the linkage between external reservations and internal reservations on core routers whereas the latter is used to manage reservations of each interior router, enabling class based reser-

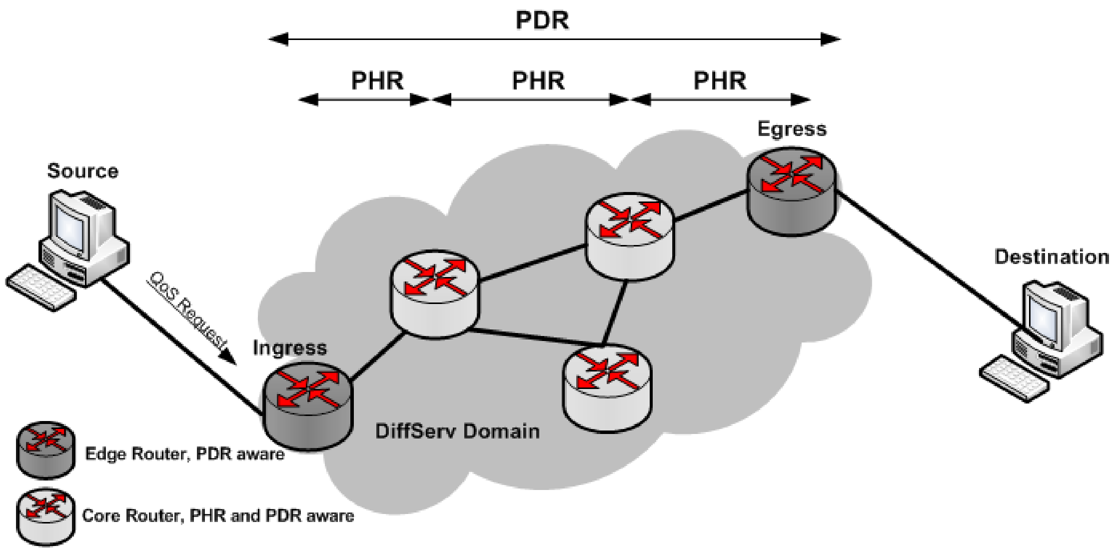


Figure 3.7: RMD Protocols

vations on DiffServ core routers. There are two groups of PHR reservations: one is the reservation-based PHR, where the reservation is done in terms of resource units; the other is the measurement-based admission control PHR, which measures resource availability before admitting any reservation request.

Therefore, each flow in an RMD domain contains two signaling sessions, one end-to-end session (PDR) and an intra-domain session (PHR) which are bounded together at the edges routers. The binding between these two sessions are made with a bound Session IDentifier (SID), this unique session identifier globally identifies a flow. Basically, when an end-to-end reservation is requested, the sender creates a reserve message containing the traffic description that will be forwarded along the path towards the destination node. When the original reserve message arrives at the ingress, a new reserve message is derived from the original. The new reserve message is sent to the intra-domain so that interior nodes can enable class based reservations.

The signaling inter-operation between PDR and PHR may operate in four functional modes: a normal unidirectional resource reservation/query - in this procedure the end-to-end reservation is initiated by the sender and is only propagated in the direction of the receiver; a normal bi-directional reservation/query- in this functional mode two unidirectional reservations/queries sessions are binded; a unidirectional severe congestion handling; and a bi-directional severe congestion handling.

The main goal of NSIS is to provide a general model capable of supporting several signaling applications by means of NSIS Signaling Layer Protocol (NSLP). The NSIS protocol suite comprises two layers: the NSIS Transport Layer Protocol (NTLP) and the NSIS Signaling Layer Protocol (NSLP). The NTLP is implemented with the General Internet Signaling Transport (GIST) [94] and the NSLP relies on GIST for carrying out the many aspects of a signaling message delivery. Each application service has its own NSLP protocol. The messages are passed from the NSLP to GIST via an application programming interface (API). This message is also specified in conjunction with additional information such as signaling application, SID, Message Routing Information (MRI) and direction (downstream or upstream). Upon reception, GIST provides the QoS NSLP with the same information. The NSLP picks the SID value used by GIST to manipulate the session state.

The IETF NSIS working group is working on several activities for realizing this framework as a whole. One of these activities is the specification of a QoS signaling protocol, denoted as QoS-NSLP (QoS- NSIS Signaling Layer Protocol) [58]. The QoS-NSLP protocol is used in combination with a QoS model (QoSM) that is also being specified within the NSIS WG. The QoS-NSLP design can be compared to the decoupling of RSVP from IntServ architecture. The decoupling of signaling and resource management permits a generic signaling protocol as well as independent resource management. This explains why a distinction has been made in the approach between the operation of the signaling protocol and the Resource Management Function (RMF).

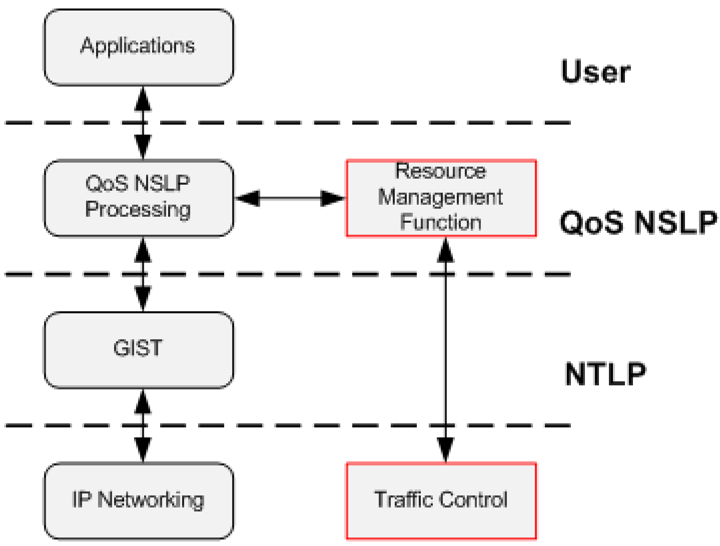


Figure 3.8: QoS-NSLP Interactions

Figure 3.8 illustrates how the QoS-NSLP inter-operates with other components. The QoS-NSLP describes the signaling protocol and QSPEC [95] describes the RMF-related information carried in the QSPEC object which will be transported in QoS-NSLP messages. QSPEC is an object of QoS-NSLP containing the traffic description that will be the input or output of an RMF operation. The QSPEC has the same purpose of the TSpec, RSpec and AdSpec in the IntServ framework which are specified in the [50, 96]. QSPEC is composed of up to four QSPEC objects (Desired, Available, Reserve, QoS minimum). Each one of these QSPECs consist of a set of QoS parameters.

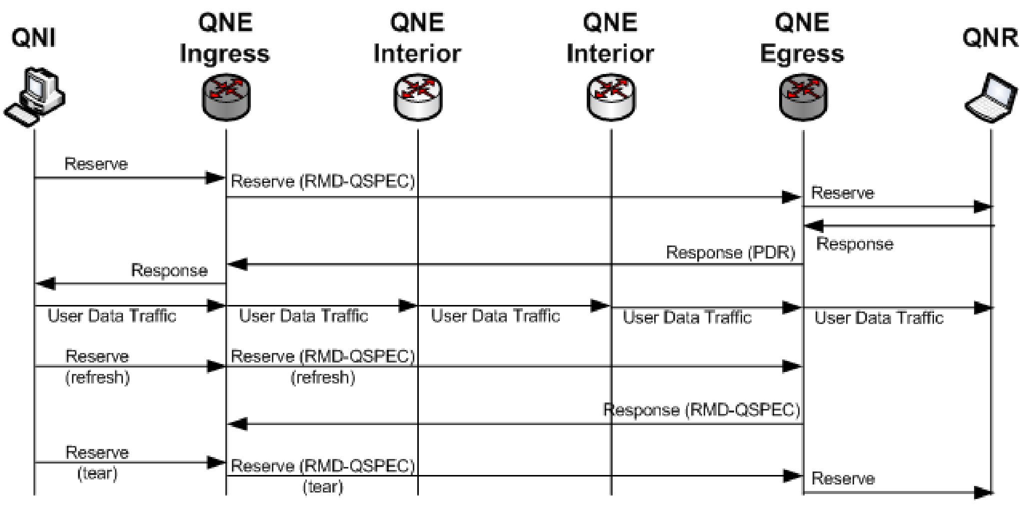


Figure 3.9: RMD QoS Signaling

Figure 3.9 shows the basic signaling messages of RMD-QoS model. The NSIS signaling in an RMD domain possesses two signaling end points: the QoS-NSLP initiator (QNI) and the QoS NSLP receiver (QNR). The QNI sets QoS desired and QoS available QSPEC objects in the initiator QSPEC. Each QoS NSIS element (QNE) on the path reads and interprets those parameters. If one or more QoS desired parameters fails to satisfy, the QNE generates a RESPONSE message to the QNI and the request is aborted. Otherwise, the QNR generates a positive RESPONSE to the QNI. The QNI sets the M flag for each QSPEC parameter it populates that must be interpreted by QNEs. A two-way transaction (RESERVE-RESPONSE) QSPEC procedure is used to populate the QoS parameters. The NSIS entities only handle the NSIS messages containing QoS parameters. The requested QoS parameters are handled by the RMF, which coordinates the activities required to grant and configure resources, such as the admission control which determines whether the node has sufficient resources to supply the requested QoS, and if the checks are succeeded.

The communication starts with a RESERVE message created by QNI (sender) with a QSPEC describing the reservation. When the message arrives at the ingress node, an RMD-QSPEC is constructed based on the sender´s QSPEC. The RMD-QSPEC is sent to interior nodes. This independent RESERVE message goes through the core routers towards QNR. This local RESERVE message uses the NTLP hop-by-hop datagram signaling mechanism. Each QoS NSLP node on the data path processes the local RESERVE message and checks the availability of resources using either the reservation-based or measurement based method. If an intermediate node cannot accommodate the new request, it indicates this by marking a single bit in the message and continues forwarding the message until the egress node is reached. A response message is sent directly from the egress node to the ingress node. When the message reaches the egress node, the original end-to-end message is forwarded to the receiver. When the egress node receives a RESPONSE message from the receiver, it forwards the message directly to the ingress node.

IP addresses change frequently in wireless networks due to mobility. This affects the NSIS reservations because they are coupled to the data path [97]. Consequently, QoS reservation became invalid because the flow no longer exists. Thus, end-to-end signaling cannot be avoided since flow state information is associated with a source and destination address. However, the NSIS protocol allows the decoupling state information and flow identification. A state information is identified by the SID and the flow associated with this state information is defined by an MRI. Decoupling is beneficial because it allows the update of a signaling state in the network due to mobility along the common path and only establishes a new NSIS state information in the new data path. As in mobile environments, only a small portion of the end-to-end signaling path is usually affected. The scope of signaling information to create and release NSIS states is limited. As additional information, the NSIS WG introduces some operations with a crossover node to enhance the release of the old state information.

Figure 3.10 shows the QoS-NSLP signaling procedure in MIPv6 with the routing optimization option on, where the data flows from the MN to the CN and the reservation is sender-initiated. When a handover event occurs, the MN must update the signaling path according to the nCoA. Therefore, the MN sends a QoS-NSLP RESERVE message con-

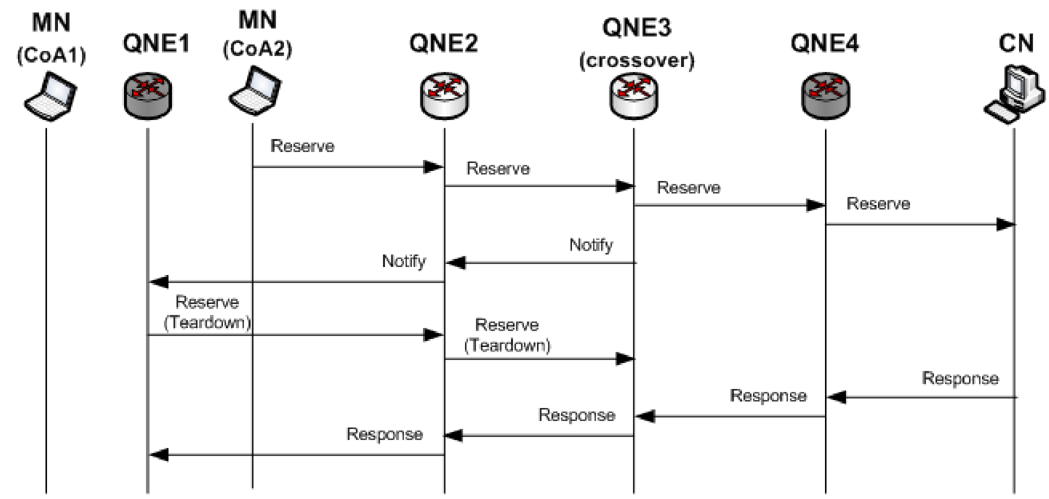


Figure 3.10: NSIS operation in MIPv6

taining the SID towards the CN. Reservation requirements as well as other identification information for the session is also sent. All the QoS-NSLP nodes along the path receive the RESERVE message and establish the corresponding NSLP state. If an NSLP state with the same SID already exists, the state will be updated and they will then forward the RESERVE message towards the CN.

In the case of all the QoS-NSLP nodes have enough resources to accommodate the required QoS, the CN responds with a positive RESPONSE message. The obsoleted state in the old path is released by a timer expiration. If a crossover node is used to speed up the release process, the RESERVE message will be intercepted by the crossover node, which is detected as the first node where the state information has the same SID of the RESERVE message, the crossover node then sends a notification towards the MN's old QNE. The old QNE receives the notification and sends a RESERVE message towards the CN with teardown bit set, in order to release the old state information. This reserve message will be stopped by the crossover node.

For future work, the NSIS WG provides some suggestions to reduce the QoS signaling latency derived from handovers in Mobile IP scenarios such as the inter-working of the GIST/QoS-NSLP protocols with a local mobility management protocol. This solution aims to further improve the re-establishment of NSIS signaling within a localized scope.

3.2.2 Mobile Extensions to RSVP

The RSVP resource reservation protocol is based on static network infrastructures, thus inadequate for scenarios with mobility where bandwidth is limited and the operating conditions are non-deterministic. Despite the fact that RSVP is flexible and robust, it does not adequately address resource reservations in mobile environments. The mobile extensions to the RSVP addresses the widely known concerns with resource reservation application in wireless mobile environments due to the maintenance of QoS guarantees following cell transition events [21].

The original RSVP is not mobility aware and therefore cannot be used directly in a mobile computing environment for the two following reasons: 1) RSVP messages are invisible to intermediate routers of an IP tunnel used in Mobile IP because the IP tunnel is implemented by an IP-in-IP encapsulation scheme [27], consequently, routers are unable to recognize RSVP QoS signaling messages; 2) after an MN moves to a new location previously allocated, resources in the old path must be released and new resource reservations must be made in the new path. The first problem can be solved with an RSVP Tunnel [98] which has been proposed for resolving RSVP signaling invisibility problem.

In order to resolve the second problem, the mobile extensions to RSVP propose several enhancements to RSVP. The extensions include attributes such as quiescent resource reservations, virtual receivers, intelligent pre-allocation of resources, based on the analysis of user mobility patterns, predictive look-ahead dynamic dormant multicast trees, fulcrum nodes, mobility management agents and extended RSVP messages. This combination of new attributes allows the MRSVP protocol to make advance resource reservations at multiple locations where an MN may possibly visit during the service time. The MN can thus achieve the required service quality when it moves to a new location where resources are reserved in advance. However, the execution of these modifications imposes a significant burden on network resources.

However, apart from the fact that it overcomes the handover impact of mobility on RSVP by making advance resource reservations in all neighboring sub-nets, these excessive resource reservations may waste significant bandwidth in turn reducing network performance.

Moreover, RSVP raises difficulties in terms of scalability because it must periodically refresh per-flow reservation states.

3.2.3 Hierarchical Mobile RSVP

Hierarchical Mobile RSVP (HMRSVP) combines Mobile RSVP with Hierarchical Mobile IP [99, 100]. The main differences between MRSVP and HMRSVP reside in the local registration of the MN and the advanced resource reservation which is only made when the MN proceeds an inter-domain handover, contrary to MRSVP which establishes reservations on all the MN's surrounding cells.

In this approach when an MN moves to an overlapped area between two domains, the HMRSVP establishes an extra passive resource reservation along the path from the sender to the adjacent cell of the MN, in order to accommodate a possible movement to the adjacent cell.

Whereas MRSVP establishes reservations on all the MN's surrounding cells inside or outside domain. When compared with MRSVP, this scheme achieves the same QoS guarantees and outperforms in terms of reservation blocking, forced termination and session completion resource reservations probabilities.

Figure 3.11a, shows an MN that is currently visiting a non-boundary cell; in such scenario they assume that the MN will only make intra-domain handovers in the near future. Therefore, the HMRSVP only establishes an active resource reservation along the path from the sender to the MN without making any advance resource reservations. Figure 3.11b, shows an MN entering an overlapped area of a boundary cell; in such scenario the HMRSVP will establish an extra passive resource reservation along the path from the sender to the boundary cell of the MN's neighboring domain. In both Figures, dark line represents active resource reservation paths while dashed line represents a passive resource reservation path.

In this scenario, the HMRSVP establishes a passive reservation because it is considered that the MN may make an inter-domain handover into the new domain.

In conclusion, although the presented solution may reduce the impact of Mobile RSVP's problems, it still inherits the same framework problems of significant processing burdens and resource waste. Moreover, the solution is restricted to HMIPv6 networks, therefore, it does not inter-operate with other mobility protocols such as MIPv6 or FMIPv6.

3.2.4 QoS-Conditionalized Handoff for Mobile IPv6

The key idea of QoS-Conditionalized Handoff for Mobile IPv6 [101] is to employ the QoS hop-by-hop option piggybacked in a binding message of mobility management to provide the QoS signaling support to make a conditional handover based on the resource availability along the new data path towards new access router. The scheme is built over the hierarchical mobile IPv6 in order to be suitable for micro-mobility scenarios.

The scheme operates as following: 1) the QoS is carried in Binding Update message; 2) then each router between the MN and MAP interprets the QoS option of the binding message and checks for its availability. If the resources are insufficient, it sends a negative response in an extended Binding Acknowledgment message and refuses the handover. If the response is positive in all routers, the handovers will proceed. This solution may reduce the signaling bandwidth on the backbone part of network communication by conducting local handovers QoS maintenance.

In conclusion, it adopts an in-band solution for QoS signaling to avoid the use of a signaling protocol. The solution has the disadvantage that all nodes must be modified in order to implement the required functionality.

3.2.5 Mobile Extensions to DiffServ using a Policy-Based Management System

A Policy-Based Management System (PBMS) is a technology that intends to facilitate the management and operation of a network. Such system, when implemented in a centralized server, can provide absolute QoS guarantees for applications.

A policy management system consists of the following major functional components:

- Policy Repository: A directory service where policy rules are stored. It is typically implemented as a LDAP[4] directory server.

- Policy Decision Point (PDP): a centralized network entity that makes policy decisions. To make the decisions, it may gather the information storage into the policy repository.

[4]The Lightweight Directory Access Protocol, or LDAP, is an application protocol for querying and modifying directory services running over TCP/IP.

- Policy Enforcement Points (PEPs): Enforce the PDP policy decisions.
- Policy Administration System: Policies are defined and administrated in a Policy Administration System. It typically provides a high-level user interface for operator input.

A policy-based management system can, for instance, be deployed with a centralized BB for performing admission control and policy delivery [102], where BB policies will be configured according to the existing Service Level Agreements (SLA) with the customers and Service Level Specification (SLS) supplied by the administrator. The COPS[5] protocol will enable the BB to communicate the policies decisions [103].

A Policy-Based Management System can be implemented for making QoS management between adjacent DiffServ domains, as described in work [104] which is one example of this type of architecture, or between a DiffServ domain and a IntServ domain [105, 106].

The former approach provides end-to-end differentiated services in Mobile IPv6 (see Fig. 3.12).

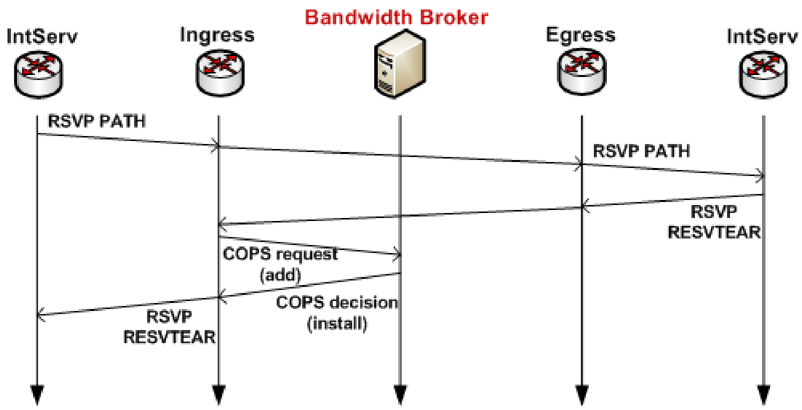
This architecture has been used with the Common Open Policy Service - Service Level Specification (COPS-SLS) protocol [103] for making inter-domain SLS dynamic negotiations across DiffServ domains and a scheme for making the DiffServ context transfer within the domain. The context is used for re-establishing DiffServ context in new data paths thus avoiding the re-initiation of COPS-SLS signaling from scratch.

The latter architecture is another possible approach for supporting dynamic resource management between DiffServ and IntServ domains (see Fig. 3.13).

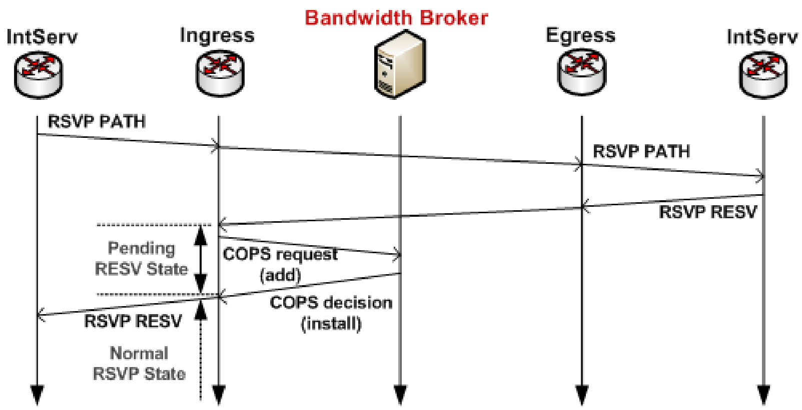
In this approach, the IntServ domain and the DiffServ domain are integrated through a centralized Bandwidth Broker in order to make resource allocations in both domains. In such an architecture, the BB with the help of ingress Edge Router, (ER) will make

[5] COPS is a very simple protocol with one message to make the queries and another to responds to those queries. The COPS protocol uses a single and persistent TCP connection between policy enforcement point (PEP) and policy decision point (PDP). The COPS protocol has two operational models, the outsourcing model and provisioning model. The former is used when there is a trigger event in the PEP that needs a policy decision. In this model PEP sends a query message to the PDP, waiting for a response decision. In the last model the PDP proactively configures the PEP. The configuration information is stored in the Policy Information Base (PIB).

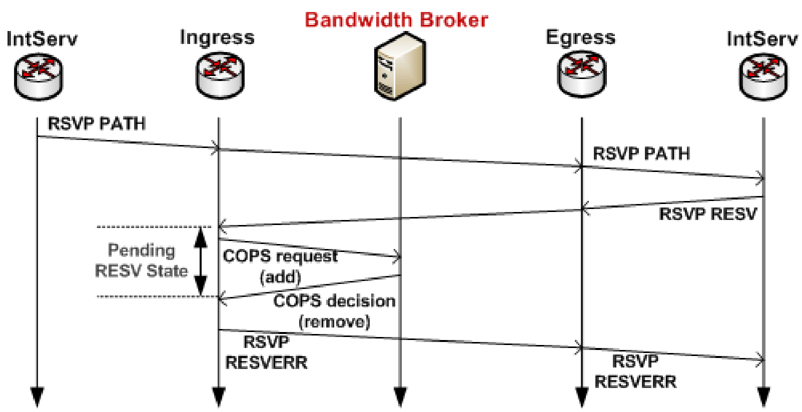
admission control decisions. The ERs communicate with the BB using the COPS protocol which exchanges simple request/response messages for making resource allocations. The ERs act as a PEP which is the client side of COPS protocol, and the BB as the role of PDP which is the server side of COPS protocol. The COPS protocol uses a single and persistent TCP connection between PEP and PDP.



(a) Reserve Messages



(b) Refresh Messages



(c) Remove Messages

Figure 3.14: Resource Reservation Signaling Messages

Figure 3.14 shows the resource reservations signaling procedures in this architecture for dynamic QoS provisioning. For instance, when the sender wants to start an application, it generates an RSVP PATH carrying the flow´s QoS requirements towards the receiver. In the IntServ domain, the message is subjected to the standard IntServ processing functions while in the DiffServ core it is transparently forwarded. After the ER receives the message, it then performs the admission control based on the requested QoS and on its resource availability in order to admit or reject the flow. If the QoS request is accepted, than the message continues towards the receiver, otherwise it is sent back to the sender as an RSVP error message.

The ERs are RSVP aware and store per-flow states. They are capable of managing packets both on a per-flow basis and on an aggregation basis. The ingress ERs have the following functionalities: classification, mapping, marking and admission control. The COPS protocol is used to communicate between the ER and the BB.

The ingress ER acting as a client PEP explicitly asks the PDP/BB for a given amount of resources from the ingress point to the egress point. The queries from ingress ER to BB are triggered when it receives an RSVP message from IntServ domain.

Some problems arise in this centralized solution: complexity, scalability, synchronization between RSVP and COPS and consistency, e.g., case a response from PDP/BB becomes temporarily unavailable the network will enter into a blocking state for new RSVP messages.

3.2.6 Projects Concerning QoS in Mobile Environments

A very popular project for QoS in mobile environments is the Broadband Radio Access for IP-based Networks (BRAIN) [107]. The key issues of the BRAIN project are the interactions between mobility and the QoS, the adaptation of protocols to several air interfaces and the unification of the Internet protocols with the mobile network. The solution proposed by BRAIN is based on: the use of RSVP for applications making their QoS requests; the use of admission control functionality in the edge routers; coupling of mobility and QoS signaling; the use of BRAIN Candidate Micro-mobility Protocol (BCMP) which is a micro-mobility solution for providing a better local mobility management.

The Mobility and Differentiated Services in a Future IP Network (Moby Dick) project defines a common architecture for QoS, IPv6 and AAAC (Authentication, Authorization, Accounting and Charging). The general network architecture includes the following entities: Mobile Terminal (MT), Access Network (AN), Access Gateway (AG), QoS Broker (responsible for the overall QoS management), AAAC server (responsible for authenticate, authorize, account and charging the user) and NMS (Network Management System). The NMS entity includes the QoS Broker, AAAC and other functions such as policy functions, service translation functions and alarm processing. The architecture provides user mobility based on mobile IP, QoS capabilities based on DiffServ and authentication, authorization, accounting and charging based on AAA procedures. The philosophy of the Moby Dick project is - "...making fast handovers between cells, in existing connections, with associated QoS parameters, keeping everything accountable." [108, 109].

The Designing Advanced Interfaces for the Delivery and Administration of Location Independent Optimized Personal Services (DAIDALOS) project [110] is a continuation of the Moby Dick project although it was focused on the signaling protocol and service aspects. In the scope of the DAIDALOS project, a QoS management system based on a PBMS that manages and configures the network resources by means of policies, admission control and resource management procedures has been proposed [111]. In order to provide an end-to-end QoS solution independent of wireless technology, the wireless QoS part of the DAIDALOS architecture has a modular architecture composed by two abstraction layers [112]. One of these abstract layers provides a generic interface for the upper layers and assures QoS functionality regardless of the access technology. The other abstract layer provides technology specific QoS functions for the technologies 802.11, 802.15.1, 802.16 and TD-CDMA.

The End-to-End Quality of Service over Heterogeneous Networks (EuQoS) project [113] aims to define an end-to-end QoS architecture for the Next Generation of Networks (NGNs) that manages the QoS across different administrative domains and heterogeneous networks. In order to provide guaranteed QoS across heterogeneous networks, a set of end-to-end network Classes of Services (CoSs) which support a broad range of applications have been specified. In this sense, the mapping functions between the DiffServ classes and the layer 2 classes available in the considered access technologies have been defined. Those mappings are the basis of the EuQoS framework. In order to have QoS across different Autonomous Systems (AS), the neighboring ASs have a pre-negotiated peering SLSs. These agreements specify the amount of traffic the neighbour AS is able to accept as well as the QoS commitments to accommodate the incoming traffic. In this system, the user runs a module which is an interface to the EuQoS services. Through this interface, the user request a QoS connection with an inter-application signaling protocol such as SIP or H.323, which are used to convey QoS requirements and negotiate the codecs between the applications. The EuQoS´ architecture slices the control plane into a technology independent layer control, where a Resource Management (RM) module is located as well as a technology dependent layer control, where a Resource Allocator (RA) module is found. The user's QoS is sent to the RM module which in turn requests the resource reservation for the new flow to the RA module. The RM is responsible for performing AC decisions and communicating the policies to one or more RAs. The signaling protocol used for the interactions between RMs modules (vertical signaling) is an extended version of the NSIS protocol. The protocol used for the communication between RMsand RA modules is an extended version of the COPS protocol. The EuQoS control plane comprises all the functions and protocols needed to create and monitor the end-to-end paths (called EQ-Paths) associated with the requested resources. Each EQ-Path corresponds to a given CoS and is built from a source access network to a destination access network, resorting to QoS routing and resource provisioning mechanisms. In order to build the EQ-Path, the EuQoS uses an enhanced Border Gateway Protocol (BGP) to select the best EQ-Path. The enhanced BGP protocol contains an additional path attribute that conveys the QoS parameters of a path. By means of this BGP, each AS knows the QoS requirements associated with the EQ-Paths for reaching a given destination.

In relation to resource provisioning, two models have been defined in the EuQoS project, namely loose and hard models. The two models have different technical solutions to build the EQ-Paths, hence they reflect different trade-off between flexibility and scalability.

DAIDALOS and EuQoS projects have similar goals and approaches to provide end-to-end QoS over heterogeneous networks. However, due to be very generic, comprehensive and open solutions, their implementation seems to be complex and costly.

The Advanced Radio Resource Management for Wireless Services (ARROWS) project deployed a QoS architecture for UMTS environment. In this approach, UMTS network has been considered an IP access network supporting IntServ model. The IntServ communicates with the DiffServ core network by means of a key component which has been introduced in the mobile terminal. This is the QoS manager. In order to overcome RSVP limitations on wireless environments, the QoS manager adapts RSVP for mobility and conducts the aggregation of IP flows into classes over Packet Data Protocol (PDP) contexts [114].

The QBone project objective is the contiguity of the QoS between different DiffServ domains in order to provide end-to-end services. Qbone architecture is developed by the Internet2 QoS IEFT Working Group [115]. They assume that each QBone domain is represented by a BB which responds to admission requests of applications for network resources. The BB of the QBone project may receive a resource allocation request from one of two sources: either by a request from an element in the domain, or by a request from a peer bandwidth broker. The inter-Domain communication protocol negotiates the Service level agreements (SLS) between peer domains, where one domain is the service provider and the other domain is the customer. SLAs are assumed to be bilateral, between peer domains, and Bandwidth Brokers are the agents whose functional responsibilities include the implementation of the technical aspects of the agreements. The SLA guarantees that the traffic offered to customer domain will be met and carried out in peer domain. However, the problem of mobility and QoS resource management for inter-domain mobility still needs to be further investigated to achieve more mature solutions.

3.3 Conclusion

This Chapter discusses some of the most relevant dynamic QoS solutions for wired and wireless networks.

Dynamic QoS provisioning architectures may be accomplished using signaling protocols and admission control policies. IntServ and Bandwidth Brokers for DiffServ were the first dynamic QoS architecture proposals that arose for wired networks.

The fact that IntServ was initially aimed to have a per-flow granularity made the framework inherently unscalable. Since IntServ has scalability problems in large scale scenarios [51, 52] same important enhancement proposals have been made in terms of core simplification (IntServ over DiffServ) and traffic aggregation (RSVP Aggregation) to turn IntServ more scalable. However, these enhancements had implementation difficulties that is why they are not widely deployed.

The use of policy-based management systems such as a centralized BB entity, for coordinating the network resources is one more element to add to the QoS architecture, therefore, it still needs a QoS model and a signaling protocol to communicate the policy information. Furthermore, BBs are centralized resource management entities. They are complex in terms of implementation because they congregate several features into a single entity, moreover in high dynamic networks such as wireless networks, rather than being a solution they may turn into the network bottleneck [116].

Furthermore, both dynamic QoS architectures are based on deterministic resource reservations for a guaranteed service model, the guaranteed service model requires the creation and maintenance of flows reservations states in all routers along the path. Thus, when an MN moves to a new location, the release of previously allocated resources in the old path is necessary and new resource reservations are made in the new path, resulting in extra signaling overhead, heavy processing and state load.

These architectures have been adapted with few improvements and adjustments for mobile reality in the research projects concerning QoS support in wireless environment mentioned in this Chapter. For instance, the fundamental concept of both BRAIN project was using the RSVP signaling protocol for communicating the application QoS requirements to the network, for admission control purposes. The Moby Dick project also uses explicit QoS signaling for providing guaranteed services to applications within its architecture.

The SCORE solution requires that each packet carry state information and also requires that all DiffServ routers should be modified. Consequently, significant scalability and transparency problems may arise with these QoS solution for wireless networks.

The QoS-Conditionalized Handoff for Mobile IPv6 makes QoS conditionalized handovers without the need of using a signaling protocol. Although, the solution has the disadvantage that all nodes must be modified in order to implement the required functionality.

On the other hand, several extensions to standard RSVP have been made in an attempt to enhance mobile networks with QoS. The first RSVP extension proposal was the Mobile RSVP, a protocol that makes advanced reservations in multiple locations where an MN may possibly go. This solution has the problem of creating excessive resource reservations causing the waste of bandwidth and reducing the network performance.

The HMRSVP solution improves the MRSVP with local MN's registrations and advanced reservations only for inter-domain handovers but still has a significant processing burden and resource waste and is restricted to HMIPv6 networks.

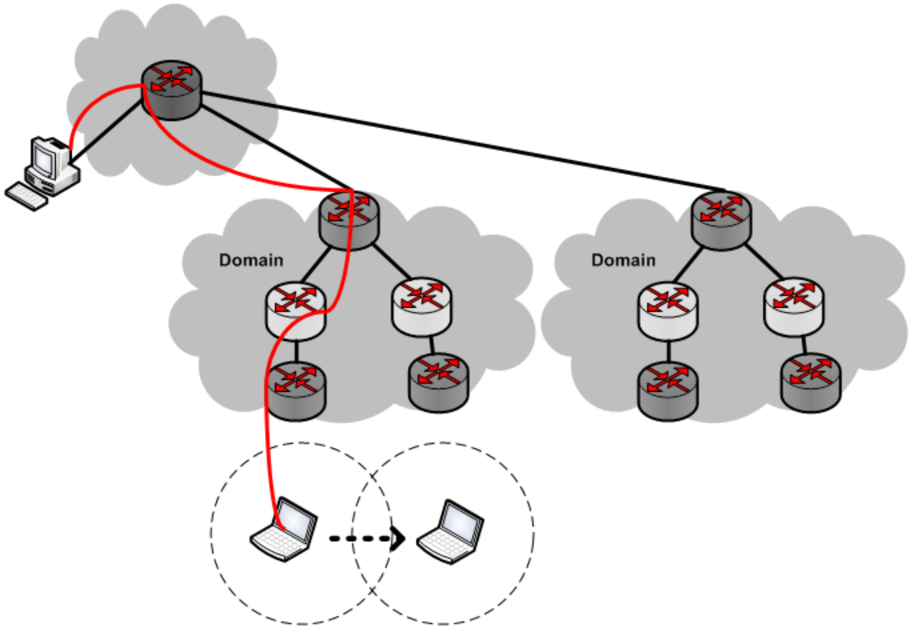
Another MRSVP derived solution is proposed in [117] where the authors introduce a Crossover Router (CR) entity to reduce tunnel distance between previous access router and new access router created by the FMIPv6 protocol. The CR is responsible for intercepting all packets sent to MN's previous CoA and forward them to the new access router. To deliver the QoS requests, they extend Fast Binding Update (FBU) and Handover Inititiate (HI) messages, which are used for informing the new access router of the MN's QoS requirements. With the information of the MN's QoS requirements, the new access router can make an advanced reservation on the common data path. This solution is claimed to outperform MRSVP in terms of signaling cost, reservation re-establishment delay and bandwidth requirements. However, the solution introduces more signaling messages and complexity.

In a more recent proposal, [118] the authors deployed a modified RSVP called Mobility-Aware Resource Reservation Protocol (MARSVP) where the binding update and the binding acknowledgment messages are conveyed in two new RSVP objects, these new RSVP objects must be added to the standard RSVP messages [50]. The solution implies modifications on MIPv6 and RSVP protocols, and on end nodes.

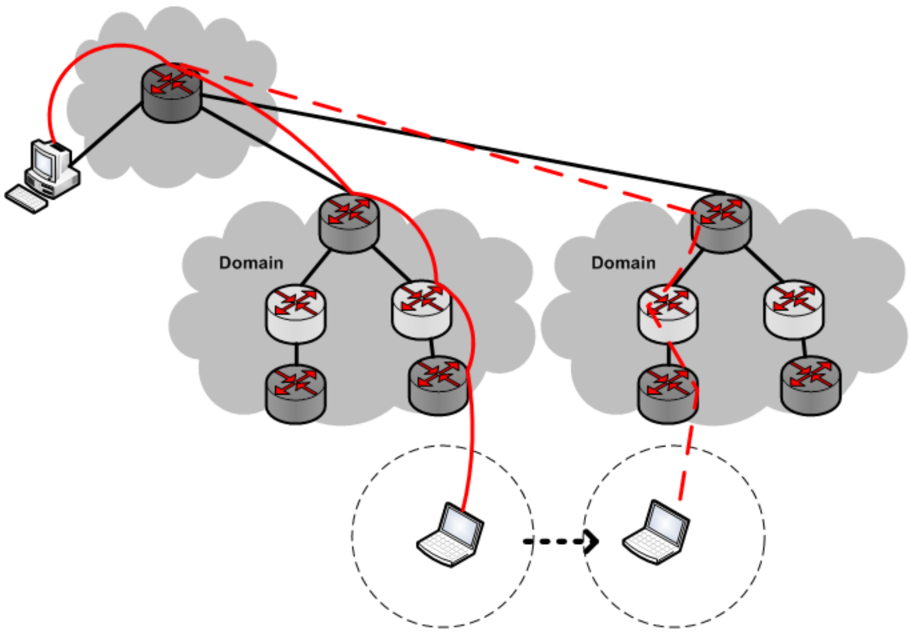
Due to the fact that the proposals mentioned above are based on the guaranteed service model when applied in high dynamic networks, such as wireless networks in the micro-mobility scenarios, significant scalability problems may arise.

In conclusion, despite unquestionable improvements achieved by the above proposals, state information overhead, signaling overhead and processing load caused by frequent handovers are still not completely solved in the existing QoS solutions for mobile environments.

Moreover, the non-deterministic nature of mobile networks makes QoS provisioning with absolute guarantees hardly possible.



(a) Intra-Domain Handover



(b) Inter-Domain Handover

Figure 3.11: Hierarchical Mobile RSVP Scheme

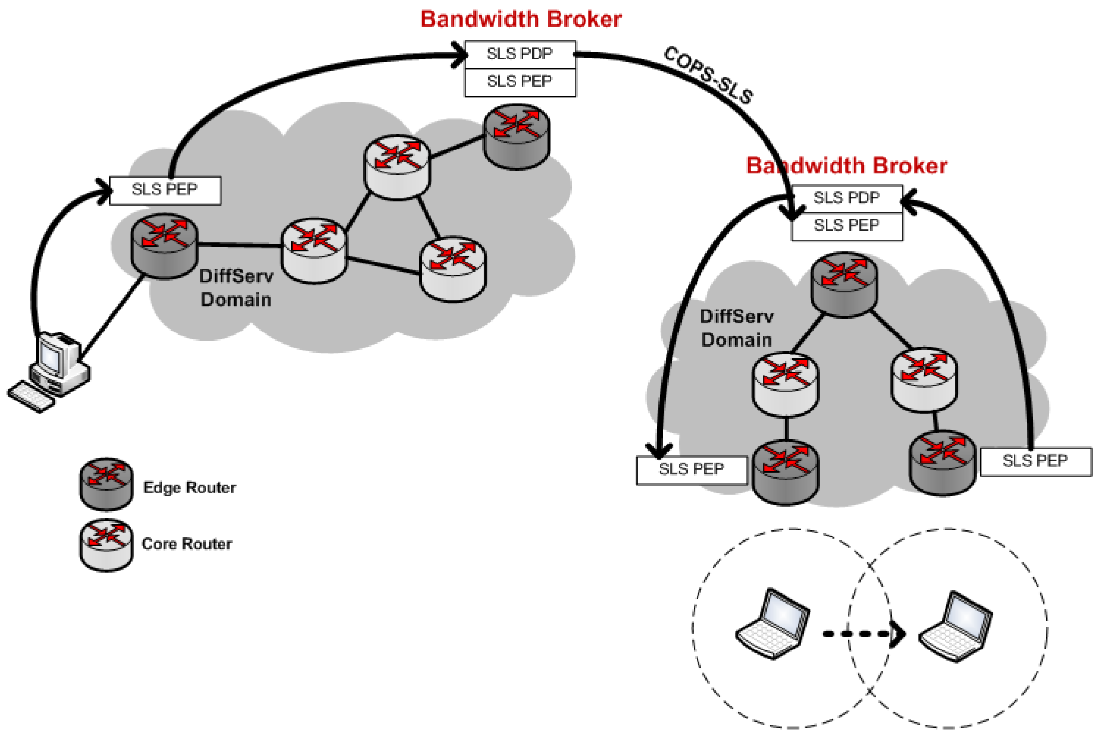


Figure 3.12: COPS-SLS for QoS Negotiation Between DiffServ Domains

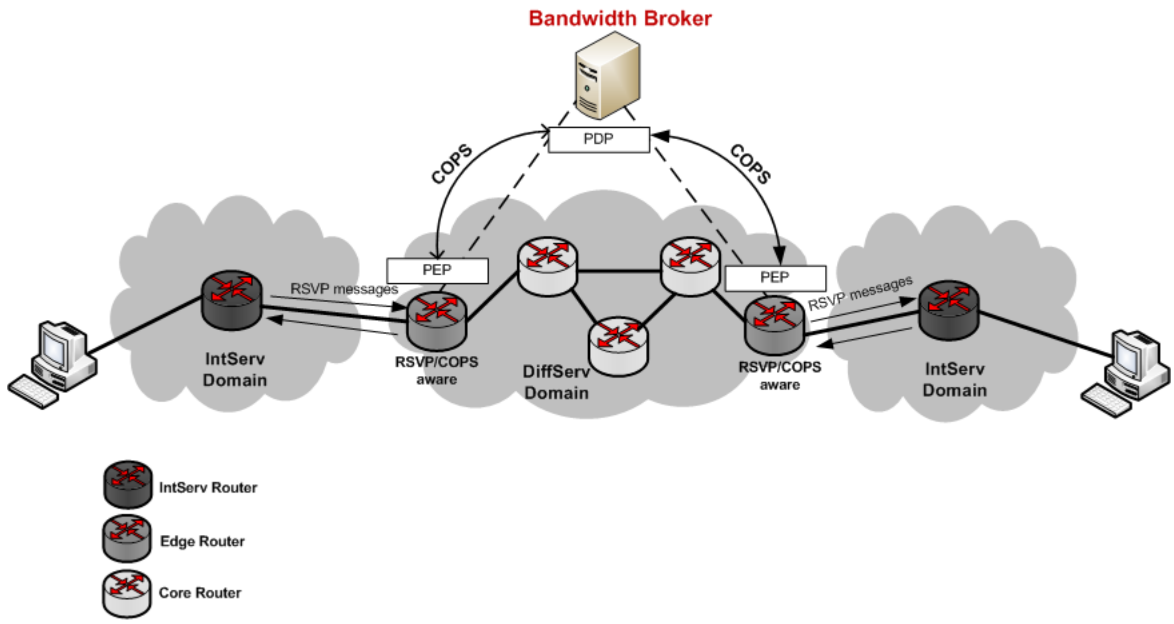


Figure 3.13: IntServ and DiffServ Inter-operation

Chapter 4

Proposed Model

In the near future, wireless networks will undoubtedly run real-time applications with special Quality of Service (QoS) requirements. In this context micro-mobility management schemes such as Fast Handovers over Hierarchical Mobile IPv6 (F-HMIPv6) will be a useful for reducing Mobile IPv6 (MIPv6) handover disruption and thereby improve delay and losses.

However, F-HMIPv6 alone does not support the QoS requirements of real-time applications. Therefore in order to meet this requirement, a novel resource management scheme for the Differentiated Services (DiffServ) QoS model has been proposed. The current proposal is to be used as an add-on to F-HMIPv6.

The new resource management scheme combines the F-HMIPv6 functionalities with the DiffServ QoS model also adding network congestion control and reallocation mechanisms in order to accommodate different QoS traffic requirements. In summary, the new proposed solution comprises enhancements regarding mobility management of Mobile IPv6 (MIPv6) as well as improvements in the resource management of the Differentiated Services (DiffServ) QoS model. The mobility management of MIPv6 has been extended with fast and local handovers in order to improve its efficiency in micro-mobility scenarios with frequent handovers. The resource management of DiffServ has been extended with adaptive and dynamic QoS provisioning which improves resource utilization in mobile networks. Furthermore, the mobility and QoS messages have been coupled together to provide a resource management scheme able to proactively react to mobile events as an attempt to optimize resource utilization.

The proposed solution is aligned with the main requirements needed to support QoS in mobile environments as identified in Chapter 2 while overcoming some of the problems identified in the solutions presented in Chapter 3.

This Chapter presents the proposed QoS solution for micro-mobility scenarios including an overall description of its fundamentals in addition to a description of the subjacent architecture.

The architecture's components such as signaling process, resource management and dynamic allocator are also thoroughly explained in this Chapter as well as the manner in which they inter-operate.

The end of this chapter describes how the proposed solution can be integrated into global mobility as well as how current signaling protocol standards can inter-operate with the proposed solution´s management function.

4.1 Overall Description

The main objective of the new model is to define a micro Mobility/QoS-aware architecture with dynamic QoS functionalities, adaptive resource management and seamless handovers. Another goal is to deal with scalability problems that may arise when handovers are frequent, by reducing the signaling overhead, the processing work and the state load.

In order to overcome the inefficiency of MIPv6 in micro-mobility scenarios, the proposed model enhances the MIPv6 protocol with a specific integration of FMIPv6 and HMIPv6 (F-HMIPv6). The F-HMIPv6 enhances the MIPv6 mobility with seamless handovers and local handover registrations. The integration follows the recommendations defined in RFC 4140 [119] except what concerns Handover Initiate (HI) and Handover Acknowledgment (HAck) messages that are maintained between the previous access router and the new access router, as seen in the FMIPv6 protocol (see Figure 4.1).

The HMIPv6, FMIPv6 and F-HMIPv6 micro mobility approaches were evaluated and compared to Mobile IPv6 in [120]. The authors claim that FMIPv6 is capable of reducing MIPV6 handover latency by a factor of 15. The HMIPv6 is also capable of reducing by a factor of 8 the handover latency of MIPv6. It is also important to note that FMIPv6 and HMIPv6 combined can reduce the overall handover latency by a factor of 18 when compared to the standard MIPv6. Similar studies regarding MIPv6, HMIPv6, FMIPv6 and F-HMIPv6 performance, as seen in [121–123], also presented very similar results.

When speaking of MAP placement, the adopted strategy was to place the MAP in a common crossover router for all ARs in the domain. In hierarchical networks the crossover router is usually found above the ARs. Therefore, the fact that the ingress node in a DiffServ stub domain is a common crossover router for all ARs, it is the best place to redirect traffic to any new data path. Furthermore for fast mobile nodes that perform frequent handovers, it is important that a more distant MAP reduce the probability of changing to a new MAP thus informing all the CNs and the HA of the new location. Therefore, the MAP agent has been placed in the DiffServ stub domain´s ingress node.

However, other solutions for the placement of MAP and more than one MAP agent per DiffServ domain are also possible [3].

Regardless of F-HMIPv6 connectivity improvement, different treatment for incoming and existing traffic with special QoS requirements is necessary. It is also crucial that support be given to QoS mobility by re-establishing the QoS context that MN had on the previous router, on the new router whenever a handover occurs in order to avoid the QoS context re-establishment from scratch. Hence, the resource management function in the nAR would benefit from receiving QoS context in advance by means of F-HMIPv6 handover layer-3 anticipation, i.e., before the MN moves there. By having the QoS context in advance, the resource management function of the nAR can perform proactive actions accordingly with the received MN's QoS context requirements and AR's status. The QoS context received in advance enables to decide beforehand the admission of new handover flows only if the QoS requirements of the existing and the incoming flows are fulfilled. Since the establishment of QoS context on nARs is made before the handover takes place, the re-establishment of MN's QoS context on nAR from scratch is avoided.

In terms of QoS architecture, the proposed model extends the resource management function of DiffServ in the edge routers with a measurement-based admission control mechanism. By taking into account the heaviness of performing admission control in all network nodes regarding the changes and overhead introduced, admission control should be left for critical points. As stated in [124, 125] the edge links are considered as being the most critical points in the domain whereas intermediate routers are over-provisioned. It was assumed that interior nodes are engineered by considering the routing behavior as well as the maximum aggregated traffic injected inside the domain through the ingress router.

As in wireless networks, the most critical points are the ARs on account of wireless link constraints. The admission control in such routers is made for both new and handover flows whereas the ingress router only makes admission control for new flows entering in domain.

In what pertains to state information overhead, signaling overhead and processing load problems caused by the guaranteed service model, our approach has been to overcome these problems with relaxed QoS requirements i.e., the predictive service model of the DiffServ QoS model. Furthermore, as the admission control scheme chosen is based on class traffic measurements, (it is a MBAC) the processing load and the state overhead caused by this mechanism are not critical [67].

The main advantage of using measurements for admission control is the fact that this scheme does not have to maintain any reservation states by means of a signaling protocol. Once an admission decision is made, than no record of the decision needs to be stored, thereby it neither requires a pre-reservation state nor an explicit release of reservation.

Another reason to use an MBAC is the fact that it is in consonance with the DiffServ service model as well as with wireless network´s nature because they only aspire to provide soft guarantees for real-time applications. The absence of rigid QoS guarantees in wireless environments does not result in a problem, since most real-time applications in wireless networks have adaptive playback times, thus tolerating occasional packet losses and varying delays. Therefore, MBAC is suited to support soft real-time application in wireless networks and can provide a higher link utilization [10, 65, 66].

A Measurement-Based Admission Control (MBAC) estimates the traffic levels i.e., predicted resource utilization and admits flows where the resource needs are within its availability at the time of request. The admission decision will be negative if the current traffic, as characterized by an estimator, added with the new flow resource request exceeds the

DiffServ class capacity. The MBAC algorithm uses prior source characterization only for incoming flows, whereas for existing flows it only uses measurements to characterize them.

The new Resource Management Function (RMF) handles the QoS input parameters that are presented in QoS signaling messages. The RMF comprises the DiffServ QoS mechanisms (policer, congestion avoidance and scheduling) and a measurement-based admission control mechanism (estimator and admission control algorithm). The RMF within the Access Routers (ARs) has an additional element - dynamic allocator - that improves network utilization with an adaptive resource management.

The state information used to perform admission control in edge routers is conceptually organized in DiffServ classes. The ARs have additional state information concerning MN's QoS context within the AR.

In what respects QoS signaling, the proposed model uses a simple signaling protocol in order to allow new flows to make their QoS requests and uses the HI/HAck messages which are mobility management messages of F-HMIPv6. These convey the MN's QoS context in order to enable handover flows to request from the new router the desired QoS.

The use of mobility messages to convey MN'S QoS context allows the coupling of mobility management and QoS management thus, levering the possibility of optimizing both managements.

Similar to NSIS framework the QoS signaling protocol used by new flows to request their services, is decoupled from the RMF [58]. Therefore, a distinction is made between the operation of signaling protocol and the RMF signifying that the RMF operability is autonomous from the adopted signaling protocol.

The major design issues in the implementation of the new resource management were: use the DiffServ mechanism as the QoS model; select the AR as the most critical point in the end-to-end path; and define the resource management function in edge routers as lower state information entity.

The RMF of ARs handles the QoS signaling messages of new flows as well as the HI/HAck messages of F-HMIPv6. These messages contain the requested QoS parameters that will be handled by a resource management function which then coordinates the activities required to grant and reallocate resources in the AR.

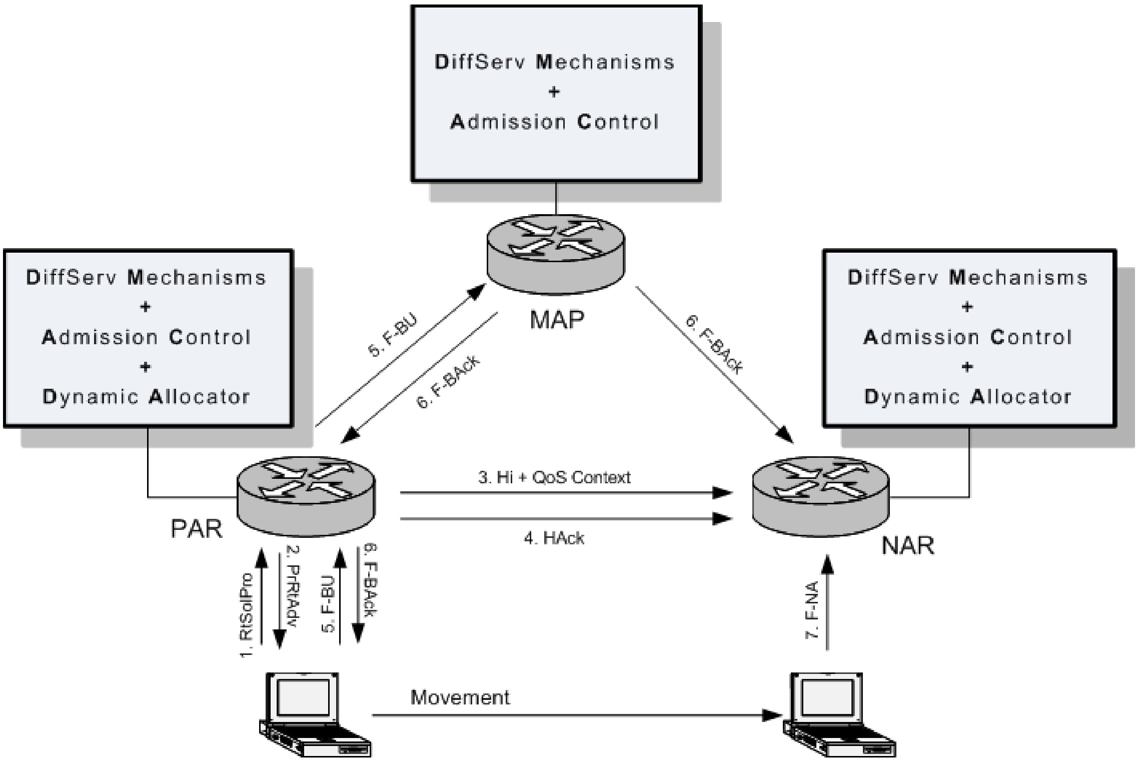


Figure 4.1: Resource Management Function Components and Handover Signaling Process

Figure 4.1 shows the implemented RMF's components and QoS signaling procedure for handovers (see the signaling details in 4.3.1).

Basically, the RMF consists of three components:

1. QoS model - Diffserv QoS mechanisms treat priority traffic differently;
2. Admission Control - Admission control determines whether or not a node has sufficient resources to support the requested QoS and;
3. Dynamic Allocator - Reallocation mechanism that reallocates more bandwidth for handover flows belonging to priority classes.

Figure 4.2 illustrates the four main RMF functions (Measure, Estimate, Police and Reallocate bandwidth). Estimators implement measurement mechanisms in order to determine the current network load in terms of DiffServ class bandwidth and DiffServ class bandwidth per MN (which is MN's QoS Context).

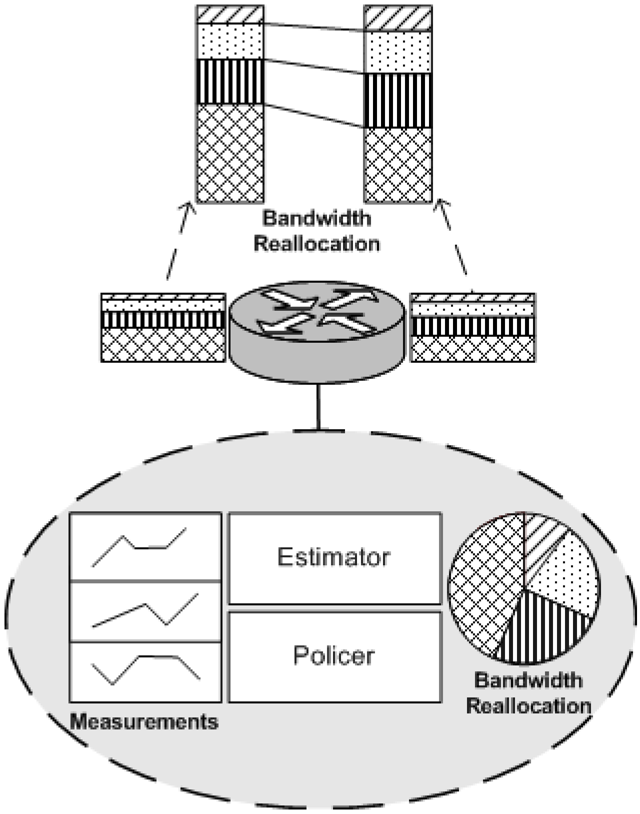


Figure 4.2: Resource Management Main Functions

The policer runs an algorithm to decide whether to admit or reject flows. For new flows, the decision is based on inputs from the traffic descriptor and on measurements of DiffServ class bandwidth against a given class threshold (which is the allocated bandwidth for that class). When speaking of handover flows, the decision is based on inputs from MN's QoS context in pAR and measurements of MN's DiffServ classes bandwidth in nAR at the time of the handover against a given class threshold.

Additionally, if necessary, the dynamic allocator, which acts as bandwidth reallocation mechanism, dynamically redistributes the allocated bandwidth for best-effort traffic among the DiffServ classes with stricter QoS requirements in order to accommodate additional incoming handover flows in higher priority DiffServ classes.

This proactive (before MN moves to a new location) and dynamic (by adjusting the load within classes for handover flows) RMF behavior may provide seamless mobility by maintaining the same MN's QoS level across ARs.

As mentioned previously in Chapter 2, another specific problem of wireless environments is that of the DiffServ packets transparency on interior nodes caused by IP tunneling. This problem has been solved in our proposed solution thanks to the propagation of DSCP information in the packet header to the outer IP header, as recommended in [11].

In summary, the model design issues are:

- Based on the DiffServ model
- Integrates HMIPv6 and FMIPv6 Handover mechanisms and uses MAP as the aggregation point
- New access router receives QoS context in advance:
- Coupling of mobility and QoS management
- Implements a proactive resource management
- Uses a simple signaling protocol for new flows
- Uses in-band signaling for handover flows
- Implements class state information in ingress
- Implements class state and class state per MN information in access routers
- Proposes a class measurement-based admission control

The model components and how they are interconnected are explained in the next sections.

4.2 Resource Management Function Behavior

Figure 4.3 shows the proposed resource management function. The resource management function located in the edge routers of standard DiffServ is comprised by a classifier which marks data packets with a DiffServ Code Point (DSCP) as well as a traffic conditioner which is compound by meter, marker, shaper, and dropper components.

Meter measures network traffic that is used to supervise whether or not the traffic stream is in compliance with a given traffic profile. According to the measured meter, an action such as remarking, dropping or shaping can be performed in order to enforce traffic to be in conformance with a given traffic profile. The interior nodes are implemented with Per-Hop Behaviors by means of buffer management and packet scheduling mechanisms.

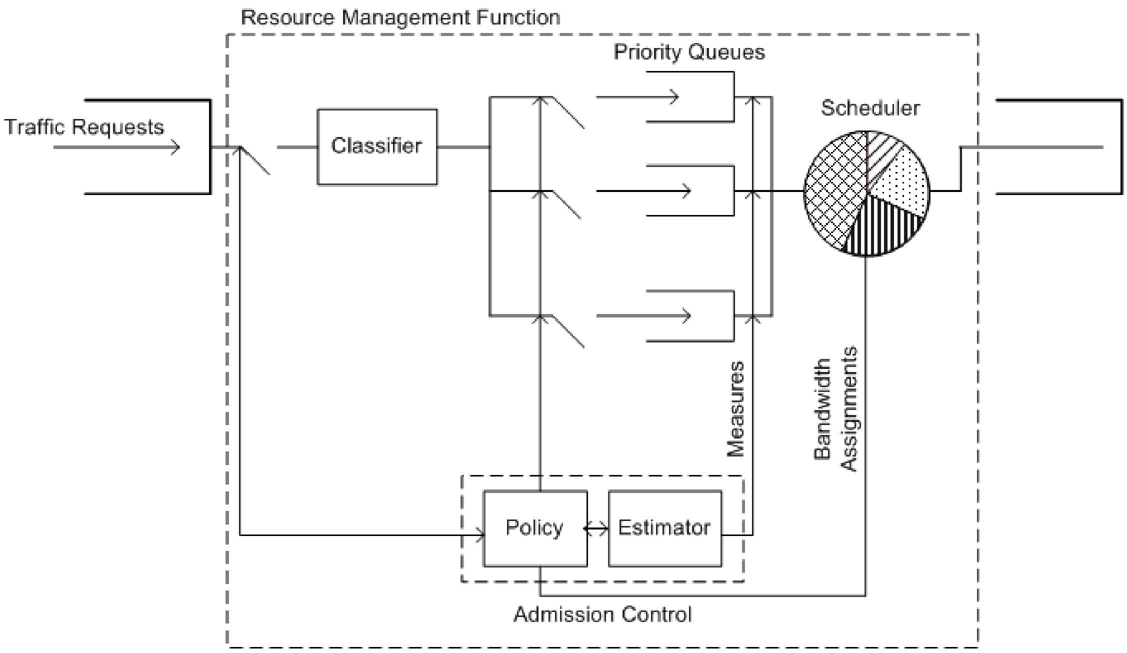


Figure 4.3: Resource Management Function Components

In the DiffServ model, the resources are allocated statically to a specific DiffServ class or allocated dynamically by means of a BB which configures DiffServ QoS mechanisms in the edge routers to a specific DiffServ class accordingly to the QoS requirements specified by an SLS. However, the BB is a centralized entity designed for fixed networks which only conduct admission control for new flows entering the domain. Therefore, when an MN moves to a new location, the BB must be informed to perform the admission control for handover flows and also to make the correspondent configuration on the new edge router. Furthermore, a resource management solely based in a centralized BB demands that each MN movement must be signaled, stated and processed in this central entity. Thus the BB itself can become the bottleneck in the edge routers´ resource allocation.

On the other hand, standard DiffServ mechanisms such as PRI scheduling, are not limited to a threshold of the amount of allocated resources that a priority DiffServ class can obtain. As a consequence, the lower priority classes can enter in starvation if the traffic of higher priority classes saturate the link capacity. Furthermore, a DiffServ queue management such as Random Early Detection (RED) is also insufficient in avoiding link congestion.

Admission control algorithms are responsible for limiting the number of flows admitted into the network. For inelastic real-time flows, such as Voice over Internet Protocol (VoIP) or video conferencing services, admission control mechanisms are essential for ensuring the availability the channel has in carrying the traffic load.

For these reasons, the resource management of standard DiffServ has been extended with explicit setup mechanisms to request resources from the network for the purpose of supporting class admission control in ingress and access routers. For admission control purposes, a new measurement-based admission control has been used.

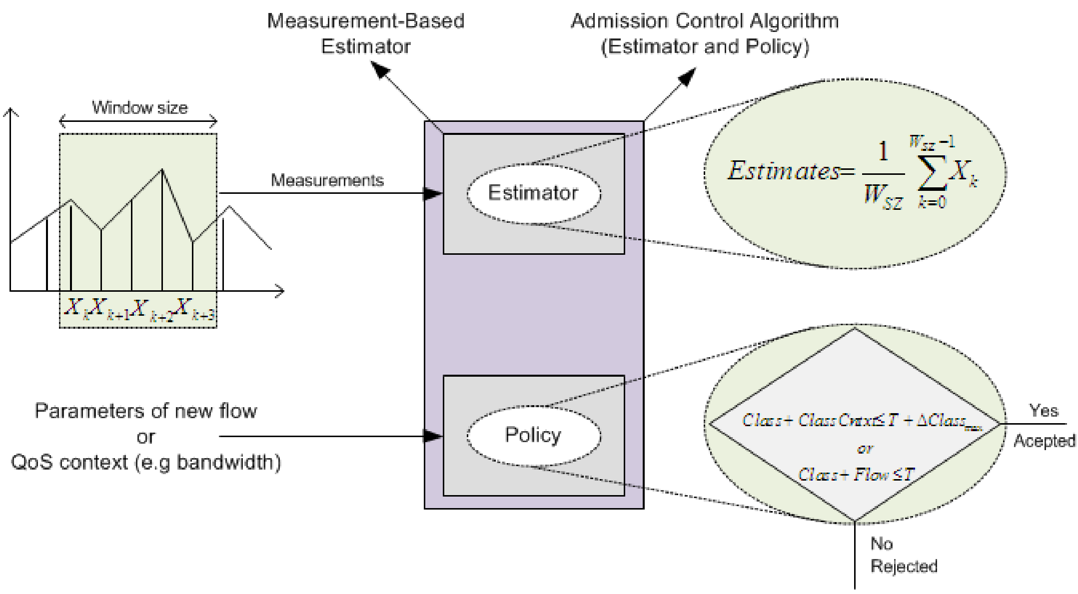


Figure 4.4: Measurement-based Admission Control

Figure 4.4 illustrates the measurement-base admission control. The new class measurement-based admission control consists of a rate estimator and an admission control algorithm / policy. The rate estimator measures the actual class bandwidth load (associated with the wired part of access router) as well as the MN's QoS context i.e., its DiffServ context. The MN's QoS context is the measured bandwidth being used within each DiffServ class on the actual AR by MN. In other words, the MN's QoS context is the measurement of the aggregated traffic in each individual DiffServ class used by MN.

In order to decide whether to admit or reject a flow, a measure rate sum algorithm has been proposed and used [62]. For new flows, the decision is made on the ingress and access router and is based on inputs from the traffic descriptor as well as on measure class bandwidth in use. The decision for handover flows is based on inputs from the MN's QoS context as well as on measure class bandwidth being used in the new access router at the time of handover.

For handover flows, the decision is made only on the new access router in case of a local handover. When speaking of inter-domain handovers, the decision is made based on the ingress router and on the new access router.

The state information required for making the admission control decision for handover flows can be therefore grouped into two categories: QoS context (Class bandwidth load for each MN class on pAR, at the handover instant) and router status (estimated Class bandwidth on nAR, at the handover instant). The admission control decision will be the result of the policy condition illustrated in Fig. 4.4.

The QoS context is extracted from a rate estimator which measures the actual bandwidth load per class on behalf of the MN. The rate estimators used in our model have already been explained in the Sub-Section2.3.3 of Chapter 2. When the MN intends to move towards a new router, the QoS context on the current AR is sent to the nAR using a mobility management message. After receiving the QoS context, nAR will compare the new QoS requirements with its resource availability (see Equation 4.4) using the admission control algorithm to compare the requested MN's bandwidth plus the estimated in the nAR, against the allocated resources for that class on the corresponding nAR. If the computed resources are insufficient, it refuses all flows belonging to the class and admits them with or without bandwidth re-allocation among the classes. The end result of this is that only MN flows within a class that does not violate the QoS resources in the nAR will be transfered and thus it can always ensure desirable QoS levels on the AR.

Therefore, MN handover flows are conditioned by class resource availability in nAR and QoS context. The following equations present the policy implementation on AR nodes

$$Bw = \sum_{i=1}^{D} Class_i \qquad where\ B_w \leq C \tag{4.1}$$

where D is the number of DiffServ classes, Bw is the occupied bandwidth, $Class_i$ is the sum of the aggregated traffic on a given class and C is the link capacity.

$$Class = \sum_{j=1}^{S} Session_j \tag{4.2}$$

Equation 4.2 determines the used class bandwidth, where S is the number of sessions on a given class. Equation 4.3 determines the occupied bandwidth by session,

$$Session = \sum_{k=1}^{F} Flow_k \tag{4.3}$$

where F is the number of flows on a given session. A session is defined as an association between a corresponding node and the MN related to a data flow. The same corresponding node and MN may have more than one session active at any one time. The session is associated with a certain flow. Often there will only be one data flow for a given session but there may be more than one (see Figure 4.5).

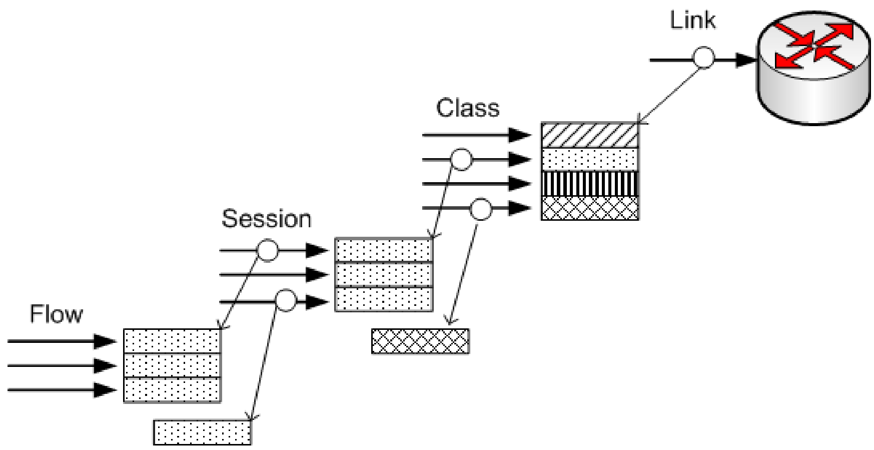


Figure 4.5: Data Hierarchy

For instance, the FTP [126] is a common example of a multiple flow based service. An FTP session is first established between a client and a server, and then one or more distinct flows can be established in order to carry data from the server to the client.

The QoS context that will be transfered to the nAR is the measured class bandwidth utilization, named $ClassCntxt_i$. Once in nAR, the AC will decide, based on the Equation 4.4, which MN class it will accept or reject, whenever a handover occurs,

$$Class_i + ClassCntxt_i < T_i + \Delta max_i \tag{4.4}$$

where $Class_i$ is the sum of the aggregated traffic on a given class, $ClassCntxt_i$ is the bandwidth required for that class i, i.e., MN's QoS context for the class in question, T is the maximum traffic admitted in the class and Δmax_i is the maximum bandwidth that can be additionaly reallocated for class i.

For new flows wishing to transmit its traffic to the network, the edge routers on the path have to make AC decisions based on the following policy:

$$Class_i + Flow_i < T_i \tag{4.5}$$

where $Class_i$ is the sum of the aggregated traffic on a given class i, $Flow_i$ is the bandwidth required for the flow and T_i is the maximum traffic allowed (the initial threshold) in class i. Therefore when speaking of new flows, the admission control has a flow granularity, whereas for handover flows, the admission control possesses a class granularity.

4.2.1 Dynamic Allocator

The admission control algorithm implemented in access routers has been extended with a reallocation mechanism based on the hysteresis method, called dynamic allocator. The dynamic allocator's main objective is to achieve an improved resource utilization in addition to a simultaneous increase of the number of accepted MN's classes meeting the required QoS. The dynamic allocator can induce the increase of the accepted handover flows by reducing bandwidth allocated for BE in favor of priority DiffServ classes.

The presented solution enhances the deterministic admission control where the system does not have any sort of adaptability to the network's condition with an adaptive admission control capable of giving an appropriate response to handover QoS requirements and simultaneously increase resource utilization without compromising the global system quality and stability.

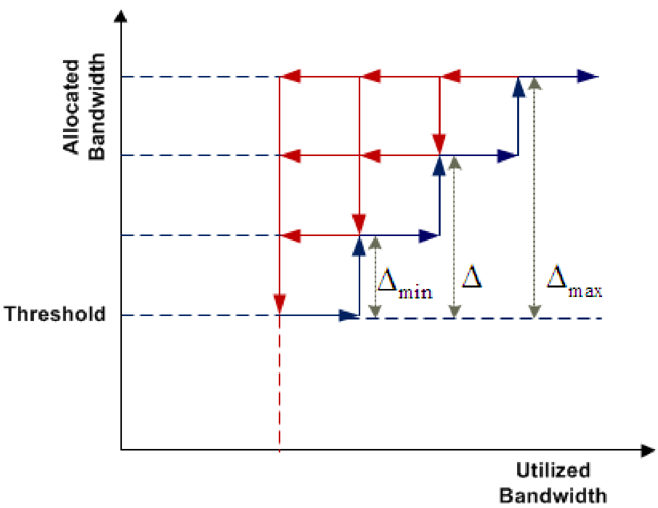


Figure 4.6: Allocated Class Bandwidth with Hysteresis

Figure 4.6 illustrates the reallocation mechanism of the dynamic allocator. Equations 4.6 and 4.7 present the policy defined by the dynamic allocator to share the uncommitted part of the best-effort class.

$$0 \leq \Delta Class_i \leq \Delta Class_{max_i} \tag{4.6}$$

$$\Delta BE_{min} \leq \sum_{i=1}^{D} \Delta Class_i \leq \Delta BE_{max} \tag{4.7}$$

The implemented scheme leads to a predictable and stable behavior on the reallocation mechanism by making bandwidth reallocations in fixed step sizes (see Equation 4.8).

$$\#steps_i = int\left(\frac{(Class_i + ClassCntxt_i) - T_i}{\triangle min_i}\right) + 1 \qquad (4.8)$$

The admission control algorithm always accepts the MN's handover flows whenever it has bandwidth available to reallocate in the required class ($\triangle max_i$). For instance, assuming that an MN starts with handover procedure to move to a new AR and at this time for MN's class i to be admitted in the new AR the number of steps necessary to reallocate is $\#steps_i = 3$, then the reallocated bandwidth will be

$$\triangle Class_i = 3 \times \triangle min_i,$$

if and only if $3 \times \triangle min_i \leq \triangle max_i$.

4.3 QoS Signaling

A two-way signaling protocol is used for new applications to express their service requests to the network. Service requests contain a traffic descriptor describing the worst case application traffic behavior and the required DiffServ class.

Signaling protocol communicates with edge router Signaling Agents (SAs) the traffic and service specification of an incoming flow (see Figure 4.7).

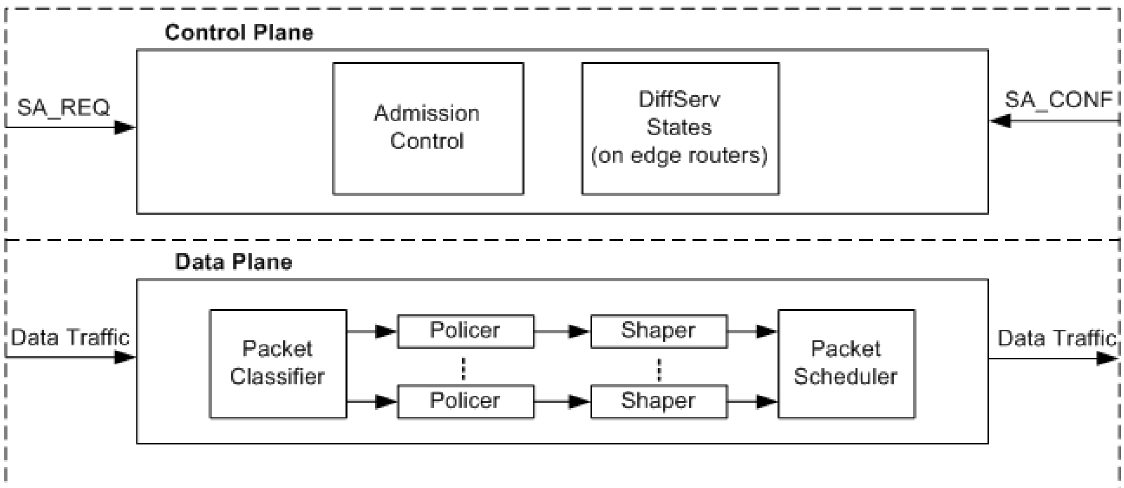


Figure 4.7: Communication Process with Edge Routers

To communicate new flows, the Correspondent Node (CN) uses its SA to request services from the network; this SA is responsible for the delivery of all service request messages. Signaling Request (SA-REQ) messages sent by CN contain the traffic description that will be the RMF input. The message contains two parameters: Desired Bandwidth and Class. The Signaling Agent sets the desired bandwidth and class such that each SA on path could read and pass those parameters to the resource management function. If one of the edge routers in the path fails to satisfy the desired QoS, the receiving Signaling Agent generates a negative Signaling Confirmation (SA_CONF) message to the SA initiator (the CN) with a negative decision, and the flow is aborted. Otherwise, the receiving Signaling Agent generates an SA_CONF with a positive decision allowing the flow to start with its traffic transmission.

Relating to handover flows, as stated before HI/HAck mobility signaling messages have been used to convey the MN's QoS context. The two following sections present the details of QoS signaling procedures for intra-domain and inter-domain handovers.

4.3.1 Intra-domain Handovers

For intra-domain handovers, the MN's QoS Context in pAR is conveyed by HI messages to nAR. The HI messages will be handled by the resource management function of nAR. The HI handover signaling message triggers the resource management functionalities in the nAR before the handover occurs resulting in a proactive behavior which allows the resource management function to adapt its configuration for incoming handover flows.

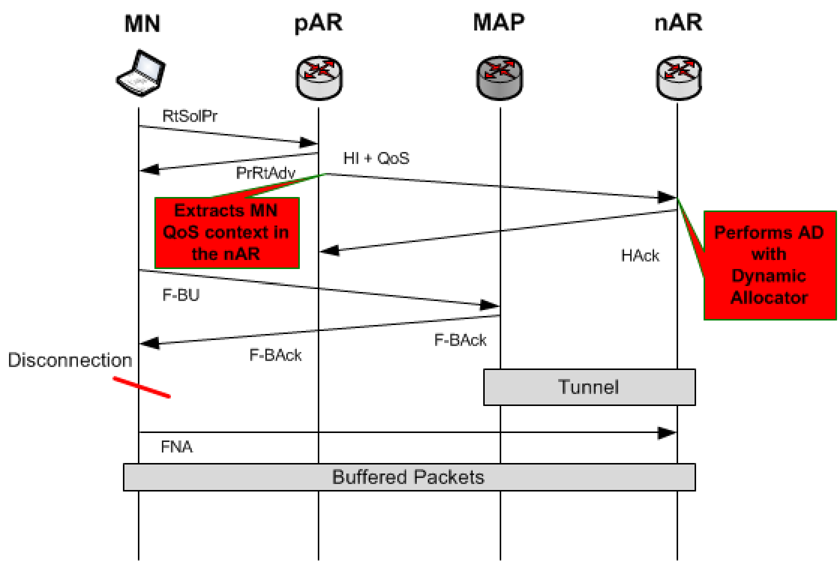


Figure 4.8: Intra-domain Handover Signaling Procedure

Figure 4.8 shows the signaling procedure for intra-domain handovers. Therefore, whenever an MN wishes to change its point of attachment, it must ask for a new CoA address to nAR by sending Router-Solicitation-for-Proxy (RtSolPro) message to pAR. The pARs receives the RtSolPro message and generates a Proxy-Router-Advertisement (PrRtAdv) message with a prospective new MN CoA and sends it to the MN. The pAR also creates an HI message containing the nAR address as well as the MN's QoS context to be sent to nAR. The MN's QoS context in the pAR is extracted from the rate estimator of the resource management function which measures each DiffServ class bandwidth in use on pAR by MN at that time.

This per-Class state information (MN's QoS context) is stored in the mobility options of the HI message´s field. The new access router receives the HI message and in turn processes the mobility and corresponding resource management functions. The resource management function decides which MN's DiffServ classes of flows it is able to accept and, if necessary, the dynamic allocator of resource management function fetches more bandwidth for classes with stricter QoS requirements to accommodate the flows belonging to those priority classes.

Next, it forms a valid Care-of-Address (CoA) or validates the prospective new CoA and places the CoA and the admission control decision on a HAck message, and returns the message to the pAR. The pAR receives the HAck, validates the new CoA address and sends a negative decision on an SA_CONF message (the message is not illustrated in the Figure 4.8) , containing the rejected flows to CN. Then, the MN sends a Fast Binding Update (F-BU), via pAR to MAP for binding its previous CoA to the new CoA. MAP receives an F-BU message and sends a F-BAck message to MN and nAR. The MN needs to wait for the F-BAck message before it makes the handover because this message indicates that MAP is prepared to make the tunneling of the packets to nAR. When the MN receives the F-BAck message, it first disconnects from the pAR and then re-attaches to nAR. Once in the nAR, MN sends an FNA message to receive the buffered packets in the nAR.

4.3.1.1 Inter-MAP Handovers

In micro-mobility scenarios where an administrative domain due to an elevated amount of mobiles requires more than one MAP entity, the proposed model architecture can be easily implemented in order to provide fast and smooth inter-MAP handovers for MNs. For this purpose, it was assumed that previous MAPs are authorized. After having received a Fast Binding Update message, to forward packets to local CoA associated with the ARs in neighbor of MAP domain.

Another consideration, a MAP node must be configured to forward packets to all ARs adjacent of its ARs on the boundary of the MAP domain. The forwarding of packets to nAR allows the MN to continue receiving packets while it is simultaneously updating the bindings in the new MAP as well as within its home agent.

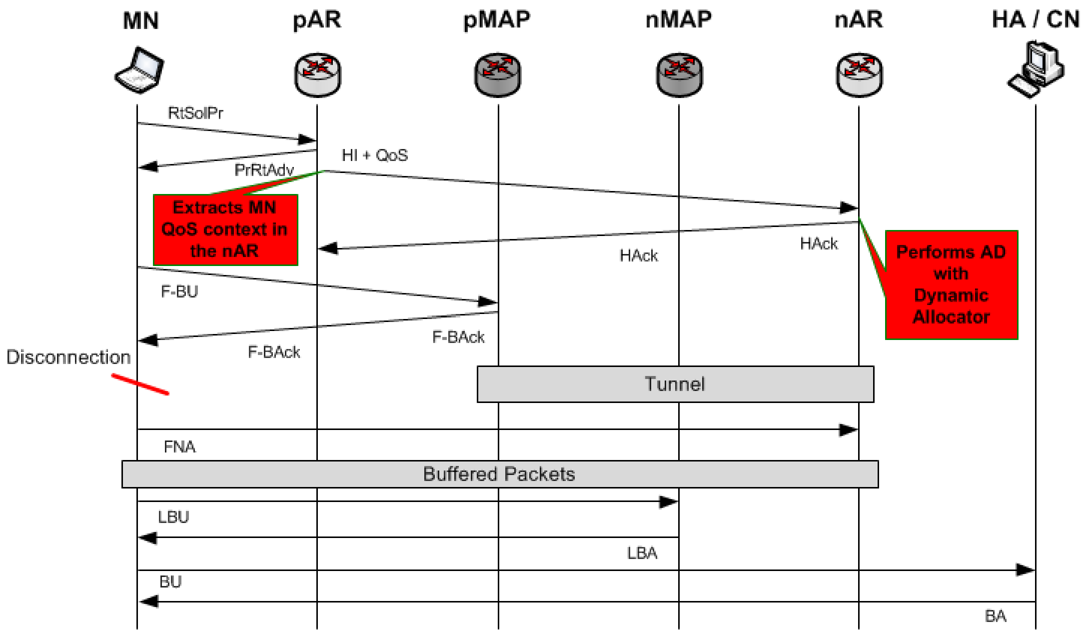


Figure 4.9: Inter-MAP Handover Signaling Procedure

Figure 4.9 shows the inter-MAP handover signaling procedure for the proposed model. So every time the MN detects movement, it must detect whether or not it is in the same MAP domain or in another MAP domain. For instance, when an MN enters a new MAP domain, it must configure the regional CoA (RCoA) address on the new MAP and the local CoA (LCoA) address. The LCoA is configured with the network prefix of the the nAR and the RCoA is configured with the network prefix of the new MAP.

4.4 Extending the Proposal For Global Mobility

A designed goal of the model is designing a micro Mobility/QoS-aware architecture capable of being extended for global mobility. Figure 4.10 illustrates the network reference model proposed for global mobility.

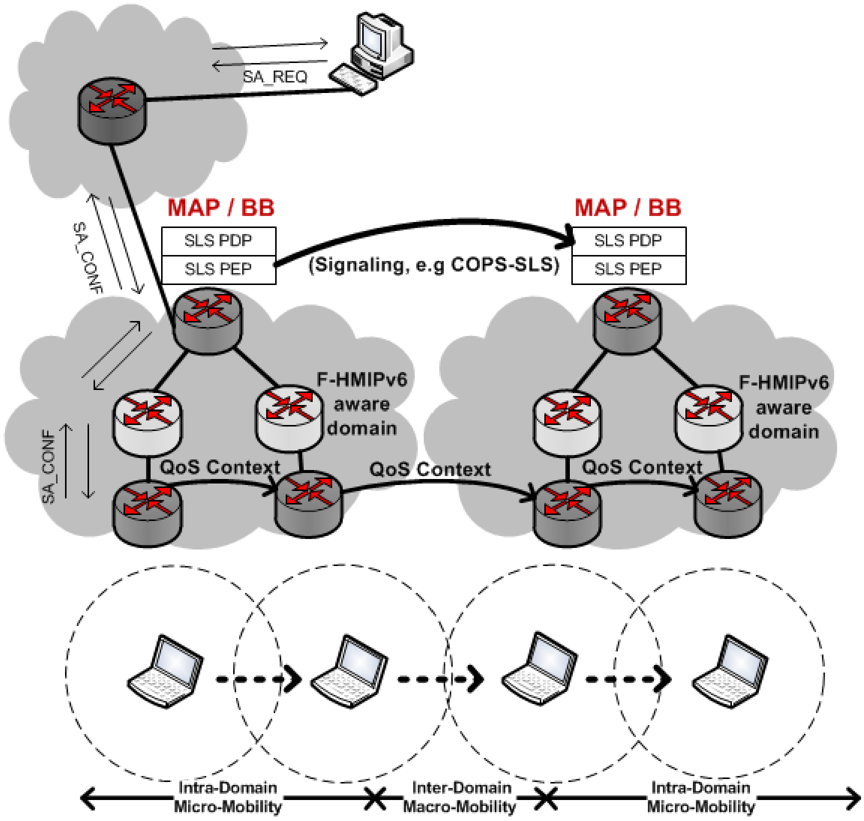


Figure 4.10: Network Reference Model Proposed For Global Mobility

In this scenario, MAP should integrate the ingress router functions: Bandwidth Broker (BB) and inter-domain signaling entity. For inter-domain communication, a signaling entity such as COPS-SLS's entity may be used. The job of the BB is to negotiate SLSs with the BBs of neighboring domains in order to provide users with end-to-end QoS. The BB translates MN's QoS Context into SLS and then negotiate SLS with its peer BB.

Therefore, when an MN moves towards a new access router in another domain its BB, as responsible for managing the Diffserv router configuration in a Diffserv domain, must be informed as to the QoS to be provided in the new router. The BB of the proposed model is only responsible at the inter-domain level which includes the negotiation of QoS parameters and the setting up of bilateral agreements with neighboring domains.

The neighboring domains should have a pre-negotiated mapping of their SLSs to avoid the reconfiguration of DiffServ routers to a new SLS. On an intra-domain level, the edge routers are responsible for the enforcement of resource allocation and admission control instead of the BB.

4.4.1 Inter-domain Handovers

In this scenario the handover flows should be subject to admission control policies within the bandwidth broker of the new domain and the new access router. For inter-domain handovers, it has been assumed that a scenario where domains are F-HMIPv6 aware and previous MAPs are configured and authorized to forward packets to local CoA associated with ARs in neighbor of MAP domain.

The forwarding of packets to nAR allows the MN to continue receiving packets while it is simultaneously updating the bindings in the new MAP (nMAP) as well as in its home agent. Therefore, when an MN enters a new MAP domain, it must configure the regional CoA (RCoA) address on the new MAP and the local CoA (LCoA) address. The LCoA is configured with the network prefix of nAR and RCoA is configured with the network prefix of the new MAP.

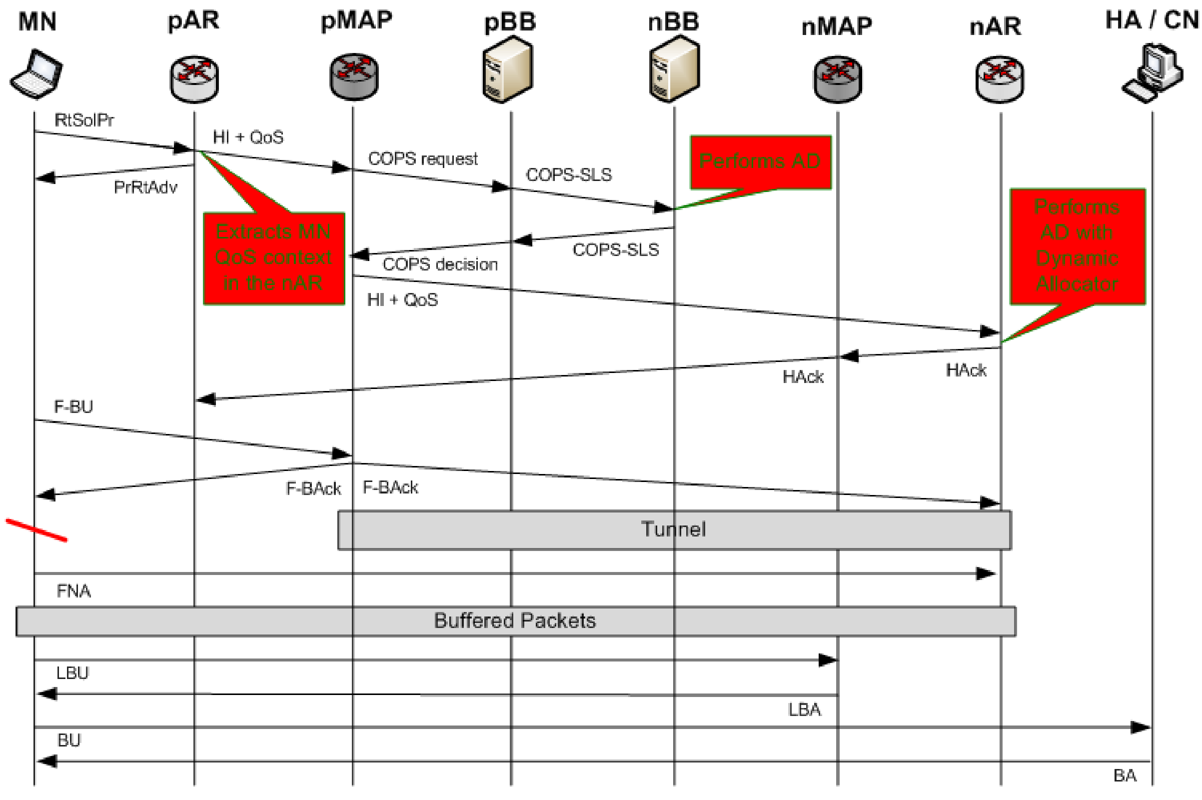


Figure 4.11: Inter-Domain Handover Signaling Procedure

Figure 4.11 illustrates an inter-domain handover signaling procedure. Thus, when an MN enters the new domain it will receive link-layer information from the available access points. The MN may discover an available access point using link-layer WLAN scan mechanisms and then request sub-net information corresponding to the access point. After that, the MN sends a RtSolPr message to pAR to resolve the identifier associated with the found access point. The pAR performs the prefix information match of the access point (provided in RtSolPr), with the prefix list of neighboring access routers, to formulate a prospective new CoA. The resolution of the identifier is a tuple containing the nAR prefix, IP address and L2 address.

The pAR responds to the MN's solicitation with a PrRtAdv message containing the prospective CoA (nCoA). The MN obtains the prospective nCoA while it is still connected to pAR, thus eliminating the need to discover the new prefix after the attachment in new subnet link.

After the MN receives the PrRtAdv message it sends a F-BU message to the previous MAP (pMAP). The MN should wait for the F-BAck message to be sent by the pMAP in response to F-BU before disconnecting from its current sub-net link. As stated previously, the F-BAck message indicates that pMAP is prepared to tunnel the packets to nAR. The pAR also generates an HI message containing the MN's QoS context and sends it to nAR. When the HI message arrives at pMAP through a common routing process, its bandwidth broker translates the MN's QoS context into SLS information and establishes a secure connection with its peer bandwidth broker in order to negotiate a rate and service class. If a request is accepted by the peer BB/MAP, the MAP of the current MN's domain is authorized to forward the MN's QoS context in the HI message to nAR.

The nAR verifies whether or not the nCoA present in HI is already in use (if it is in used it forms a valid new CoA) and checks for its capabilities using the resource management function proposed by the model and then makes an AC decision. Additionally, the nAR, if necessary, can dynamically adapt its configuration by using the reallocation mechanism in order to accommodate the incoming handover flows which belong to priority classes. Then, the nAR respond to the HI message using a HAck message containing the admission control decision.

In the new domain, after the L2 handover, the MN sends an FNA message which receives the buffered packets in the nAR and performs the registration procedures with nMAP and HA. Regarding the correspondent nodes, the MN may send a Binding Update containing its LCoA instead of its RCoA, if they are in the same MAP link. After the correspondent nodes have received a BU, they than can deliver the packets directly to the MN.

According to the proposed extension for global mobility, new flows and inter-domain handover flows will be subject to admission control decisions in the BB and access routers whereas intra-domain handover flows will be only subject to admission control decisions in the access routers.

In the case of the new domain not being F-HMIPv6 aware, no QoS handover will be performed thus resulting in a typical MN handover using MIPv6 protocol for its mobility management and without any QoS support.

4.5 Model Interoperability

A demanding consideration when speaking of the design of a new model is its alignment or compatibility with the current standards. When speaking ofthe interoperability with most relevant QoS signaling protocols such as RSVP and NSIS the proposed model can adopt either NSIS or RSVP for signaling the new flows because the RMF operation has nothing to do with the signaling protocol. Despite the fact that these signaling protocols have been designed for providing services based on the guaranteed service model and the proposed solution intends to offer an alternative service based on the predictive service model, they could be used to convey the new flow QoS requirements. The RSVP or NSIS signaling entity in the edge routers will handle QoS signaling messages and instead of applying the conventional RSVP and NSIS resource management processing, they will pass the QoS information to resource management function of the proposed model in order to be processed.

The use of RSVP and NSIS signaling protocols somewhat changes or extends the usual RSVP and NSIS reservation model where reservations are normally set resorting to deterministic resource reservations. Another useful change in these signaling protocols is the inclusion of the use of proxies agents [97, 127] in the access routers in order to preserve the wireless link. These proxies agents will initiate or terminate the signaling messages generated by the senders on behalf of an MN. The path-triggered receiver proxy approach could be used in access routers.,In this approach, the proxy is activated when a signaling message (e.g. RSVP PATH message) arrives. The RFC [84] document discusses and gives the details of using different resource management approaches in DiffServ domains. In the proposed model, the resource management in the DiffServ domain is dynamically provisioned by means RMF.,In this approach only the edge routers of DiffServ domain will be

RSVP or NSIS signaling aware.

Therefore, the RSVP or NSIS request messages shall be interpreted and processed by the the edge routers RMF. A brief description of how RSVP and NSIS messages could be used in the model is presented:

i) - with RSVP

Figure 4.12 illustrates the model signaling process for new flows with the RSVP protocol. The following sequence describes the process by which an application requests a QoS when RSVP is used by a CN.

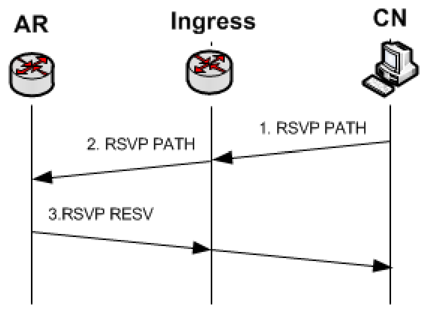


Figure 4.12: Signaling Messages for New Flows Using the RSVP

1. The CN sends a RSVP PATH message describing the traffic which will be generated by an application running in the CN.
2. The PATH message is sent toward the receiving MN. At the F-HMIPv6 domain ingress node, the RSVP PATH is processed and the PATH state is installed within the router. The PATH message is ignored by interior routers and is then processed within the access router according to standard RSVP specification.
3. When the AR receives the PATH message, its RSVP receiver proxy sinks the PATH message and behaves as if an RESV message has been received and performs the admission control without establishing any RESV state. If the RESV message is admitted, it forwards the RESV message upstream to the sender. If it rejects the QoS request, the RESV is not forwarded. If the sender receives the RESV message, it interprets it as an indication that the QoS request has been admitted.

ii) - with NSIS

Figure 4.13 illustrates the signaling process with the NSIS protocol when a CN requests resources for new flows. The following sequence describes the process by which an application requests a QoS when NSIS is used by a CN.

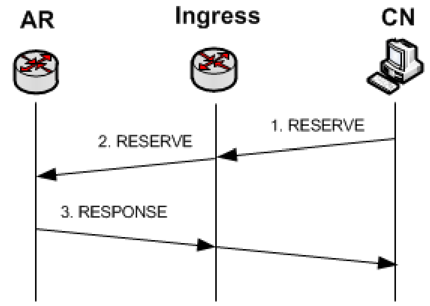


Figure 4.13: Signaling Messages for New Flows Using the NSIS

1. The CN sends a RESERVE message containing QSPEC description with PHB class and bandwidth in the QoS desired object.
2. When the RESERVE message is received at the DiffServ domain ingress node, it passes the message as input to the RMF so that it will be subject to the admission control policy. After the admission control decision, the message is forwarded towards MN.
3. When the message reaches the AR containing a proxy NSIS agent, it passes the RESERVE message to the RMF. If the admission control decision is positive in the both routers, it sends a positive RESPONSE message back to the CN. Otherwise, it sends a negative RESPONSE message.

4.6 Conclusion

This work proposes a model which aims to provide dynamic QoS provisioning to local mobility. The model encompasses two IP mobility enhancements. The first enhancement is a specific integration of FMIPv6 and HMIPv6 (F-HMIPv6) which aims improve handover latency and reduce registration time. The second enhancement is the coupling of mobility management with a specific resource management, where mobility management is based on F-HMIPv6 and resource management is based on a novel DiffServ resource management scheme. Since the standard DiffServ model has its resources statically provisioned, the new DiffServ model resource management has been enhanced with an adaptive and dynamic QoS provisioning.

In order to accomplish this goal, a combination of Fast and Hierarchical Handovers, in-band signaling, DiffServ resource management, QoS context transfer and a Measurement-Based Admission Control (MBAC) algorithm have been integrated to design a QoS framework solution for mobile environments. This symbiotic combination of components has been optimized to work together in order to support seamless handovers with suited QoS requirements for mobile users running multimedia applications.

In summary, the model proposes an extension of the MIPv6 mobility protocol with F-HMIPv6 as well as that of the DiffServ QoS model with QoS signaling and admission control. Furthermore, the model can also be easily extended to global mobility by assuming that each domain handles a bandwidth broker responsible for the inter-domain level where MN's QoS context negotiations between administrative domains are made with an inter-domain signaling protocol such as COPS-SLS.

Finally, the proposed model also provides a QoS solution for mobile environments which is in compliance with the current mobility standards thus it possesses inherent inter-operation with Mobile IPv6. Furthermore, the proposed model can also be in compliance with RSVP and NSIS signaling protocols with the extensions mentioned in the previous section.

Chapter 5

Model Implementation

The purpose of designing a simulation model is to imitate the real system in order to study its behavior. The model is similar to a mini replica or representaion of the real system. The creation of a simulation model makes it possible to perform experiments that are impossible, impracticable, or prohibitively expensive in the real world. Real system simulations are common in the computer communications field. There are several different simulation techniques including, for instance, stochastic modeling and discrete simulation that perform simulation models. Despite the differences among the different simulation techniques, all have a common objective which is to model a real system as realistically as possible.

In this chapter the model architecture implementation in the Network Simulator version two (NS-2) is discussed along with the manner in which the data collected from NS-2 simulations was processed in order to obtain relevant statistical information and graphics, and to perform a rigorous and extensive evaluation of the model performance.

Therefore, a brief description of how mobility protocols is supplied. DiffServ mechanisms and heterogeneous access technologies (NIST[1]) are implemented in NS-2 as well as the manner in which these components are integrated in the proposed simulation model. Then, the admission control algorithm used in the dynamic allocator element is presented.

The end of the chapter presents the selected tools for treating the data collected from NS-2 simulations as well as how the data process was organized for the statistical analysis

[1]The NIST extension has been developed in the Seamless and Secure Mobility Project as NIST, it is located on http://www.and.nist.gov/seamlessandsecure/toolsuite.html.
The NIST patch integrates different access technologies like ethernet, WiFi, Bluetooth and UMTS into a single platform, making possible simulate heterogeneous mobile environments.

and graphics production.

5.1 NS-2 Basics

NS-2 is an event-driven simulator written in C++ where the greater part of C++ hierarchical classes as a mirror class (one-to-one) in OTcl[2], which serves as an interpreted front-end. The root of the hierarchy is the class TclObject (see Figure 5.1).

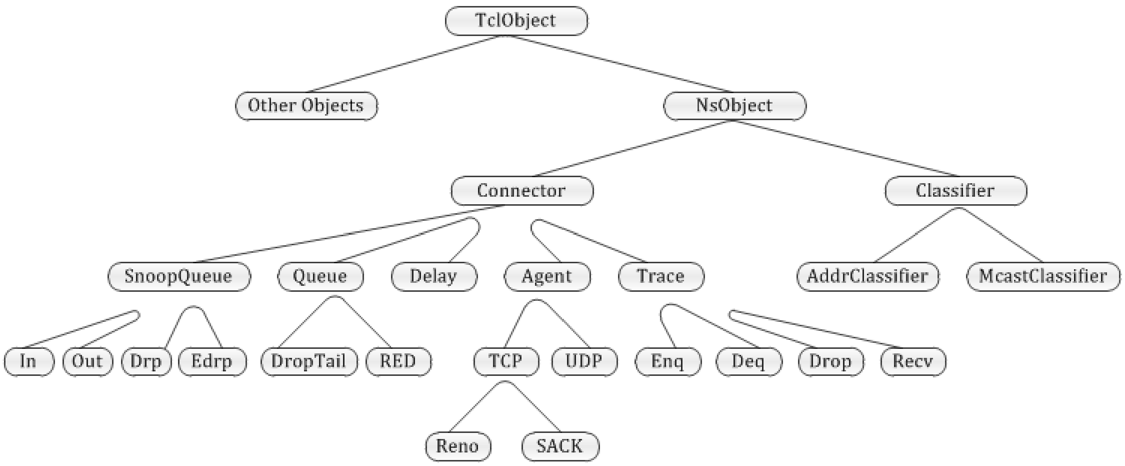


Figure 5.1: Class Hierarchy in NS-2

In order to create simulation scripts, the user must instantiate simulator objects and parametrize them in OTcl language. These objects are instantiated by the interpreter and normally mirrored to a corresponding object in the compiled C++ class hierarchy. The NS-2 uses two languages because it aims to accomplish two types of requirements. One of these requirements is to possess a programming language able to efficiently manipulate large data sets and the other requirement is to have a programming language capable of giving a the user quick interaction in the manipulation and configuration of different simulation scenarios. C++ language is fast to run but slower to change whereas OTcl runs slower but can be changed quite quickly.

In order to accomplish the model architecture implementation on NS-2 the addition and modification of some files in the NS-2 library was necessary. These extensions and modifications were made in the NS-2 version 2.29 and installed in the Red Hat operating system

[2]OTcl is an object oriented extension of Tcl/Tk created by David Wetherall at MIT, to be used in network simulator NS-2

with compiler gcc version 4.3.0. To extend the NS-2 library with the model architecture the following changes and modification were necessary: :

- the addition of new mobility processes in the F-HMIPv6 and QoS context transfer during handovers;
- the addition of new functionalities in DiffServ QoS model such as flow marking over an IP tunnel;
- the implementation of a signaling protocol in order to make the application QoS requests ;
- the implementation of the new resource management function;
- the incorporation of different access technology interfaces into a single node (ethernet, bluetooth, WiFi and UMTS);
- the integration of all the functionalities mentioned above towards the global model functionality objective.

The first step in the implementation of proposed model was to extend the NS-2 with the available patches for F-HMIPv6, DiffServ model and heterogeneous access technologies. Then, the second step was to add more functionalities in the F-HMIPv6 and DiffServ mechanisms and to deploy the new resource management function. The last step was the integration of these extensions regarding the overall architecture functionality of the proposed model.

In order to carry out the F-HMIPv6 implementation on an NS-2 plataform the original F-HMIPv6 patch, developed by Robert Hsieh [128] for ns-2.1b7a version, was modified to work in the ns-2.29 version.

The F-HMIPv6 patch was also extended to work with n mobile nodes since the original was limited to a single mobile node. This extension was provided the necessary support to make extensive evaluations of micro-mobility management schemes with n mobile nodes.

In order to have a better understanding of how wireless extensions have been processed in NS-2 [129] and also before starting to explain how the wireless model was implemented in NS-2, some important aspects such as creating a simple wired topology and how nodes are connected to form links will be briefly described beforehand.

Each NS-2 simulation requires only a single instance of the class simulator to control, operate and store all internal references to each network element.

The NS-2 software supports unicast and multicast nodes within the simulations, however, by default it uses unicast nodes. The multicast option must be used in order to make multicast simulations. . Nevertheless, being as the multicast nodes outside the scope of this research work, only the unicast structure will be described. Figure 5.2 (Figure reprinted from [129]) shows the programing structure of a unicast node. This structure is formed by two TclObjects: the address classifier (*classifier_*) and a port classifier (*port_*). These classifiers are responsible for the distribution of the incomig packets to the target agent or to the outgoing link.

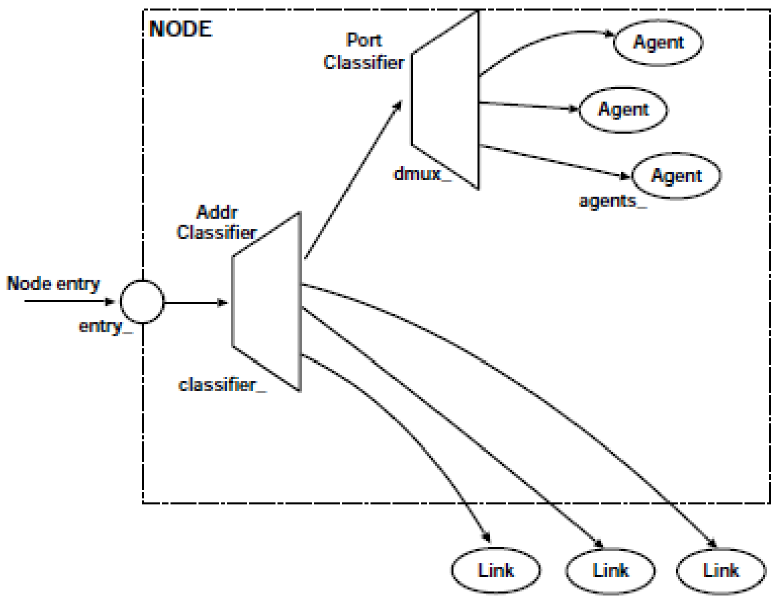


Figure 5.2: Structure of Unicast Node

When a node is created in a TCL script, the instance procedure of node constructs at least the following components:

- an *address* or *id_*, which identifies a node in the simulation,
- *neighbors_*, which contains a list of neighbors,
- *agent_*, which contains a list of agents,

- *nodetype_*, which identifies the node type, and
- a routing module.

The procedures that configure an individual node can be classified into:

Control Functions The *$node entry* returns the entry point for a node, where the *entry_* node instance variable is responsible for storing the reference of this element. This element is the first which will handle the packets arriving at node. For unicast nodes, the reference stored in the *entry_* variable corresponds to the classifier address (*classifier_*). Another control function is the *$node reset* which is responsible for resetting all the agents attached to the node.

Address and Port Number Management The procedure *$node id* returns the node number assigned to the node. The class simulator stores an instance variable, *node_*, which is indexed by the node id, and contains a reference to the node with that same *id*. The procedure *$node agent (port)* returns the handle of the agent on the specified port. The procedures *add-route* and *add-routes* are used to add routes for populating the *classifier_*.

Agent Management The procedure *attach{}* adds an agent to the list of *agents_*, assigns the agent a port number, sets the agent a source address, sets the target of the agent to be the *node entry{}*, and finally adds a pointer of the agent to the corresponding slot in the *dmux_* classifier. In order to remove a given agent from *agents_* list the procedure *deattach{}* must be used.

Tracing Neighboors Each node possesses a list of its neighbor nodes in its instance variable, *neighbor_*. The procedure *add-neighbor{}* adds a neighbor to the list and the procedure *neighbors{}* returns the list of neighbors.

In a network, the function of a node when receiving a packet is to inspect the destination address and then map that address to an outgoing interface. In the NS-2, this task is made by a *classifier* object. The *classifier* allows the retrieval of the packet reference to some simulation object based on the same criteria. The *classifier* has a table indexed by a slot number with the simulated objects. Therefore, the *classifier* function is to determine the slot number associated to the received packet and forward that packet to the object indexed by that slot. As can be seen in Figure 5.2 a unicast node by default contains one address classifier and one port classifier. Thus, when the objective is to add more functionalities to the node more classifiers must be added to the basic node. For instance, the hash classifier is an object used for classifying a packet as member of a particular flow and can be easily adopted to add more funcionalities to the basic node.

Another important aspect in the definition of a network topology is the creation of links that inter-connect nodes in a network. For creating a simple point to point link, the NS-2 provides the class *SimpleLink* which is a class derived from class *Link*. The class *SimpleLink* forms an unidirectional link from one node to another. The *Link* is defined with the following five instance variables (see also the Figure 5.3, Figure reprinted from [129]):

head_ the *head_* is the entry point of link.

queue_ the *queue_* contains a reference to the link queue.

link_ the *link_* contains a reference to the model which characterizes the link in terms of delay and bandwidth.

ttl_ the *ttl_* contains a reference to the element which calculates packet life time.

drophead_ the *drophead_* contains a reference to the element which is the head of the queue that will process the link drops .

While nodes are composed by classifiers, links are built by a sequence of connectors. The connectors may generate data either to the *target_* neighbor, or to the *drop-target_* to be droped.

A derived class from the connector base class is the queue class. The purpose of this class is to hold and eventually mark or discard packets. The implementation of packet schedulings and buffer managements are made inside this class.

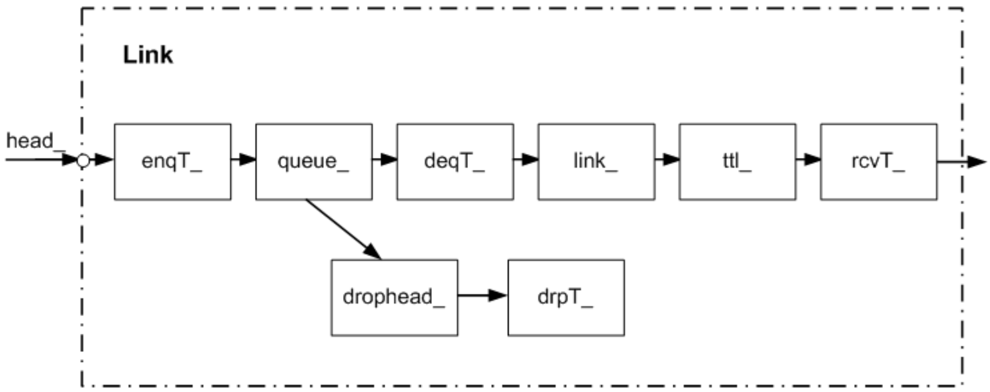


Figure 5.3: Instance Variables that Compose a Unidirectional Link

The Differentiated Service mechanisms were integrated in the NS-2 software since the ns-2.1b8 version, this software module of the Differentiated Services mechanisms was originally implemented by the Advanced IP Networks group in Nortel Networks [130]. The DiffServ queue implemented , i.e, class *dsREDQueue* is derived from the base class *Queue*. The class *dsREDQueue* added the following new features into NS-2 library:

- multiple physical RED queues on a single link
- multiple virtual queues within a physical queue each one with an independent set of parameters.
- capacity to determine which virtual and physical queue a packet should be enqueue based on its DiffServ code point.
- capacity to determine from which virtual and physical queue a packet should be dequeue based on the selected scheduling scheme.

Just as in a DiffServ domain of a real nework, the core and the edge router have different roles. Each role is defined in NS-2 within its own class which are *coreQueue* and *edgeQueue* classes. These classes are derived from the *dsREDQueue* class. The packets are marked in the *edgeQueue* class with a given code point which is determined according to the policy defined in the *policy* class and its sub-classes. Then, the packet is placed into the corresponding virtual and physical queue. By default, in NS-2 a DiffServ policy is established between the source and the destination node. Therefore, all flows matching the source-destination pair will have the same perhop-behavior treatment (PHB). Each policy has a source destination pair associated to its policy as well as a meter type, and an initial code point.

Therefore, the DiffServ functionality will be captured in a Queue object called *dsQueue* and in a *policy* class. The *policy* class handles the creation, manipulation, and enforcement of edge router policies.

The agent concept is another important concept in the design of an architecture in the NS-2. Agents act as endpoints in the network layer where packets are generated and absorbed. They are used to implement protocols in any layer level. The *agent* class includes several internal states to assign several fields to simulate a packet before it is sent, such as the node address, destination address, packet size, packet type, IP flow identifier, IP priority field, packet flags and IP ttl value. These variables can be modified by any class derived from *agent* class. To create and receive packets through an *agent* class, the NS-2 provides the *allocpkt*() and *recv*() methods. In order to The implementation of a new agent includes the following steps:

- choose its inheritance, and define the class
- define the *recv*() and *timeout*() methods
- define OTcl linkage functions
- develop the OTcl code for accessing C++ agent code

The objects from *Packet* class correspond to the basic unit of exchange between all objects in the simulation. Figure 5.4 (Figure reprinted from [129]) illustrates an object from *Packet* class.

This class consists in a generic pointer where packet header fields are stored, a pointer to packet data, and a pointer to next packet in the list or in the queue.

To create new protocols it may be necessary to define a new packet header or extend an existing header with some additional fields. By assuming that the creation of a new header is necessary, the following steps should be performed:

- create a new structure with the name of fields, define the *offset_* and the access methods

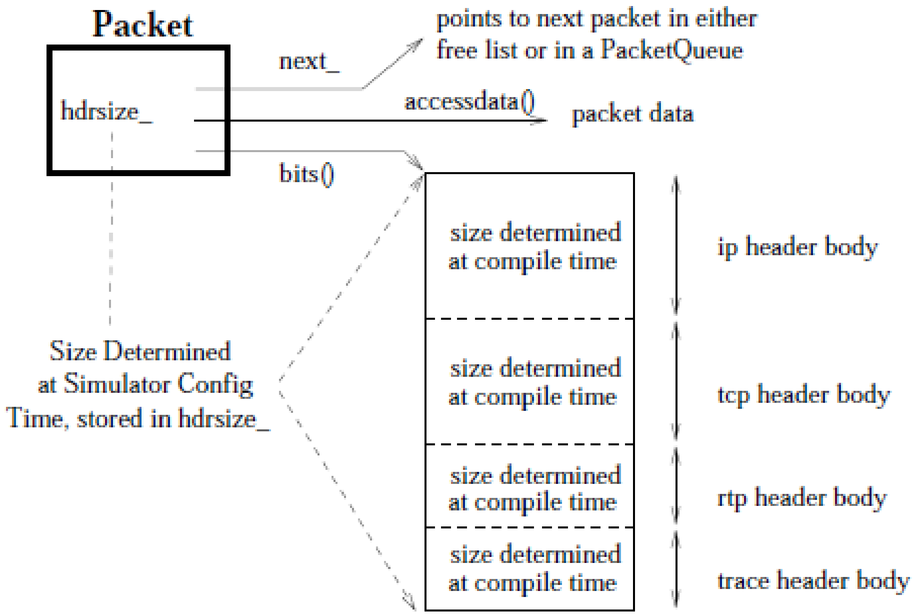


Figure 5.4: An Object from *Packet* Class

- define member functions for needed fields
- create a static class to perform OTcl linkage
- enable the new packet header in *~ns/tcl/lib/ns-packet.tcl*

5.2 Wireless Networks in NS-2

After presenting some of the most important concepts in order to better understand the NS-2 operations, the wireless implementation in NS-2 will be described. The first mobility extension to NS-2 was made by the CMU's Monarch groups. This wireless model enables NS-2 with simulations of wireless LANs and ad-hoc networks. However, this model does not permit simulations using wired and wireless nodes at same time, thereby SUN (Sun MicroSystems - Charlie Perkins et al. [131]) was extended to this model with the wired-cum-wireless feature. This new feature enables the simulations of using both node types in the same simulation. The SUN was also integrated in the Mobile IP protocol in wireless models enabling the Mobile IP to run over mobile nodes.

Figure 5.5 shows the historical evolution of the wireless model in NS-2.

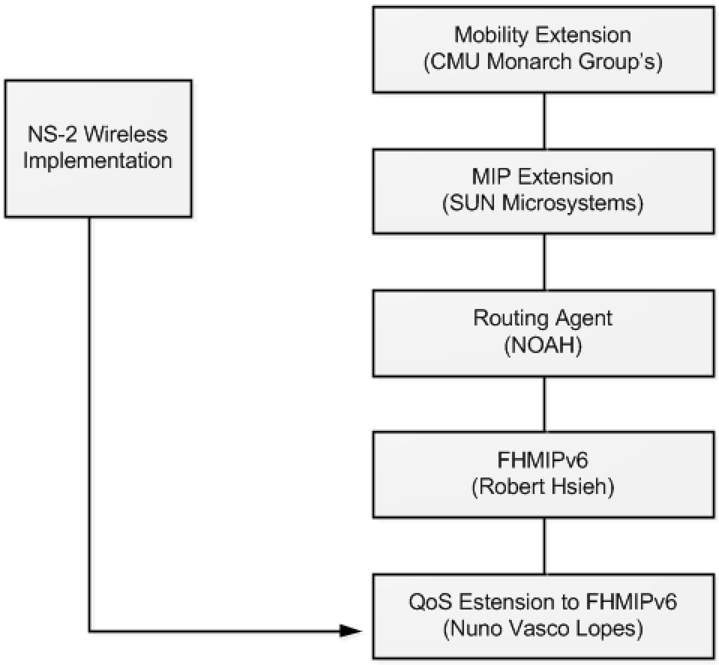


Figure 5.5: History of Wireless Model in NS

As a mobile node in NS-2 derives from the basic node object it can be said that it is a basic node with new wireless functionalities. The main difference between them is the fact that instead of a mobile node being connected to a link or to other mobile nodes, it transmits signals to and from a wireless channel. The mobility features in NS-2 are implemented in C++ and in the OTcl code. Figure 5.6 (Figure reprinted from [129]) illustrates the mobile node structure.

The network components of a mobile node consists in a Link Layer (*LL*), an *ARP* module connected to *LL*, an interface priority queue (*IFq*), a mac layer (*MAC*) and a network interface (*netIF*) connected to a channel. These components are created and integrated in order to work together in OTcl. A brief description of each component will be made now.

Link Layer The link-layer object is responsible for simulating the data link protocols and setting the *MAC* destination address in a packet MAC header. This task includes finding the next hop node and resolving the IP address in the correct *MAC* address (*ARP*). The mobile *ARP* module resolves all IP conversions. For down packets, the *LL* hands the packets to interface queue, for incoming packets the *MAC* layer hands up to the *LL* which then sends them to the *entry_* point.

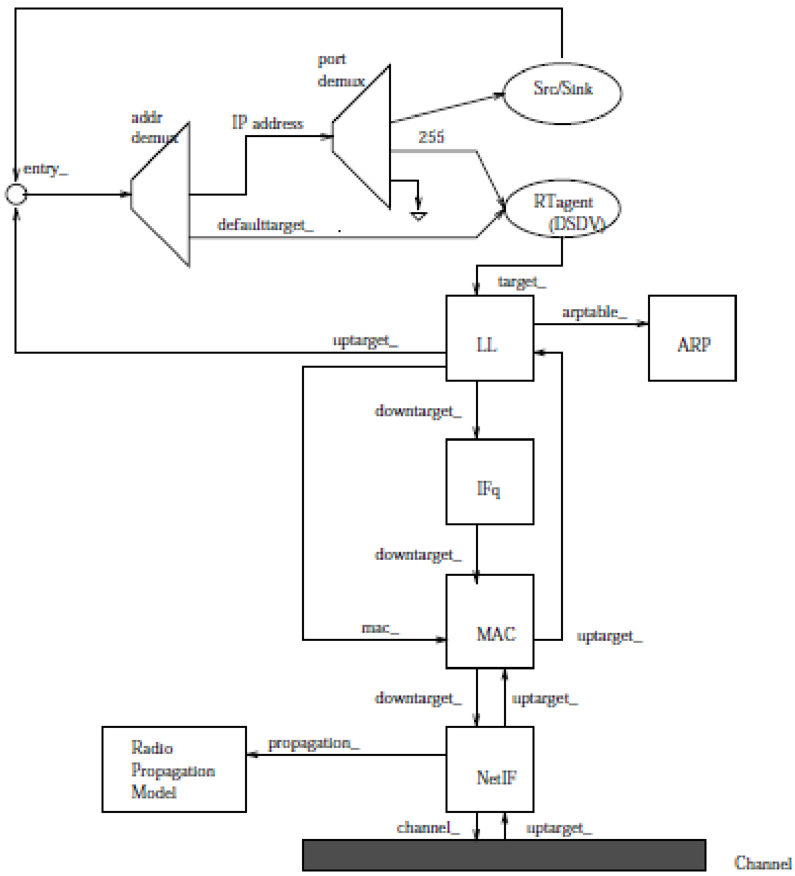


Figure 5.6: Structure of a CMU Mobile Node on NS-2

ARP This protocol receives queries from *LL* and conducts the IP conversions and writes the *MAC* address destination into the *MAC* header of the packet. The packet is then inserted into the interface queue.

Interface Queue The *IFq* is implemented with a priority queue which provides higher priority to routing protocol packets.

MAC Layer The NS-2 releases prior to ns-2.33 version use the IEEE 802.11 implementation which has been designed by CMU.

Network Interfaces The network interface simulates the hardware interface which is used by a mobile node to access the channel. The wireless interfaces are implemented in the *Phy/WirelessPhy* class. The network interface implementations are located in *~ns/phy.{cc,h}* and *~ns/wireless.{cc,h}*.

Currently, the basic NS-2 installation has implemented two MAC layer protocols, namely the 802.11 and TDMA, and four ad-hoc routing protocols which are DSDV, DSR, AODV and TORA.

The mobile nodes described so far support multi-hop ad-hoc networks however, they do not run the Mobile IP when wireless networks are connected to wired nodes, therefore SUN Microsystems extended the CMU wireless model to overcome this limitation. The main difficulty in the implementation of these extensions was to figure out how to make the routing from the wireless part to wired parts and vice-versa. Usually, the routing information is made based on the connectivity of the topology. This means that it is made based on how nodes are connected to each other through a link.

However, the link concept does not exist in mobile nodes. They need to use their own routing protocols. Therefore, a node called the base station node was created to serve as a gateway between wired and wireless domains.

The wireless domains along with its base stations have a unique domain address assigned to them. All the packets destined to a mobile node go through the base station attached to the domain of that mobile node. Mobile nodes route their packets to an outside domain through its assigned base station.

The struture of the *MobileNode/MIPBS* node is shown in Figure 5.7 (Figure reprinted from [129]). The *MobileNode/MIPMH* structure is similar to the MobileNode/MIPBS but without the encapsulator and decapsulator agents. In a scenario with wired and wireless nodes, the structure of mobile nodes and base stations are similar because both require hierarchical addressing/routing.

The Mobile IP implementation in NS-2 consists in a Home-Agent (HA), Foreign-Agents (FA) and Mobile-Hosts (MH) moving between their HA and FAs. The HA and FA are base stations and MHs are mobile nodes. Mobile IP extension methods and procedures are implemented in *~ns/mip.{cc,h}*, *~ns/mip-reg.{cc,h}*, *~ns/tcl/lib/ns-mip.tcl* and *~ns/tcl/lib/ns-wireless-mip.tcl*.

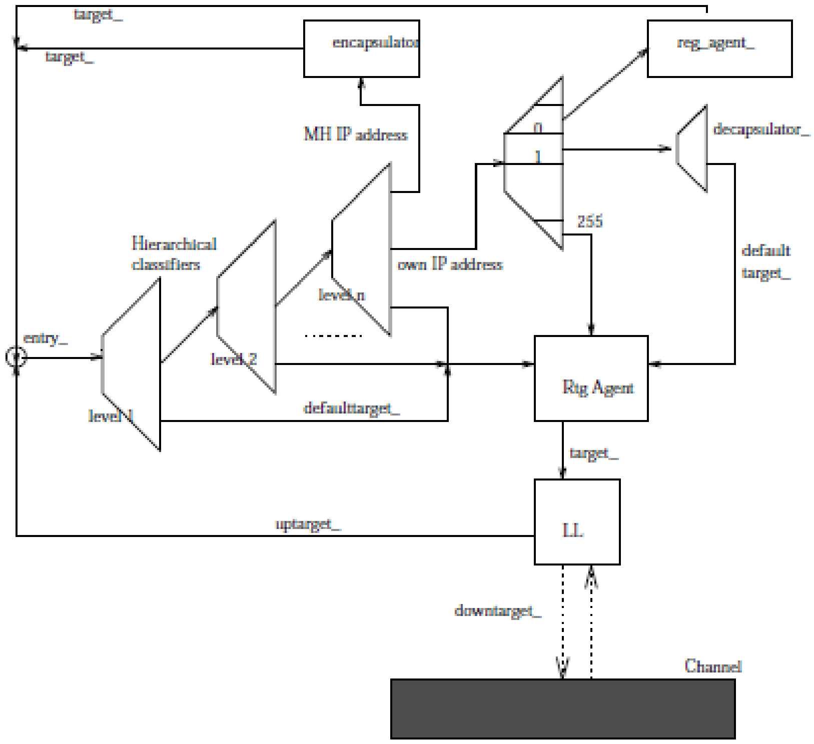


Figure 5.7: Structure of Mobile IP Base Station Node

The HA and FA are defined in the NS-2 code as *MobileNode/MIPBS* possessing a registering agent (*regagent_*) which is responsible for sending out beacon messages to mobile nodes, encapsulating and decapsulating packets, and replying to mobile node solicitations.

The MH nodes are defined as *MobileNode/MIPMH* having also a *regagent_* to receive and respond to beacon messages as well as make solicitations to base stations.

The *MobileNode/MIPBS* periodically broadcast beacon and advertisement messages to MHs. If an MH makes a solicitation to a base station, the base station sends an advertisement message directly to the MH that made the request. The beacon messages are used by the MH to extract the new CoA (Care-of-Address).

Therefore, whenever an MH moves from one base station to another its CoA also changes. When a base station receives a registration request (*reg_request*) it verifies the HA of the MH that is requesting the registration. If not, the FA is placing the request, thus it must send a message with a registration request to its decapsulator and forward it to the HA of the MH.

All the packets destined to MH that reach the HA are tunneled through the encapsulator which encapsulates the IP with the current CoA of the MH. When the FA receives these packets, it sets up its decapsulator and removes the encapsulation IP header and sends it to MH. The data sent to wired nodes from an MH is routed by the MN's base station.

5.3 Model implementation

The model implemented on NS-2 encompassed the addition of new features on the Nortel's DiffServ module and Robert Hsieh's F-HMIPv6 patch as well as the development of a new resource management function.

The code developed during the model implementation was carried out in both C++ and OTcl programming languages. The next paragraphs provide a detailed description regarding the new features added to the NS-2 simulator.

i) - DiffServ Mechanisms

Being as a DiffServ policy in NS-2 is established between the source and destination node, this results in a problem for wireless simulations because of both IP tunneling, which changes the IP source address, and handovers which change the IP destination address (CoA).

Therefore, to overcome the problem caused by handovers, the code of class *policy* was changed in order to add a DiffServ policy also based on *flow_id*. The IP tunneling problem was solved by assigning the same DiffServ policy mark contained in the packet header of data packet to the encapsulator header. This way, the DSCP information in the packet header is propagated to the outer IP header allowing DiffServ's packets to be recognized in the presence of IP tunnels by interior DiffServ routers [11].

ii) - F-HMIPv6 Mobility Management

The basis code used for the implementation of Fast over Hierarchical Mobile IPv6 micro-mobility approach was the Robert Hsieh's F-HMIPv6 patch. This patch implements a new mobile agent of the HMIPv6 protocol, called the MAP agent for providing local registration functionalities on MAP. The schematic structure of the MAP node is showed in Figure 5.8. The adopted strategy was to simulate in the NS-2 the behavior of 802.11b standard protocol with the MAP agent making an emulation

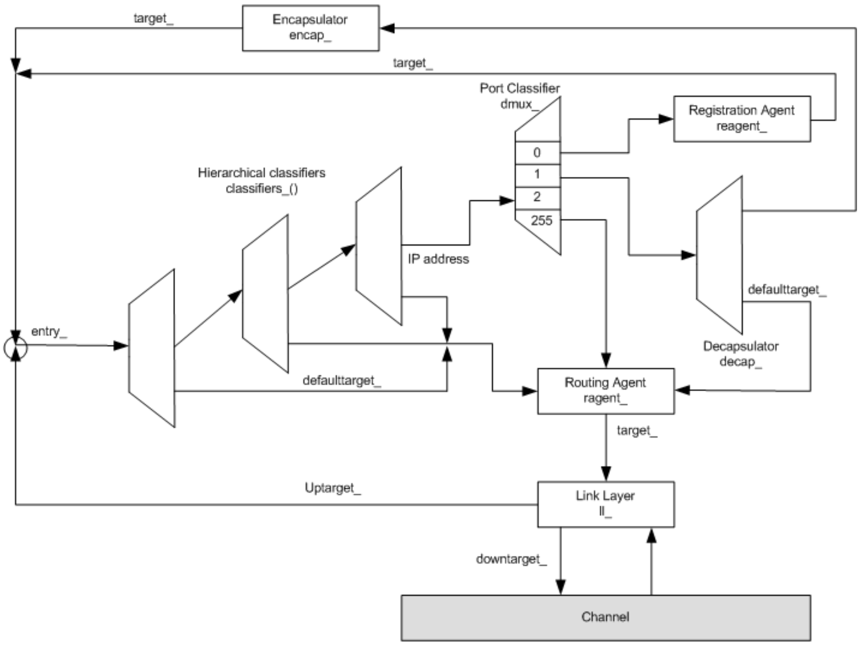


Figure 5.8: MAP Node Structure

of this standard instead of making a complete installation. This was due to the fact that the wireless simulation model in NS-2 was developed for ad-hoc networks, i.e, for broadcast mode. Therefore, the strategy was to add a new Connection Monitor (*CMon*) entity in the MN structure (see Figure 5.9) between the port classifier (*dmux_*) and the receiving agent(s) (e.g. TCP/UDP). The *CMon* entity must control the MN's packets reception during the existence of a flow. Thus, when an MN makes one L2 handover to a new access network, *CMon* is set for dropping any of the received packets until the L2 handover is completed. This way, it is possible to treat the receptions of control messages as if it were operating in the periodic channel scanning mode. This entity also emulates the channel changes which occurs when the MN is switching between two access networks by blocking the agents from receiving any packets during this period.

The Robert Hsieh's F-HMIPv6 patch was extended in this research work for working with n mobile nodes and for conveying the MN's QoS context in the Handover Initiate (HI) message of the FHMIPv6 mobility management.

In order to implement the proposed model signalization, the F-HMIPv6 mobility signaling messages HI and HAck were also changed so that they could be sent between the access routers.

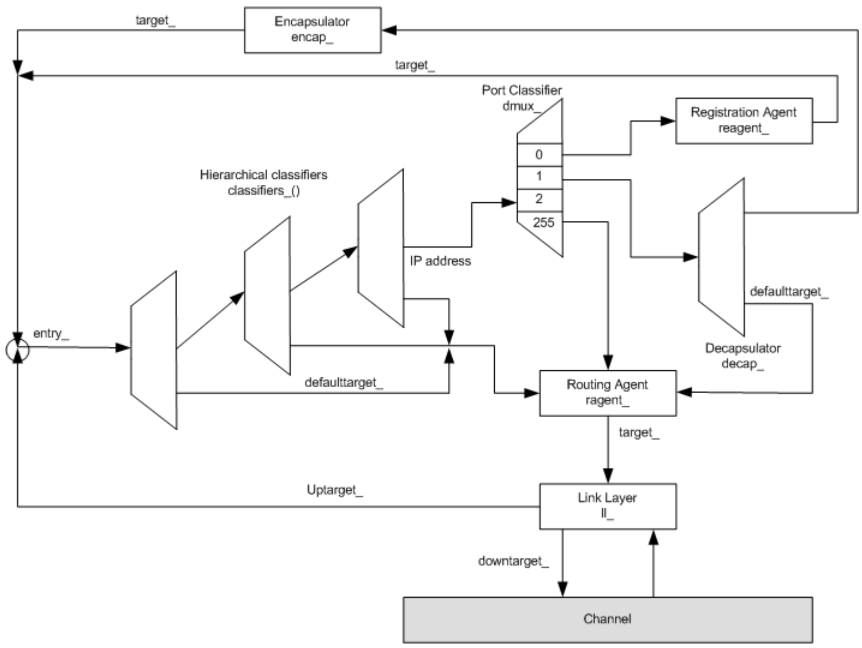


Figure 5.9: MN Structure with the F-HMIPv6 Implemented

$\#Add\,MIH\;function\,to\,a\,function$
$set\,mihf\;[\$node\,install - mih]$

$\#Link\,MIH\;function\,to\,MAC$
$set\,mac\,[\$node\,set\,mac - (0)]$
$\$mac\,mih\,\$mih$
$\$mih\,add - mac\,\$mac$

Figure 5.10: MIH Configuration in OTcl

The mobility management of F-HMIPv6 were coupled and synchronized in order to work with the new resource management function of the DiffServ model. The resource management function is triggered in AR when it receives an HI message or an SA-REQ message.

The modifications were made specifically on the *MIPBSAgent* agent of the *MobileNode/MIPBS* node. The *MIPBSAgent* agent is responsible for passing the MN's QoS context to the resource management function of the new access router. After the new access router receives the MN's QoS context sent from the previous access router, its resource management function proceeds to the admission or rejection of mobile node flows.

iii) - Heterogeneous Handovers

This functionality was added to the installation of National Institute of Standards and Technology (NIST) patch. The patch implements a media independent handover (MIH) in the NS-2 for providing seamless mobility to the MN between heterogeneous mobile technologies [132].

The MIH function is implemented as an agent and denominated as an *MIHAgent*. This Agent possesses a list of the local interfaces in order to get their status and also controls their behavior. Figure 5.11 shows a high level representation of the MIH function with others node components.

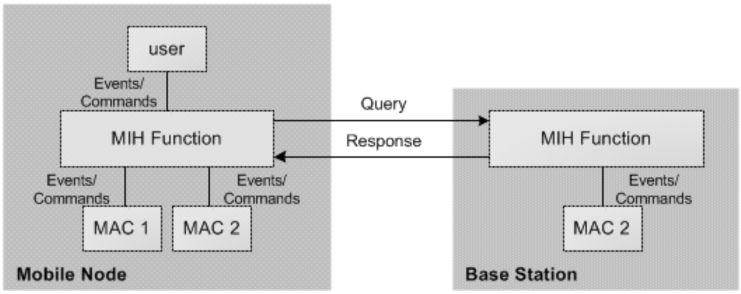


Figure 5.11: MIH Diagram Design

The MIH functionalities were added to NS-2 resorting to an abstract class called *MIHUser*. The Figure 5.10 illustrates the OTcl code for adding the MIH function to a node, and the corresponding code to bound the MIH function to MAC interfaces.

In order to create a multi-interface node, the NIST designed a virtual node for linking nodes configured with different access technologies. These nodes were considered virtual node interfaces.

Figure 5.12 shows the virtual node structure from NIST integrated with the previously described F-HMIPv6 structure node (illustrated in Figure 5.9). NIST adopted the approach of creating a virtual node because the EURANE[3] package for UMTS, the UCBT package from university of Cincinnati[4] and IEEE 802.16e developed by

[3]The EURANE has been developed by the European Commission 5th framework project SEACORN and comprises the addition of three new nodes, namely the Radio Network Controller (RNC), Basestation (BS) and the User Equipment (UE).
http://eurane.ti-wmc.nl/eurane/

[4]UCBT - Bluetooth extension for NS2 at the University of Cincinnati is a ns-2 based Bluetooth network module which simulates the Bluetooth network operations.
http://www.cs.uc.edu/~cdmc/ucbt/

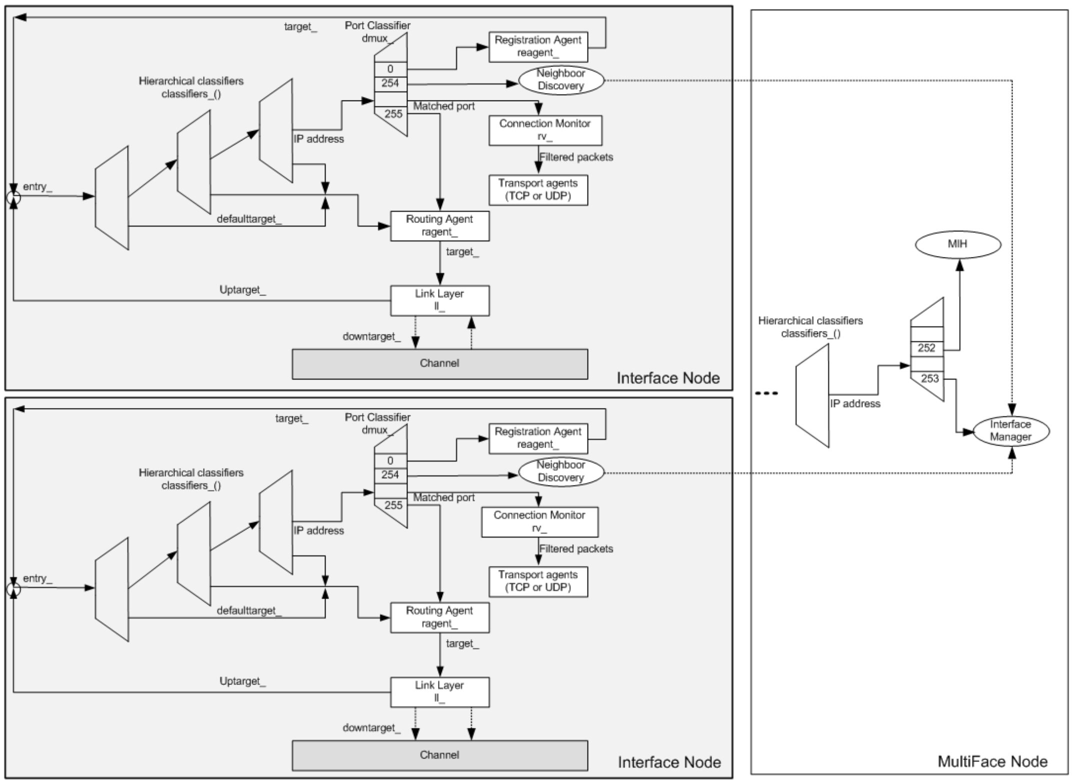


Figure 5.12: NIST Virtual Node Design

NIST do not follow the same node structure of the basic node structure defined in NS-2 therefore, its integration in the basis node structure would be quite difficult to implement.

The *neighbor discovery* (ND) agent in NIST structure node provides the layer 3 movement detection. By analysing the routing advertisements, an MN can determine if it is in a new base station.

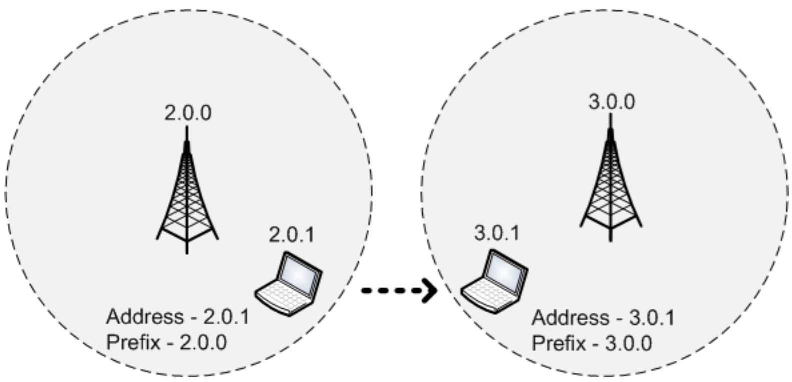


Figure 5.13: MN Handover with NIST Implemented

Figure 5.13 shows an example of how this occurs. This implementation also supports router solicitations from MNs allowing an MN discover a new base station after an handover.

iv) - Resource Management Function

The current NS-2 implementation of measurement-based admission control algorithm in *~ns/adc/*.{cc,h}* was modified for working on the *MobileNode/MIPBS* nodes (access routers). The MBAC was also changed in order to support two levels of granularity in the control of admissibility, i.e., per flow for new flows and per class for handover flows.

For identifying a packet as member of a particular class, in order to allow packets beloging to the same class to be forwarded to the estimator object of the class, the addition of more classifiers into MobileNode/MIPBS node of the NS-2 was necessary.

For the effect, the hash classifier has been used to classify a packet as member of a class, so that packets marked with a given class can be forwarded to object estimator of that class. A description of the classes that were involved in the implementation of the resource management function in the NS-2 is made below.

Signaling Agent The signaling agents are attached to the CNs and ARs. They are responsible for making the QoS communications between them.

Signaling Agent Link Creates a new type of *Link*, called *saLink* class, derived from NS-2 class *connector*. When an *saLink* is created, an instance of the class *ADC* (admission control class) is created as well.

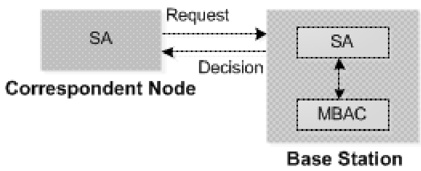
Admission Control The *ADC* class derives from *the NsObject* class. This class contains the admission control algorithm and an instance of the class *Estimator*.

Estimator The *Estimator* class is responsible for estimating the amount of aggregated traffic in each class as well as the amount of aggregated traffic being used in each class by a mobile node. The class *Estimator* has an instance of the class *Measurement*. This instance is used by the class *Estimator* for making, its traffic estimations based on the measures.

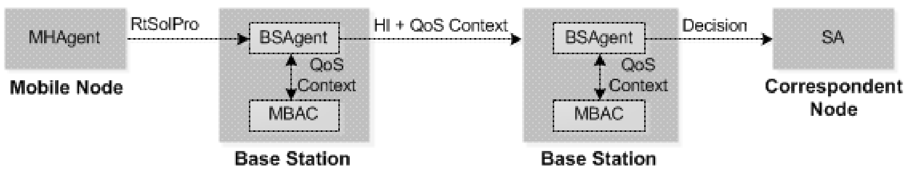
Measurement This class measures the amount of traffic being used in each DiffServ class as well as the amount of aggregated traffic that is being used in each class by a mobile node.

The admission control algorithm 5.1 implemented in the proposed model is based on a measure sum admission control algorithm. This algorithm has been endowed with a reallocation mechanism that is controlled by the hysteresis method.

The behavior of this algorithm has already been explained in the subsection 4.2.1 of Chapter 4.



(a) New Flows



(b) Handover Flows

Figure 5.14: Interoperation Between DiffServ, F-HMIPv6 and MBAC modules

Figure 5.14 shows the interoperation between the DiffServ, F-HMIPv6 and MBAC in the NS-2.

For new flows, see Figure 5.14a, the *SA* (Signaling Agent) generates a message containing the application's QoS request to the network. All SA agents along the path receive the request message and then pass the information to MBAC in order for it to proceed with the admission control decision.

The decision is sent towards the mobile node (receiver). The mobile node sends back a response message with the decision to the correspondent node (source).

For handover flows, see Figure5.14b, the *MIPMH* (Mobile IP Mobile Host) agent sends a Router-Solicitation-for-Proxy (RtSolPro) to *MIPBSAgent* of pAR. The pAR sends an HI message with the MN's QoS context, which has been extracted from the estimator at the time of handover, to nAR.

The *MIPBSAgent* agent of the new access router receives the HI message with MN QoS context and passes it to the MBAC module in order for it to proceed the admission control decision.

The decision is sent in a response message to the correspondent node.

The packets belonging to flows that were admitted by the MBAC, see Fig. 5.14, are placed into the virtual queue of the *dsREDQueue* class that corresponds to DiffServ codepoint that packet are marked with.

In order to create an *ADC* object from OTcl front-end, an *AdmissionControlLink* class in OTcl has been created. This is a subclass of *SimpleLink* class. This class provides the access and control of C++ objects and variables inside the MBAC module.

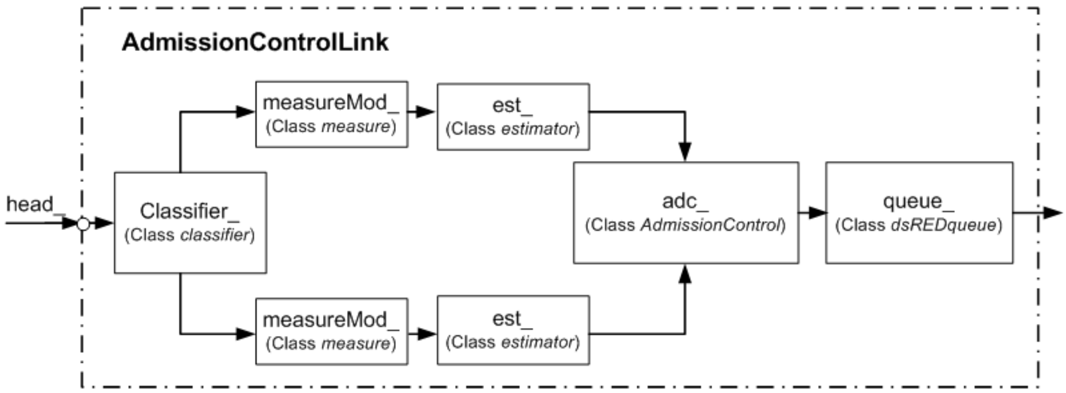


Figure 5.15: Instance Variables of AdmissionControlLink

The instances that define an *AdmissionControlLink* link are illustrated in the Figure 5.15.

$$\vdots$$
$$\$ns \quad simplex-adc-link \quad \$node \quad \$BS \quad Rate \quad Delay \quad dsRED \quad SALink \quad \$adc \quad \$est$$
$$\vdots$$

Figure 5.16: Admission Control Link in OTcl

The Figure 5.16 illustrates the command to create an *AdmissionControlLink* class instance in OTcl, with the respective estimator and admission control algorithm attached to a link with a DiffServ queue.

5.4 Data Processing

Simulation scenarios were designed and simulated in order to assess the proposed model. However, in the NS-2 simulator whenever a simulation is made, the simulation events results are stored in a trace file that is generated at the end of simulation. The trace files generated by NS-2 contain information reagrding all the events occurred during the simulation period in the nodes of the simulated topology. This results in a huge amount of unnecessary data, at least for this research study, per trace file. Therefore, if the number of scenarios to be simulated is very high, the quantity of information that must be processed will also be elevated. For this reason, a cluster of computers called SeARCH [5] essentialy for simulation data processing will be used.

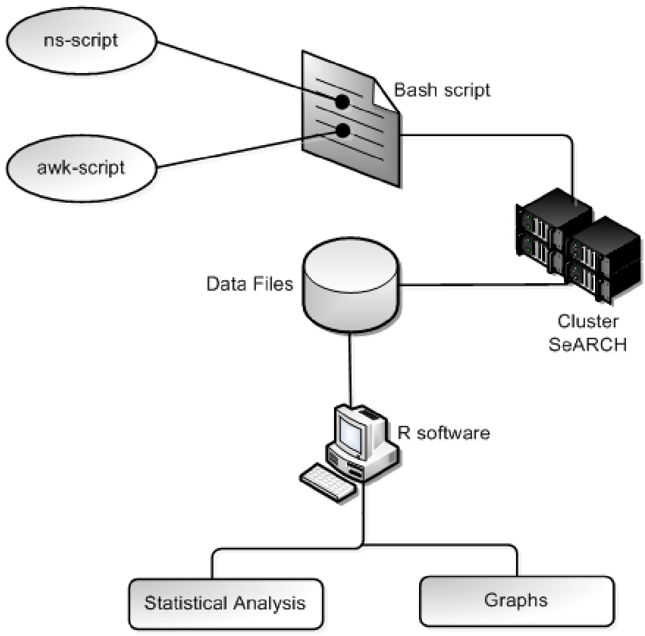


Figure 5.17: Data Processing

Figure 5.17 shows a diagram which illustrates how the data has been processed.

In order to run the several simulations in SeARCH a bash script containing an algorithm that processes n simulation iterations, applies a filter to the trace files to obtain only the

[5] SeARCH "Services and Advanced Research Computing with HTC/HPC clusters". (Funded by FCT, CONC-REEQ/443/EEI/2005, co-funded by Program POCI 2010, via FEDER)

relevant information and stores the data files obtained from the filtered trace files into a local directory has been elaborated.

The AWK language has been used to filter the revelant data from trace files. The algorithm also produces some basic statistical results such as average, standard deviation, among others, which has been calculated in AWK.

Furthermore, the algorithm also generates some simple plots for a first view and interpretation which has been performed with gnuplot scrips.

After the simulations and data filtering in the cluster have been finished, the resulted data files storage in a local directory are then loaded and merged into the R software [6].

In the R software, the data has been organized into flat tables structured only with the key fields that has been considered important for this research work.

After, the various databases being loaded and organized in R, several R functions that have been developed for our research purposes have been used to generate graphics and to obtain important statistical information from the databases.

This working methodology has been adopted for the treatment of data generated by simulations permitted a more extensive statistical analysis and evaluations of the simulated scenarios.

[6] R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.
The current R is the result of a collaborative effort with contributions from all over the world. R was initially written by Robert Gentleman and Ross Ihaka—also known as "R & R" of the Statistics Department of the University of Auckland. Since mid-1997 there has been a core group with write access to the R source
http://www.r-project.org/

Algorithm 5.1 Admission Control Algorithm

$if\ (BWutilized_i + QoSContext_i \leq BWalocated_i)\ \{$

$BWutilized_i = BWutilized_i + QosContext_i$
$\Delta_i = 0$
$\Delta Total_i = 0$
$\Delta Classes = 0$

$\}\ else\ if\ ((i < 4) \wedge (BWutilized_i + QoSContext_i > BWalocated_i) \wedge \ldots$
$\ldots (BWutilized_i + QoSContext_i < BWalocated_i + \Delta Total_i))\ \{$

$df = (BWalocated_i + \Delta Total_i) - (BWutilized_i + QoSContext_i)$
$steps = floor\left(\frac{df}{\triangle min_i}\right)$
$\triangle_i = steps \times \triangle min_i$
$\Delta Total_i = \Delta Total_i - \triangle_i$
$\Delta Classes = \Delta Classes - \triangle_i$
$BWutilized_i = BWutilized_i + QoSContext_i$

$\}\ else\ if\ ((i < 4) \wedge (BWutilized_i + QoSContext_i > BWalocated_i + \Delta Total_i))\ \{$

$df = (BWutilized_i + QoSContext_i) - (BWalocated_i + \Delta Total_i)$
$steps = ceil\left(\frac{df}{\triangle min_i}\right)$
$if\ ((steps \times \triangle min_i + \Delta Total_i \leq \Delta max_i) \wedge (steps \times \triangle min_i + \Delta Classes_i \leq \Delta max_{BE}))\ \{$
$\triangle_i = steps \times \triangle min_i$
$\Delta Total_i = \Delta Total_i - \triangle_i$
$\Delta Classes = \Delta Total_i - \triangle_i$
$if (BWalocated_{BE} - BWutilized_{BE} \geq \triangle_i) \{$
$BWutilized_i = BWutilized_i + QoSContext_i$
$\}\ else\ \{$
$BWutilized_{BE} = BWutilized_{BE} + QoSContext_i$
$BWutilized_{BE} = BWutilized_{BE} - \triangle_i$
$\}$
$\}$

$\}\ else\ \{$

$reject\ (handover\ flows\ in\ the\ class\ ith)$
$\}$

Chapter 6

Simulation Results And Analysis

This chapter presents several simulations results regarding model performance and parametrization. For this purpose a simulation model for representing the architecture in real scenarios has been designed. The goal of the simulation model is to assess the performance improvement achieved by implementing the proposed model in mobile environments. The evaluation of the model is conducted with the respective data statistical analysis in order to validate and consolidate the conclusions. The deployment of the simulation model intended to achieve the following objectives: 1) choose the best rate estimator for the model's architecture; 2) evaluate the model and dynamic allocator to assess the influence of this element in the Model's performance; 3) evaluate the model performance under different parametrization values in order choose the best values based on objective criteria.

In summary, a detailed discussion regarding the efficiency of rate estimators in a mobile environment will be presented along with a trial of the model performance and dynamic allocator evaluation.

Finally, by the end of the Chapter proposes a criteria to parametrize the architecture, based on network stability, maximum bandwith utilization and on optimization algorithm. This criterion intends to provide a guide that allows network administrators to correctly parametrize the architecture´s input parameters. For this purpose, the performance analysis of the model with distinct parametrization values has also been conducted. It would be better to have a comparative analysis of the results obtained with this solution in comparison to any other solutions for QoS in mobile environments. However, the lack of the proposed solutions´ implementations have been a serious obstacle in addition to the lack of common metrics in similar solutions, the majority of which are based on the guaranteed service model.

6.1 Preliminary Study on Rate Estimators

In order to prevent QoS deterioration, the proposed model evaluates the impact of accepting new traffic in the wireless network. This evaluation and posterior decision of whether or not to accept any, or all, of this new traffic is based on a measurement based admission control procedure. Thus, all DiffServ class-of-service must be measured to then be used by the admission control algorithm as input parameters. Consequently, the rate estimator mechanisms will have an essential role in the efficiency of the QoS-aware overall architecture. Therefore, in an attempt to choose the best rate estimator for the global QoS architecture, two rate estimators - Time Sliding Window (TSW) and Exponential Moving Average (EMA) - were modified so that they were able to work in class basis and in MN's class basis and then be evaluated in a QoS-aware mobile simulation scenario.

The Time Sliding Window (TSW) estimator provides a running average bandwidth of traffic classes over a window of time (see Figure 6.1).

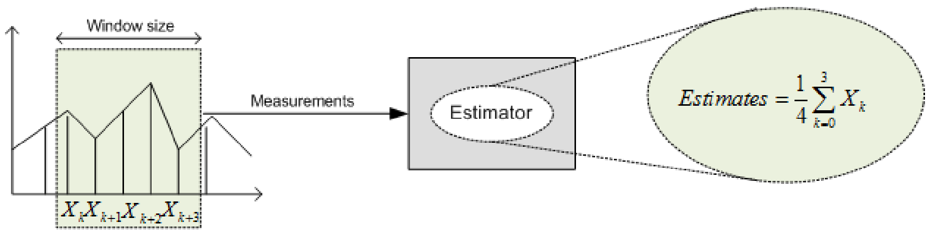


Figure 6.1: Time Sliding Window

It uses all packets in order to determine the current rate. This rate estimator takes into account burstiness and smooths out its estimate to the long-term measures within each class rate.

The Exponential Moving Average (EMA) applies weighting factors which decrease exponentially thus providing an increased importance to recent observations (see Figure 6.2).

An interesting discussion regarding the most appropriate parameter values for rate estimators can be found in [133] and [134].

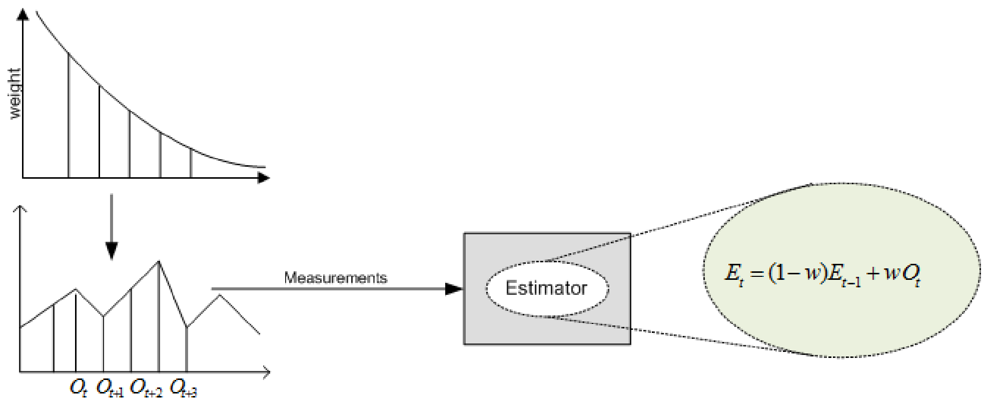


Figure 6.2: Exponential Moving Average

6.1.1 Simulation Scenario for Rate Estimators Evaluation

In order to study the reaction behavior of rate estimators during handovers, a simple topology was set up containing two access routers (pAR and nAR) and two MNs. Figure 6.6 displays the simulated scenario. Initially, one of the MNs is located in the pAR and the other is being served via nAR. Both are receiving CBR traffic marked with different DiffServ Classes. The traffic is originated from fixed correspondent nodes located in another DiffServ domain. The generated flows initiate at different points in time marking a time period between 0 and 80 seconds.

Mobile Node 1 (MN_1) traffic includes one 13kbps flow in class 1, two 15kbps flows in class 2, five 30kbps flows in class 3 and three 60kbps flows in class 4, leading to a grand total of (all classes aggregated) of 373kbps.

Whereas, Mobile Node 2 (MN_2) traffic includes one 18kbps flow in class 1, one 30kbps flow in class 2, two 20kbps flows in class 3 and two 40kbps flows in class 4, making a grand total (all classes aggregated) of 168kbps.

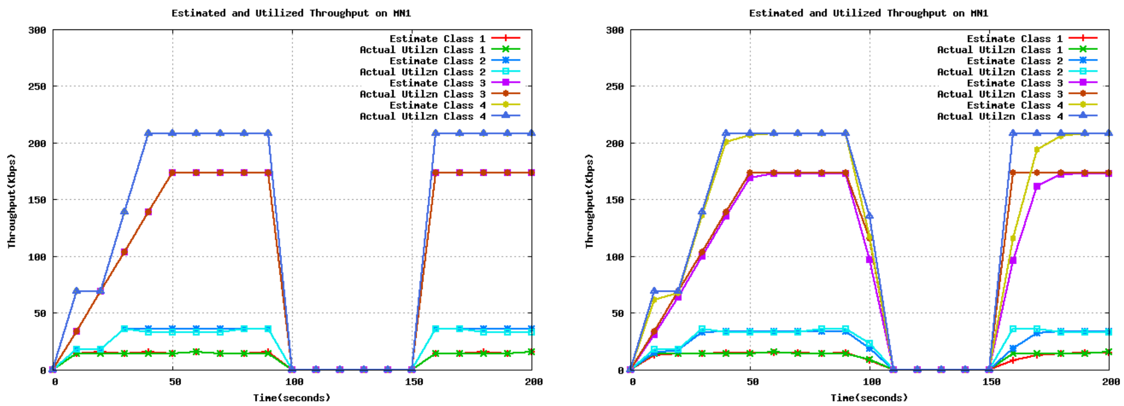
Once 80 seconds have passed, MN_1 starts to move at different speeds towards a region within the nAR scope. Simulation tests have been carried out using the mentioned topology as well as two different rate estimators used in measuring data traffic: Time Sliding Window (TSW) and Exponential Moving Average (EMA) rate estimators. For the experiments containing each rate estimator, the MN velocity during handover varied from 1m/s, to 3m/s and to 10 m/s.

Both rate estimators were evaluated and compared in order to analyze its behavior and determine which is the best in mimicking the traffic dynamics in order to take part in the overall architecture.

There has been a special concern in evaluating the performance of rate estimators during handover periods and assessing their responses. Therefore, only rate estimations for MN_1 and new access router (nAR) were analyzed and commented on. Results were analyzed in terms of traffic dynamics and rate deviation between estimated and real used.

6.1.2 Result Analysis on Rate Estimators

The selected speeds for mobile node handover correspond to the mobility speeds of: walking (3.6 Km/h), running (10.8 Km/h) and cycling (36.0 Km/h).



(a) Estimated and Utilized Throughput on MN1 with TSW Estimator

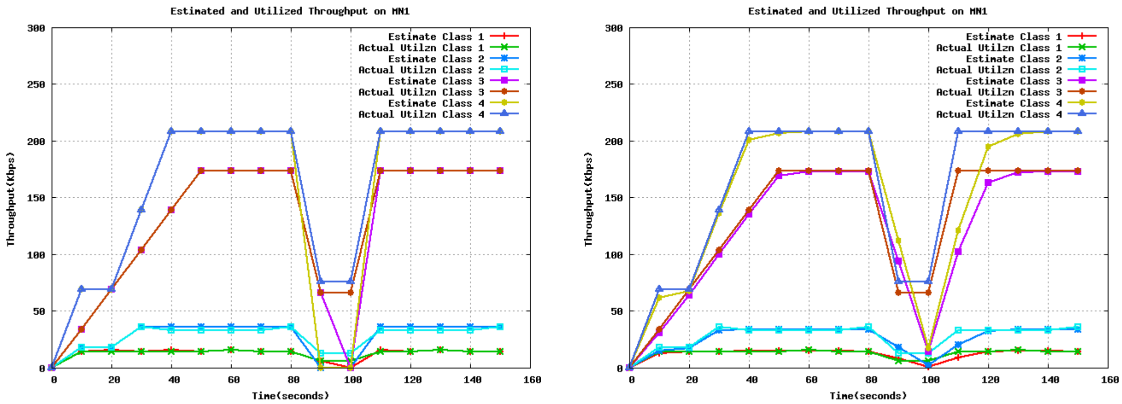
(b) Estimated and Utilized Throughput on MN1 with EMA Estimator

Figure 6.3: Estimated and Utilized Throughput on MN1 with velocity of 1 m/s

Figure 6.3 displays the bandwidth estimated and utilized by mobile node 1 when it moves with a velocity of 1 m/s towards new access router. As can be seen, the TSW estimator sightly under estimates classes with less traffic before the occurrence of handover, whereas the EMA estimator under estimates classes with more traffic.

Figure 6.4 shows the estimated and utilized bandwidth by mobile node 1 when it moves at a velocity of 3 m/s towards the new access router. When the mobile node is moving at the velocity of 3 m/s it receives traffic from both access routers because it arrives at the new access router before finalizing the simulation time (150 seconds). In the case of the TSW rate estimator, the estimation is almost equal to bandwidth utilization. There is only a little decrease in estimation values during handover but it rapidly recovers from this little shifting.

In the case of the EMA rate estimator, its estimation during handover is significantly delayed in relation to the current class load. Contrarily to the TSW rate estimator which



(a) Estimated and Utilized Throughput on MN1 with TSW Estimator

(b) Estimated and Utilized Throughput on MN1 with EMA Estimator

Figure 6.4: Estimated and Utilized Throughput on MN1 with velocity of 3 m/s

immediately follows the actual class bandwidth utilization, the EMA rate estimator only converges its estimation values to the utilized bandwidth at the end of the simulation.

As shown in Figure 6.5, at this velocity the TSW estimator was acceptably accurate in testimating the actual class bandwidth utilization, whereas the EMA estimator expresses a significant delay when it comes to following the current class traffic load (see Figure 6.5b).

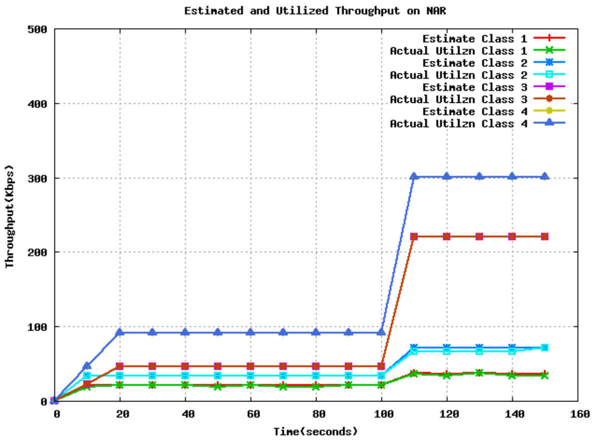
Even if the mobile node moves at a velocity of 10 m/s it shows a significant delay in achieving a reasonable accuracy as can be seen in Figure 6.5c.

6.2 QoS Architecture - Model Evaluation

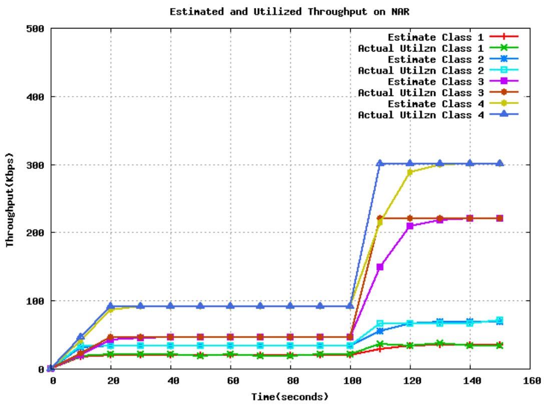
After elaborating an evaluation of rate estimators, the next step is aimed towards the objectives defined at the beginning of the Chapter: the implementation of the QoS architecture in the simulator in order to be able to evaluate it. Therefore, only the QoS signalization, DiffServ and admission control mechanisms of the proposed model are implemented in the NS-2 simulator in the next simulation scenario.

6.2.1 Simulation Scenario for Model Evaluation

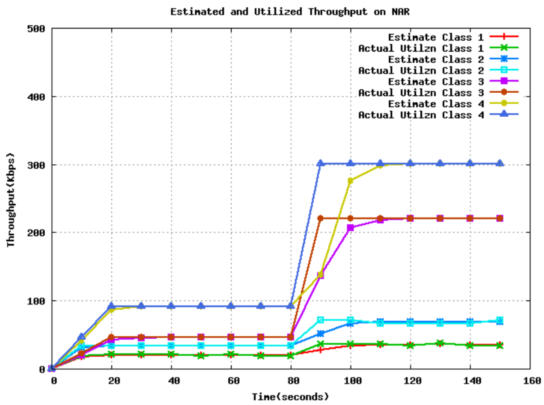
For this simulation scenario, topology was set up with: two ARs (pAR and nAR) and two MNs (MN_1 and MN_2). Figure 6.6 shows the simulation scenario used to evaluate the model without the dynamic allocator element.



(a) Estimated and Utilized Throughput on New Access Router with TSW Estimator with velocity of 3 m/s



(b) Estimated and Utilized Throughput on New Access Router with EMA Estimator with velocity of 3 m/s



(c) Estimated and Utilized Throughput on New Access Router with EMA Estimator with velocity of 10m/s

Figure 6.5: Estimated and Utilized Throughput on New Access Router with both Rate Estimators

The MNs are receiving Constant Bit Rate (CBR) flows from Correspondent Nodes (CNs) located at another DiffServ domain of the global internet. This traffic transmission is one to one (CN→MN). Each CN of the two CNs is generating CBR flows marked with a different DiffServ Code Point (DSCP).

Being as the bottleneck is in the last hop (wireless link), all flows will be accepted by precedent posts of AC until the AR. Initially, one MN is located in the pAR and the other MN is in the nAR. All flows start at different points in time, within the period of 0-80 seconds.

MN_1 traffic includes one 13kbps flow in class 1, two 15kbps flows in class 2, five 30kbps flows in class 3 and three 60kbps flows in class 4, making a grand total (all classes aggregated) of 373kbps. MN_2 traffic includes one 18kbps flow in class 1, one 30kbps flow

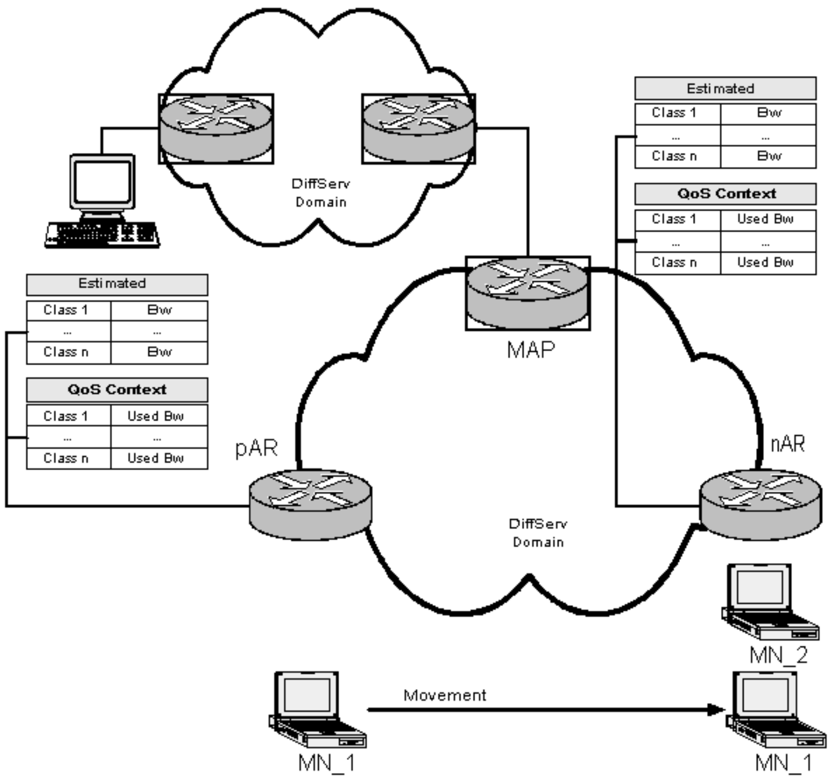


Figure 6.6: Architecture with per Class Tables

in class 2, two 20kbps flows in class 3 and two 40kbps flows in class 4, leading to a grand total (all classes aggregated) of 168kbps.

After 80 seconds have passed, MN_1 starts moving towards a region within the nAR scope. Therefore, if admission control is applied, some traffic classes may be eventually rejected so that pAR is able to move to nAR.

To evaluate the architecture without the dynamic allocator three distinct scenarios were implemented. The first scenario was implemented usingwith the proposed FMIPv6 and HMIPv6 integration.

The second implements the DiffServ over in F-HMIPv6 micro-mobility protocol for resolving the DiffServ signaling invisibility problem over IP . The standard DiffServ QoS mechanisms have been configured with four DiffServ classes that have been set up according to the QoS requirements of UMTS classes [12]. Here, the highest priority class (class 1) is configured with Expedited Forward (EF) service, the lowest priority class (class 4) is configured with Best-Effort (BE) service and the other two classes (class 2 and 3) are configured with Assured Forward (AF) service.

The third scenario was designed to illustratethe performance of the DiffServ resource management with the proposed dynamic QoS provisioning and admission control scheme.

i) Scenario A - F-HMIPv6;

ii) Scenario B - Scenario A + DiffServ over ;

iii) Scenario C - Scenario B + Admission Control.

6.2.2 Result Analysis on Model Evaluation

According to the proposed model, in the latter scenario when an MN intends to move to the new access router, its QoS requirements, i.e. MN's QoS context, will be submitted to an admission control decision in the nAR to assess its availability for receiving the new traffic. Consequently, only the mobile flows belonging to the accepted classes will be transfered.

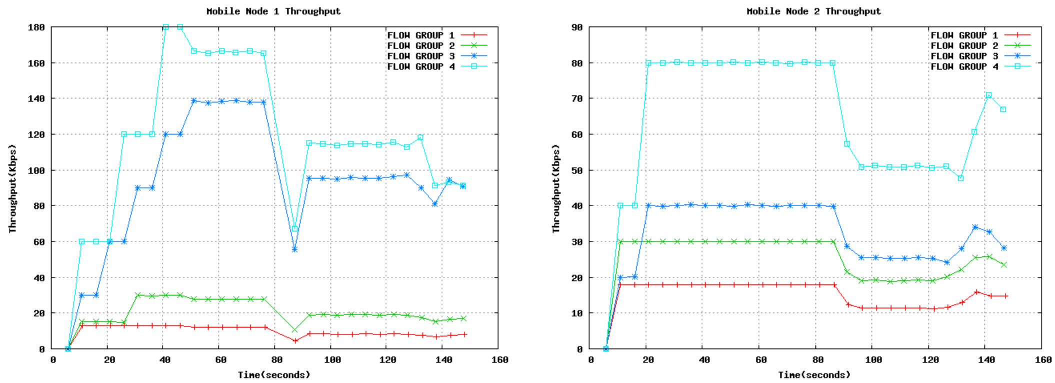
Despite the fact that link capacity for IEEE 802.11 was set with 1Mbps on NS-2, the available bandwidth on medium i.e., the maximum guaranteed throughput that can be transmitted between the base station and the MN without causing the disruption of any ongoing flow in the network, is only 35% (350kbps) of the base bandwidth [135, 136]. This is due to the carrier sense mechanism. In this situation, whenever a node needs to send a frame, itmust contend for medium access and it cannot transmit its frame until the medium is totally free [137].

For a clearer comparison between the others scenarios, flows in scenario A were aggregated as in the scenarios where DiffServ is implemented however, they do not have a differentiated treatment in the routers. The results obtained from simulations will be analysed and examined in terms of throughput, delay, loss and jitter.

6.2.2.1 Throughput

Figures 6.7a and 6.7b illustrate that in this scenario flows are equally treated when the link is congested because flows are decreasing their throughput in equal percentages.

Figure 6.7a illustrates that before handover occurs, when a link reaches its maximum capacity near of 45 seconds, flows start to reduce their throughput because they begin loosing their packets. After handover, with MN_1 at nAR, throughput sharply decreases for

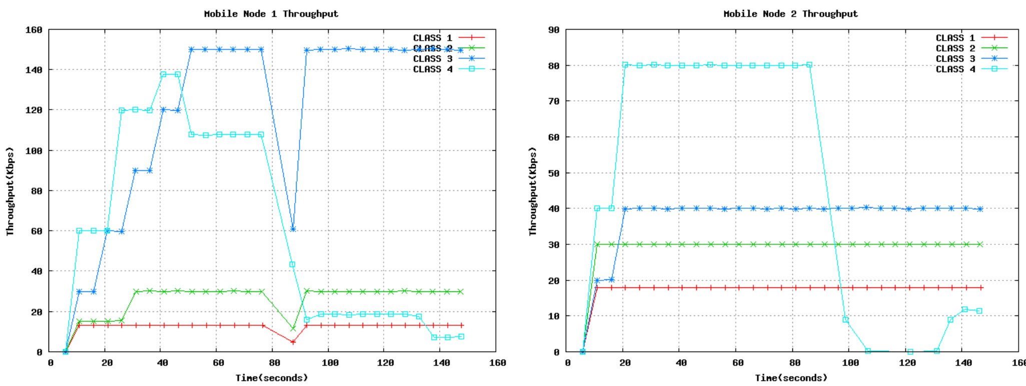


(a) Mobile Node 1 (b) Mobile Node 2

Figure 6.7: Mobile node 1 and 2 Throughput in Scenario A

all the flows since bandwidth in nAR must be shared by both MNs.

Figure 6.8 shows MN_1 and MN_2 throughput in the scenario B. The implemented scheduling mechanism in the DiffServ model was the Priority Queueing (PRI).

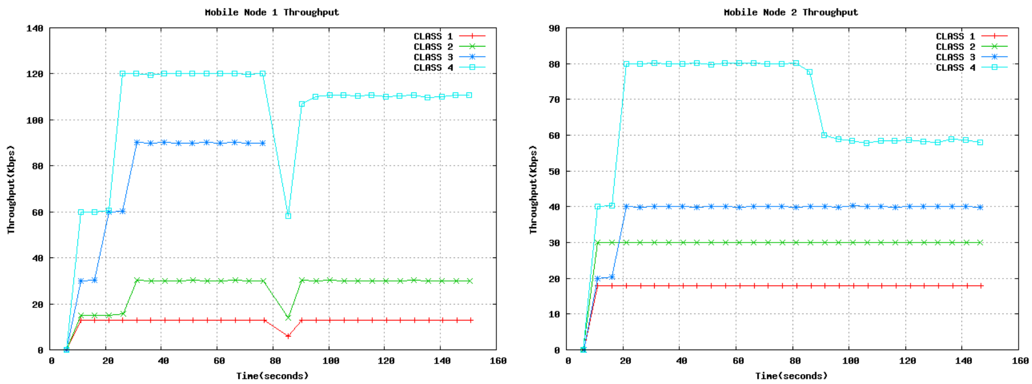


(a) Mobile Node 1 (b) Mobile Node 2

Figure 6.8: Mobile node 1 and 2 Throughput in Scenario B

Figures 6.8a and 6.8b show a very aggressive behavior with BE traffic (class 4) when the channel becomes congested causing the so called traffic starvation. On the other hand, the priority classes maintain their throughput at the expense of the BE class.

In scenario C, the admission control components were configured with the following parameters: 1) Measure Sum algorithm - 10% of BA (Bandwidth Allocation) for class 1, 20% of BA for class 2 and 30% of BA for class 3 and the remaining 40% for class 4 (BE); 2) Time Sliding Window - 3 seconds of window size (T), 0.7 seconds of sample period (S) and 0.0625 seconds of average arrival rate estimation (W).



(a) Mobile Node 1 (b) Mobile Node 2

Figure 6.9: Mobile Node 1 and 2 Throughput in Scenario C

Figure 6.9a shows that even before the handover and contrary to the precedent scenarios, when the link limit is reached, new flows are rejected in order to avoid the throughput deterioration of the existing flows. At the handover, the admission control algorithm rejects the class 3 causing the deletion of flows belonging to this class. This rejections result from the fact that the allocated resources for this class in nAR were not enough to acommodate the traffic of both mobile nodes. Being as the BE traffic is not subject to the admission control decisions, it will never be refused. Therefore, a throughput reduction in both mobile nodes occurs being as the available bandwidth for BE class on the nAR is shared between MNs.

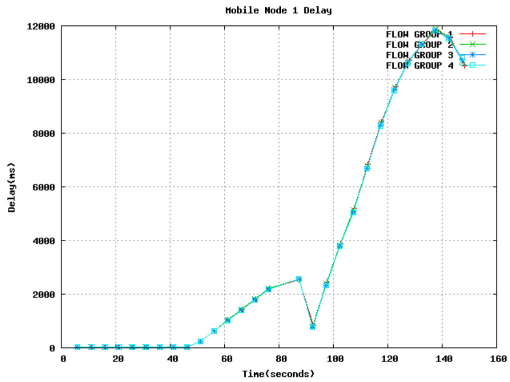
6.2.2.2 Delay

Figures 6.10a and 6.10b indicate that in scenario A, after MN_1 handover takes place, the delay dramatically increased to 12 seconds for all MN_2's flow groups. This occurs because the link becames congested.

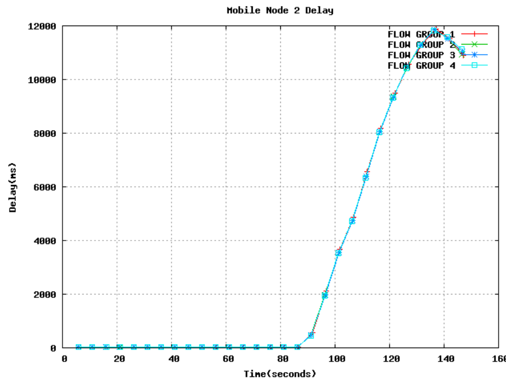
The 6.11 shows that scenario B presents a significant increase in the delay of MN_1's BE traffic (see Figure 6.11a) before handover because the link is saturated whereas the others classes only show a slightly increase in the traffic delay at this specific point. After handover, the link becames more congestedand in turn, BE traffic climbed to 2.8 seconds whereas the remaining class slightly grew.

The figures 6.11a and 6.11b illustrate that the delay of classes with higher priority were less affected.

Figure 6.12 shows that with scenario C the delay was significantly reduce when compared

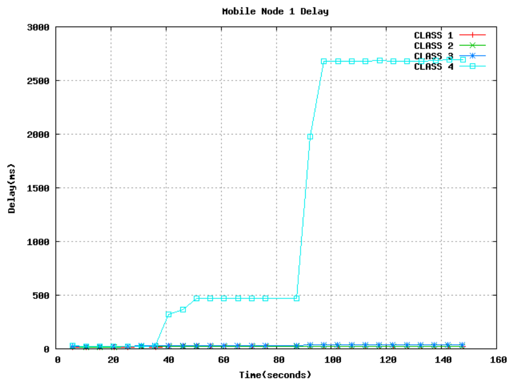


(a) Mobile Node 1

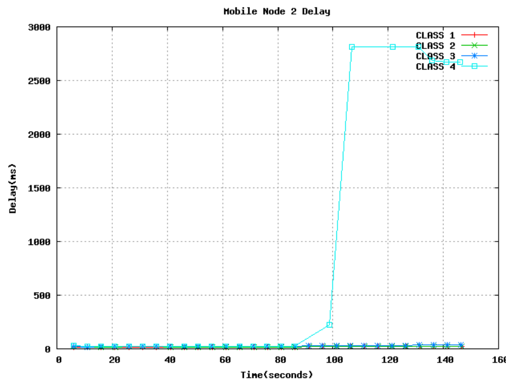


(b) Mobile Node 2

Figure 6.10: Mobile node 1 and 2 Delay in Scenario A



(a) Mobile Node 1



(b) Mobile Node 2

Figure 6.11: Mobile node 1 and 2 Delay in Scenario B

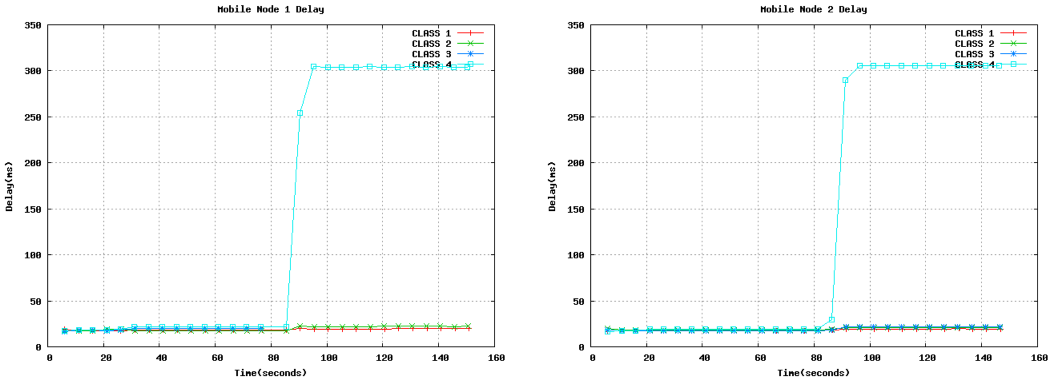
that of scenario B because the introduction of the admission control mechanism avoids link congestion and prevents that priority classes exceed their allocated bandwidth.

6.2.2.3 Loss

In what respects losses, results show a significant packet loss in scenarios A and B when the link is congested. The DiffServ mechanisms in scenario B did not reduce losses because they were not capable of avoiding link congestion. The losses seen in secnario C are substancially improved since the amount of injected traffic is bounded by the admission control algorithm.

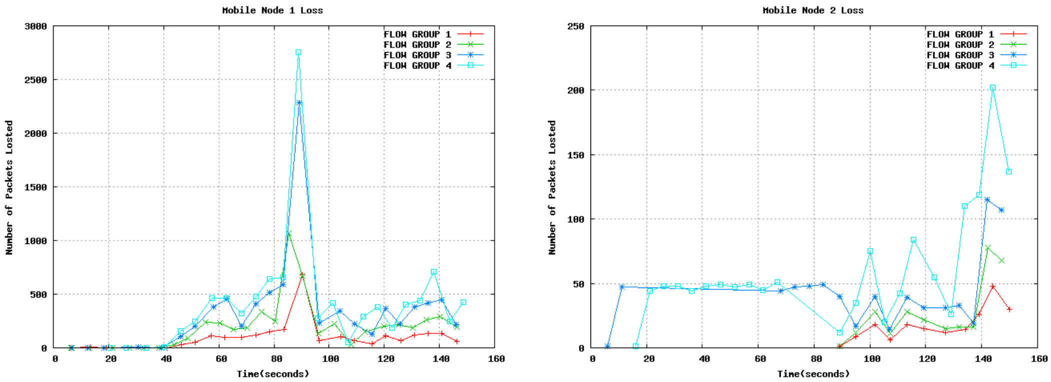
Figure 6.13 presents the mobile node losses in scenario A. Figure 6.13a shows that the losses are significant when the link is saturated. This is increasingly evident during handover. Figure 6.13b shows that around second 100, MN_2's losses begin to increase due to the arrival of MN_1 at nAR.

Figure 6.14 demonstrates that with scenario B, the level of MN_1's losses (see Figure



(a) Mobile Node 1 (b) Mobile Node 2

Figure 6.12: Mobile node 1 and 2 Delay in Scenario C



(a) Mobile Node 1 (b) Mobile Node 2

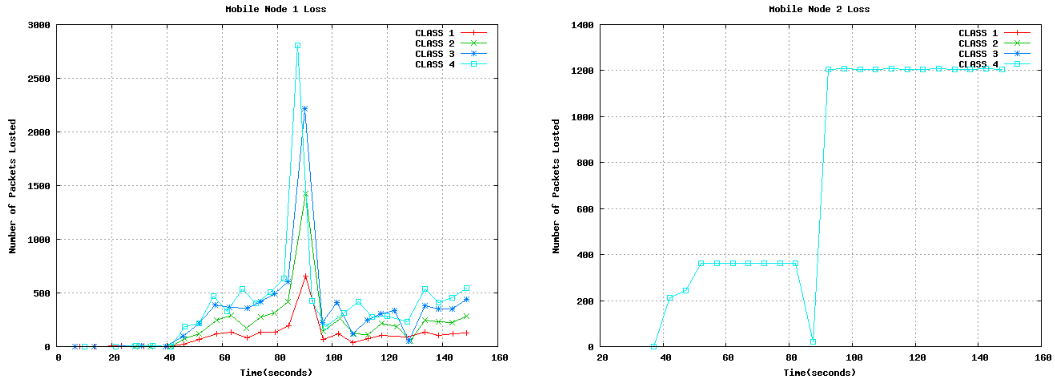
Figure 6.13: Mobile node 1 and 2 Losses in Scenario A

6.14a) did not improve with DiffServ mechanisms whereas in MN_2 (see Figure 6.14b) the priority classes did not loose any packets however, the BE class was increased to aproximately 600%.

Figure 6.15 provides the losses of MN in scenario C. Figure 6.15a shows that MN_1's losses were substancially improved with the bounding of injected traffic in the network made by the admission control mechanism. As best-effort traffic only transmits its packets after all priority packets have been transmitted, it is not subject to admission control restrictions. However, best-effort traffic also benefits from admission control mechanisms because it reduces the amount of priority traffic in network transit.

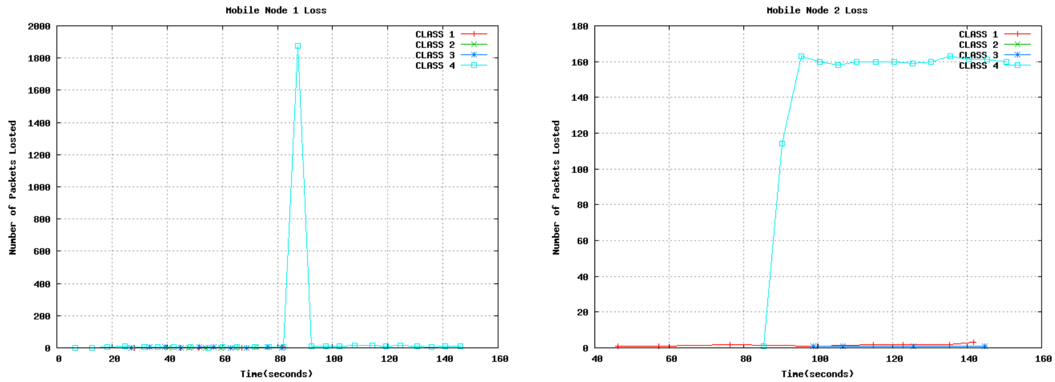
6.2.2.4 Jitter

When speaking of jitter, results showed a significant increase in jitter after the MN_1 handover due to link congestion for scenario A. In scenario B, jitter values for priority



(a) Mobile Node 1 (b) Mobile Node 2

Figure 6.14: Mobile node 1 and 2 Losses in Scenario B



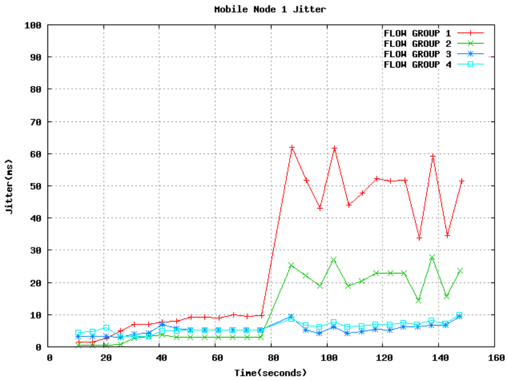
(a) Mobile Node 1 (b) Mobile Node 2

Figure 6.15: Mobile node 1 and 2 Losses in Scenario C

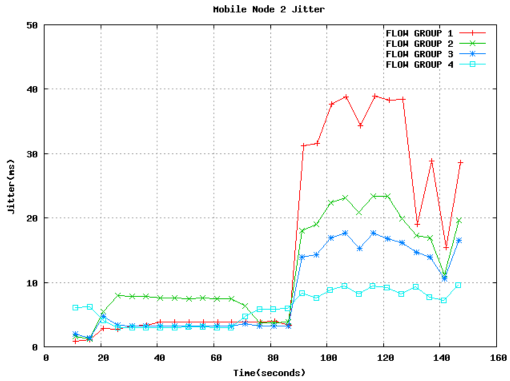
classes are maintained even when the link is congested. Although, the BE classe met what was expected by the DiffServ mechanisms, i.e. the jitter was severely aggravated. Scenario C reveals a significant jitter enhancement for MN_2 before and after handover when compared to scenarios A and B. Furthermore, the priority classes can almost maintain the same jitter before and after handover. This is due to the supervision of the admission control algorithm which limits the amount of traffic in the network.

Figure 6.16 provides the traffic jitter behaviour of the mobiles nodes in scenario A. In this scenario, results indicate a significant increase of the jitter in both MNs after the MN_1 handover.

Figure 6.17 suggests that in scenario B the DiffServ mechanisms maintain the jitter values of the priority classes even when the link is congested for both MNs. Although the BE traffic behaved according to what was expected by the DiffServ model, aggravating the jitter values

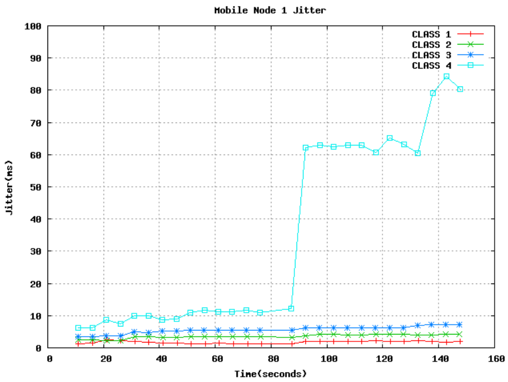


(a) Mobile Node 1

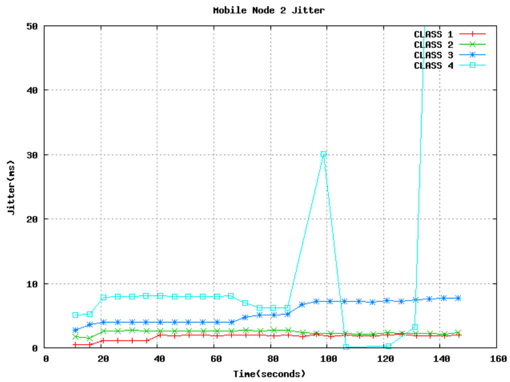


(b) Mobile Node 2

Figure 6.16: Mobile node 1 and 2 Jitter in Scenario A



(a) Mobile Node 1



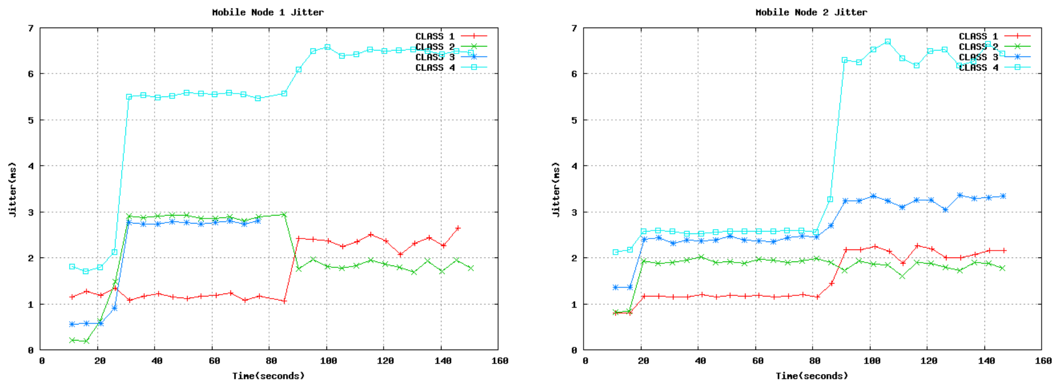
(b) Mobile Node 2

Figure 6.17: Mobile node 1 and 2 Jitter in Scenario B

Figure 6.18 reveals that in scenario C, jitter is significantly enhanced for both MNs with admission control supervision. Furthermore, the priority classes almost maintain the same jitter before and after handover.

6.3 QoS Architecture With Dynamic Allocator

In the continuity of pursuing the objectives outlined in the begining of this Chapter, the next set of simulation experiments intend to evaluate the model with the dynamic allocator. In order to assess the dynamic allocator´s influence on the system, a new scenario called scenario D has been added to the previous scenarios (A, B and C). This scenario contains the DiffServ mechanisms, admission control policy and dynamic allocator. Scenarios A, B, C and D were simulated one hundred times and then submitted to statistical treatment using R software [138].



(a) Mobile Node 1 (b) Mobile Node 2

Figure 6.18: Mobile node 1 and 2 Jitter in Scenario C

6.3.1 Simulation Scenario for Dynamic Allocator

The simulation scenario of the previous section has been enriched in order to produce realistic, generic, and richer results. Figure 6.19 shows the simulated topology for an intra-domain scenario.

The simulation scenario consists in a part with the global internet where ten CNs and HA's are located and another part with a F-HMIPv6 aware DiffServ domain with two access routers and ten MNs.

The MNs receive Constant Bit Rate (CBR) flows from CNs located at another DiffServ domain of the global internet. This traffic transmission is one to one (CN→MN). Each CN generatesfour CBR flows and each one is marked with a different DiffServ Code Point (DSCP). Therefore, forty flows have been generated in total (see Table 6.1).

MN	*Class 1 (kbps)*	*Class 2 (kbps)*	*Class 3 (kbps)*	*Class 4 (kbps)*
1	10	6	10	12
2	10	6	6	8
3	10	11	6	8
4	10	7	6	8
5	10	10	6	8
6	10	4	6	8
7	10	12	6	8
8	10	12	6	8
9	10	20	30	40
10	10	20	30	40

Table 6.1: CBR Flows received by MNs

For real-time applications, the user expects a flow of audio or video that is presented in

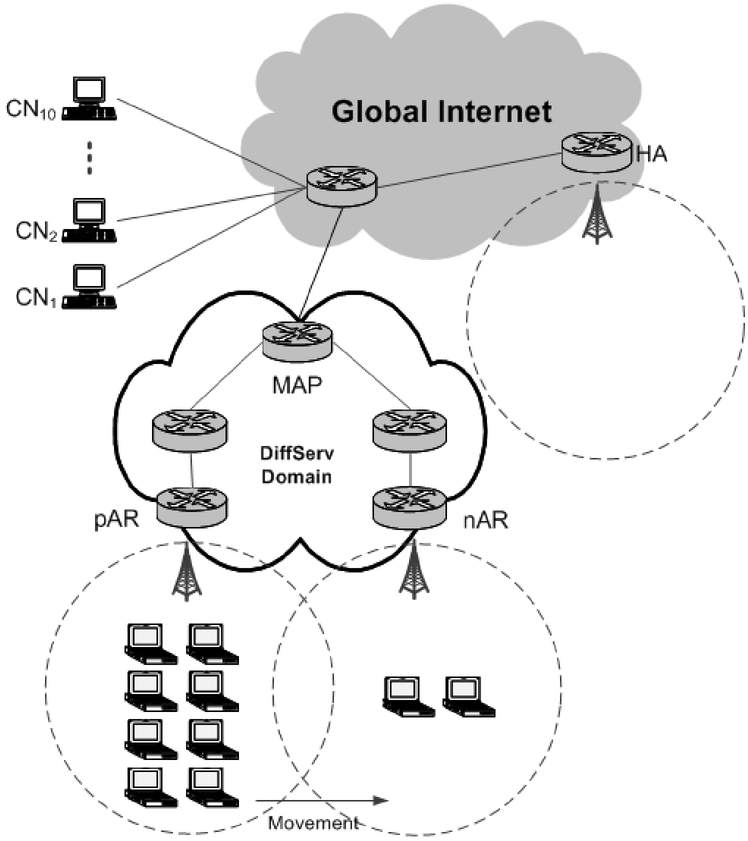


Figure 6.19: Simulation Model

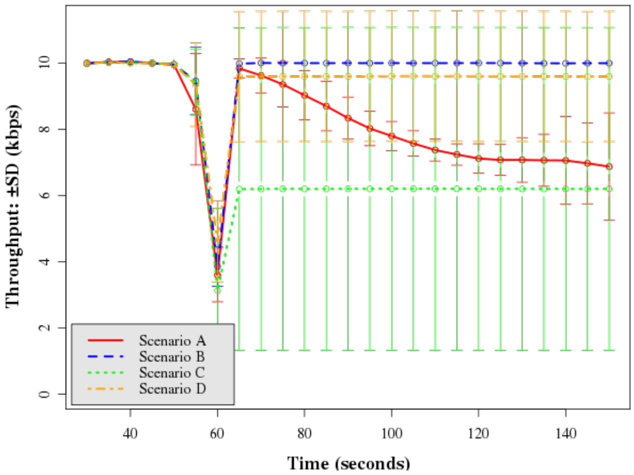
a continuous and smooth fashion. The CBR service is one commonly used for audio and video infomation because these kind of applications require a fixed data rate during the connection lifetime. These were reasons which explained why CBR flows were choosen to model the real-time applications' traffic in the priority classes.

Being as the bottleneck is in the last hop, (wireless link) all flows will be accepted by the ingress. Eight MNs are initially located in pAR and two MNs are fixed in nAR. One of the MNs in pAR is moving at fixed time (60 seconds) while the others are moving randomly in a range time between 50 and 100 seconds to nAR. The network load on nAR after MNs handovers is 131.6%.

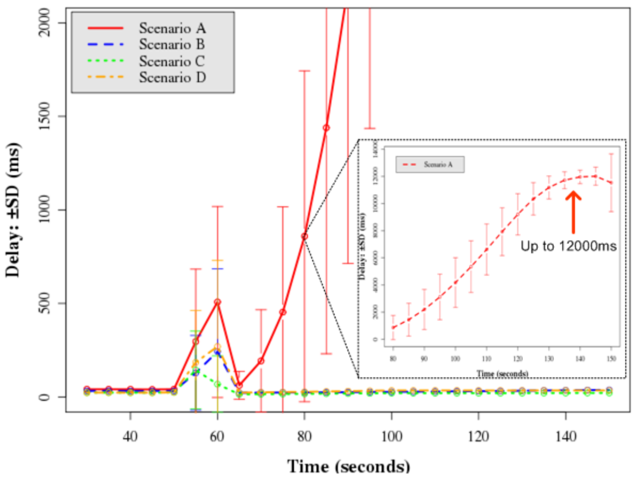
6.3.2 Result Analysis on Dynamic Allocator

Figures 6.20 and 6.21 illustrate the class 1 and 3 mean throughput as well as delay distributions with its respective standard deviation around the mean.

Figure 6.20 shows that after MN's handover, which started at 60 seconds, scenario B has achieved the best mean throughput. An explanation for this is that the standard DiffServ



(a) Class 1 Mean Throughput and Standard Deviation

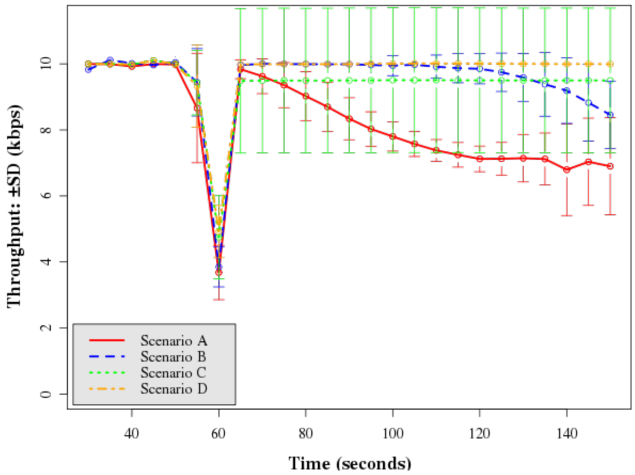


(b) Class 1 Mean Delay and Standard Deviation

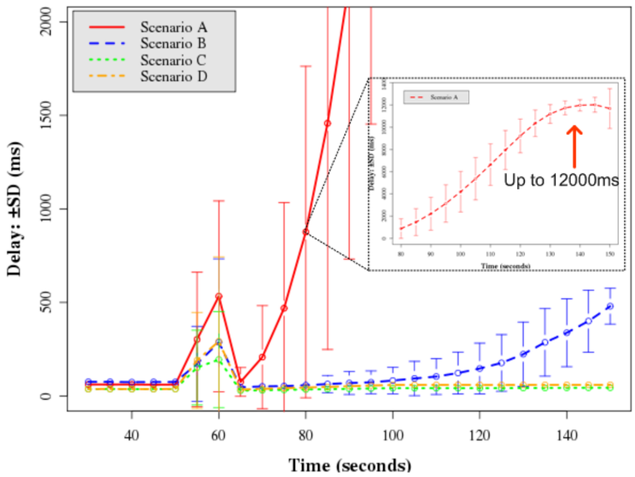
Figure 6.20: Class 1 Throughput and Delay with Standard Variation in the Four Scenarios

mechanisms did not have any threshold limit in the class which results in the admission of all generated traffic. Scenario C presents a mean throughput which is almost half of the initial mean throughput (before handover) and high standard deviation due to an admission control scheme that limits the amount of traffic in class 1 rejecting surplus traffic. Scenario D presents a slight decrease in the initial mean throughput and a standard deviation similar to scenario B. This is due to the dynamic allocator that reallocates more bandwidth for class 1 to accommodate more traffic in this class thus resulting in a small traffic rejection. Scenario A presents a gradual mean throughput decrease which is proportional to the link saturation since traffic is treated equally in the four classes.

In what pertains to delay, scenario A has sharply increased the mean delay as well as its variability with the link saturation and scenarios B, C and D present a similar mean delay.



(a) Class 3 Mean Throughput and Standard Deviation



(b) Class 3 Mean Delay and Standard Deviation

Figure 6.21: Class 3 Throughput and Delay with Standard Variation in the Four Scenarios

Figure 6.21 shows that scenarios B and D, after MN's handover in the initial period, have approximately the same mean throughput however, scenario D shows that the mean throughput remains constant. In scenario B, the mean throughput starts to decrease at 100 seconds which is when all MNs have proceeded their handovers. This occurs because class 3 is the DiffServ class with less priority and when the link starts to become saturated, the class with less priority is the first to be affected. Scenario C presents a mean throughput decrease which is due to the AC scheme that rejects some flows during the handover, such as in class 1 mean throughput. Scenario A presents a mean throughput distribution for class 3 similar to class 1.

When speaking of delay, the figure shows that scenarios C and D maintain the same delay while scenario B increases delay with the arrival of the eight MNs. In scenario A, one can observe that classes 3 and 1 (see Figure 6.20) have an identical mean delay distribution curve and standard variation on account that both classes have the same forwarding treatment on routers.

6.3.3 Traffic Behavior during MNs Handover

This section analyses the behavior of different traffic classes during the MNs handovers. For this purpose, one of the MNs moving to nAR has been selected. Therefore, the graphics presented in this section were generated from classes of traffic from this single MN.

6.3.3.1 Scenario A

In order to facilitate the analysis, traffic flows in this scenario have been aggregated in the same manner as in the DiffServ configurations even though they do not have any differentiated treatment in a congestion scenario

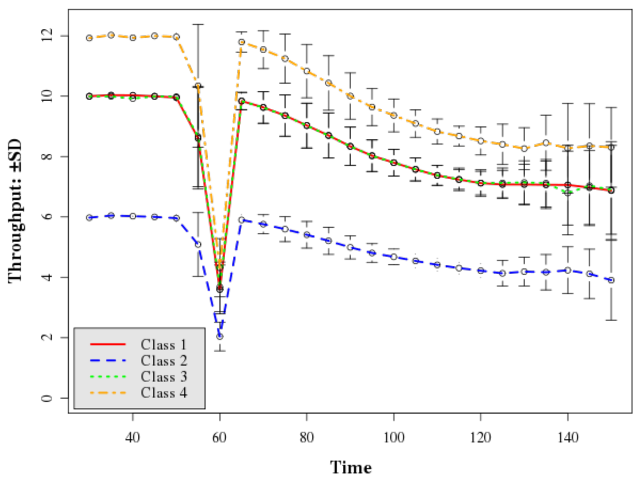


Figure 6.22: Classes Mean Throughput(kbps) with Standard Deviation for MN in Scenario A

Figure 6.22 illustrates the mean throughput of the four traffic classes and their standard deviation in scenario A. Results suggest that when an MN makes a handover to nAR, classes are equally treated, merely gradually decreasing their throughput when the link becomes more congested. The standard deviation is enhanced when the MN starts the handover. After the handover, a slight variation in the transmission rate is presented.

Figure 6.23 shows the cumulative function of the class 1 throughput for scenario A as well as its correspondence to the normal distribution when set with the same mean and standard

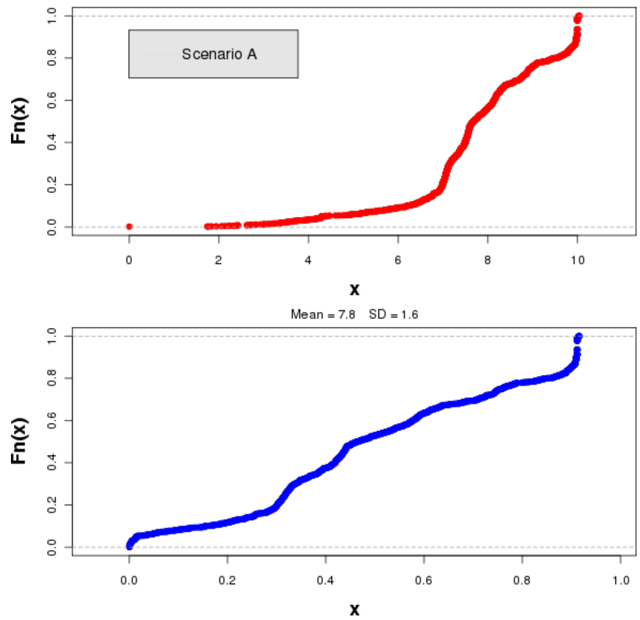


Figure 6.23: Class 1 Throughput(kbps) Cumulative Function for MN in Scenario A and its Respective Normal Distribution

variation of the class 1 throughput. As can be seen in this scenario, throughput varies from 4kbps to 10kbps and only about 10% of flows can obtain the required 10kbps of transmission rate. This is easily explained because all traffic classes received the same treatment thus, when the link becomes congested, traffic competes for bandwidth without any kind of privilege for traffic belonging to applications that require same QoS requirements.

6.3.3.2 Scenario B

Figure 6.24 shows the cumulative function of class 1 throughput for scenario B and its correspondence to the normal distribution.

DiffServ has been configured with a Priority Queueing (PRI) scheduling mechanism in order to decide which packet should be transmitted when the link experiences congestion [12].

The figure illustrates that in this scenario, approximately 90% of flows in class 1 had the required 10kbps of transmission rate. However, it should be noted that class 1 is the highest priority DiffServ class and does not have any traffic limit within its class therefore, it could affect or even cause starvation in lower priority classes. These starvation effects in lower priority classes may occur if the resource management does not control the excess of traffic within a class.

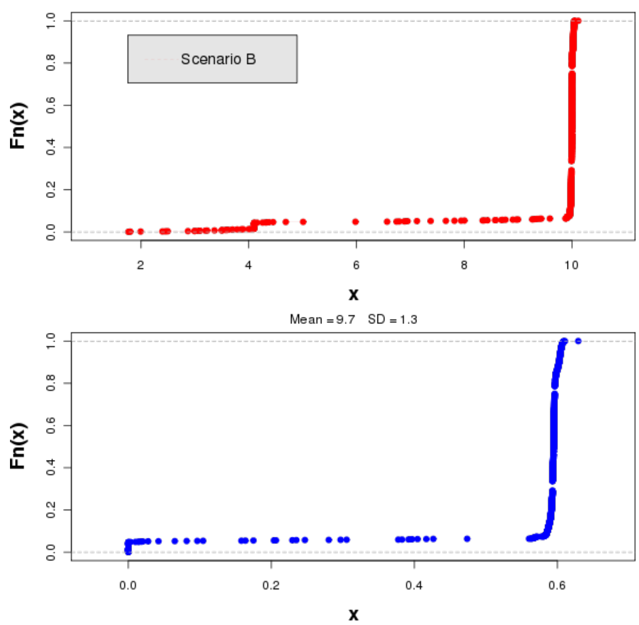


Figure 6.24: Class 1 Throughput(kbps) Cumulative Function for MN in Scenario B and its Respective Normal Distribution

In Figure 6.25, when handovers occur, the link becomes saturated consequently, class 4 starts to decrease its throughput until it enters starvation around second 145 and class 3also starts to decrease its throughput. After the MNs handovers, the standard deviation also increases in two lower classes (3 and 4).

6.3.3.3 Scenario C

Figure 6.26 provides the mean throughput results of MN's classes for the scenario C. The results indicate that all priority classes, with the exception of class 4, because this class is for Best Effort traffic and therefore it is not subject to admission control decisions, had rejected flows thus, the mean throughput of classes were substantially reduced, this is specially true for class 1 whose mean decrease to 64% when compared to the initial throughput.

Figure 6.27 shows the cumulative function of the class 1 mean throughput for scenario C and its correspondence to the normal distribution when set with the same mean and standard deviation throughput values of class 1.

The AC components i.e., its measure sum algorithm and TSWE have been configured with the following parameters: 10% BA (Bandwidth Allocation) for Class 1, 20% BA for Class 2 and 30% BA for Class 3 and the remaining 40% for Class 4 (Best Effort); 3 seconds

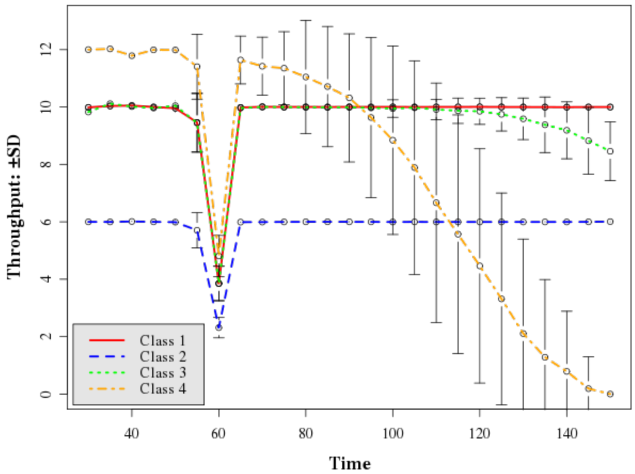


Figure 6.25: Classes Mean Throughput(kbps) with Standard Deviation for Scenario B

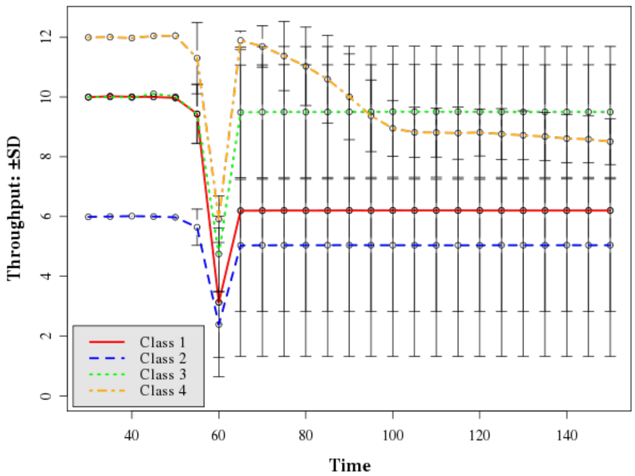


Figure 6.26: Classes Mean Throughput(kbps) with Standard Deviation for Scenario C

for window size (T), 0.7 seconds for sample period (S) and 0.0625 seconds for the average arrival rate estimation (W). The figure illustrates that almost 40% of flows were rejected although almost all accepted flows have obtained a transmission rate of 10kbps. MNs have achieved the required QoS for class 1 because the amount of traffic in this class has been limited by the AC algorithm. Being as a significant percentage of moving flows have been rejected to preserve the QoS level in class 1 without taking into account if there is available bandwidth in other classes, the resource utilization is not optimized therefore, resources were wasted.

6.3.3.4 Scenario D

In Figure 6.28 results show that in scenario D there has been a significant improvement on the amount of accepted flows belonging to priority classes. In the priority classes 1, 2

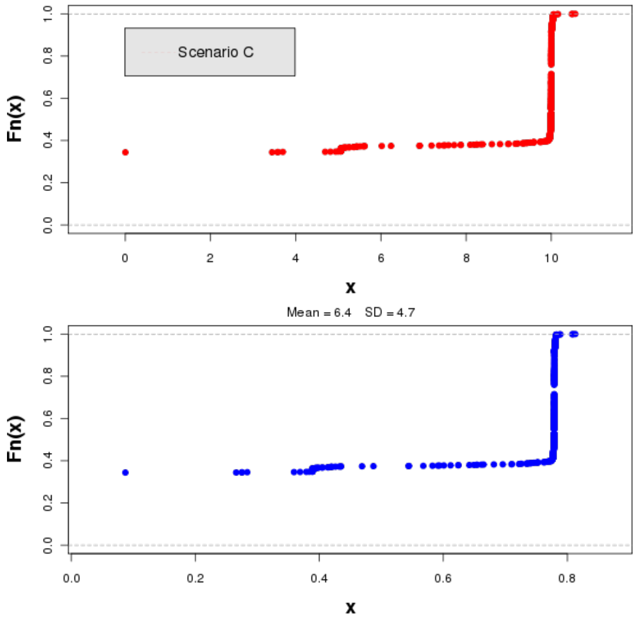


Figure 6.27: Class 1 Throughput(kbps) Cumulative Function for MN in Scenario C and its Respective Normal Distribution

and 3, the mean throughput after MNs handovers was sharply increased when compared to scenario C. Thus, as a consequence of the growth of accepted flows, the link utilization in the AR also increased.

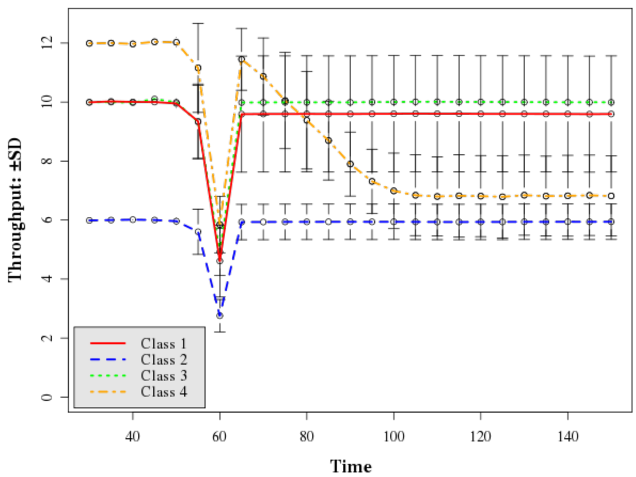


Figure 6.28: Classes Mean Throughput(kbps) with Standard Deviation for Scenario D

Figure 6.29 shows the cumulative function of the class 1 mean throughput for scenario D and its normal distribution correspondence set with the same mean and standard deviation values.

It is interesting to observe from the Figure 6.29 a significant improvement of rejections that dramatically decreased to approximately 2%. Furthermore, almost all of the accepted

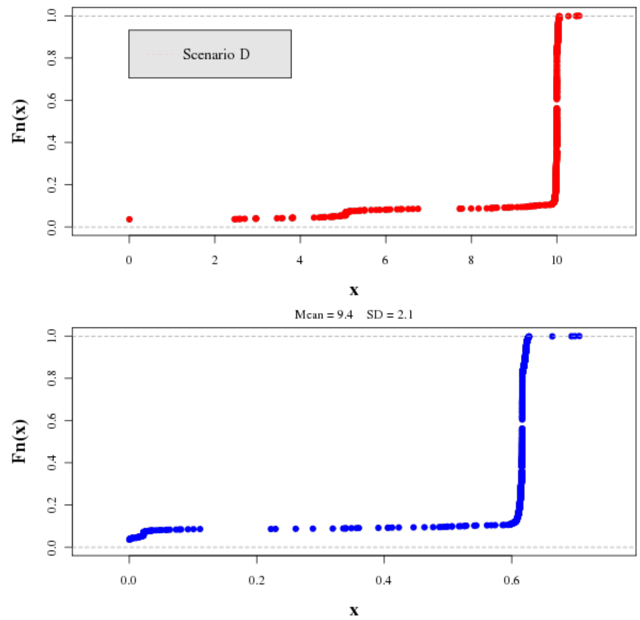


Figure 6.29: Class 1 Throughput(kbps) Cumulative Function for MN in Scenario D and its Respective Normal Distribution

flows obtained a transmission rate of 10kbps which results in a normal distribution for class 1 with a mean throughput of 9.4kbps and a standard deviation of 2.4kbps. This improvement is due to the dynamic allocator that makes a more efficient use of resource utilization by making re-allocations when a given priority class requires more bandwidth.

As class 1 throughput that has been collected from MNs follows a normal distribution, it is possible to apply parametric statistical tests. The statistic test that is most commonly used for assessing if there is a significant difference between two sample means is Student's test. In this work, a two sample location test of the null hypothesis was used to compare the mean throughput of scenarios C and D. The t student p-value for a confidence interval of 95% was 2.2e-16 what means that the null hypothesis is false, i.e., as expected the mean throughput of class 1 of scenarios C and D are not equal.

The throughput quartiles for both scenarios were also computed. The scenario C has showed a lower quartile of 0.00, a median of 9.94 and upper quartile of 10.00. Whereas scenario D showed a lower quartile of 9.98, a median of 9.99 and upper quartile of 10.00. The quartiles reveal that in scenario C at least 25% of flows had 0.00kbps of throughput, what means that were rejected (see Fig. 6.30).

By analyzing the results one can conclude that scenario D has achieved approximately the same mean throughput of scenario C, but having much more flows accepted.

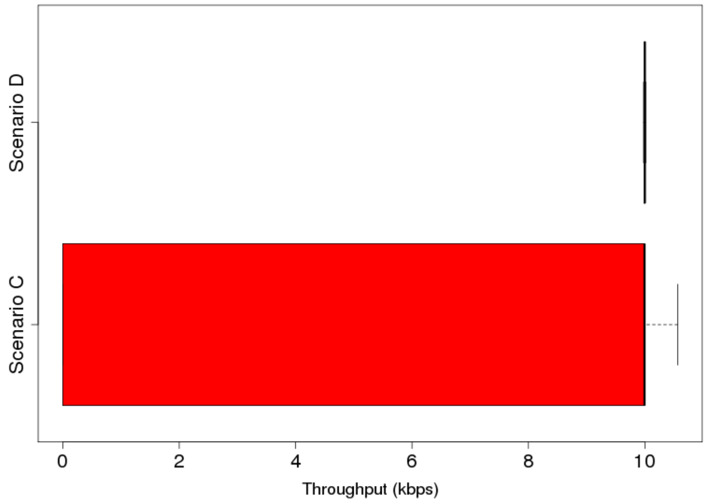


Figure 6.30: Bloxplot of MN Throughput(kbps) for Scenarios C and D

6.3.4 Impact of Incoming MNs on the Existing Traffic

This section analyses the MN's traffic behavior in nAR in order to evaluate how the existing traffic is affected by the incoming MN's traffic. For this purpose one of the MNs in the scope of nAR has been selected to collect data. Therefore, the results shown in this section refer this MN. As stated before, MNs start to move towards nAR randomly between second 50 and 100 therefore, nAR will become increasingly congested with the arrival of eight MNs.

6.3.4.1 Scenario A

In this scenario, results indicate that the existing traffic was severely and equally reduced with the incoming MNs. The class 1 throughput was affected by the incoming MNs traffic causing a significant decrease of 25% in its throughput. With regards to the standard deviation in class 1 throughput, there is a variation of 5% around the mean throughput caused by the arrival of MNs at nAR which begin to congest the link.

In terms of delay, the results show that after the MNs arrived at nAR the delay in class 1 of the existent traffic dramatically increased to 12 seconds and had high standard deviation values (1000ms) with the link congestion.

6.3.4.2 Scenario B

This scenario turned out as expected. Results show the typical behavior of DiffServ with PRI scheduling, where the high priority classes, class 1 and 2, are primarily served and therefore, maintain their throughput. However, class 3, a lower priority class, begins to decrease its throughput and class 4 with BE traffic enters when the link becames congested. The results also show a slight decrease of 0.2% in the class 1 average throughput as well asa slight standard deviation of 0.3% with link congestion.

The class 1 delay in this scenario increased gradually to approximately 57ms with link congestion while the standard deviation maintained itself regular and at reasonable values (5ms) of variation.

6.3.4.3 Scenario C

Results reveal that priority classes were not affected with the incoming MNs, only class 4 (Best Effort traffic) has been affected as expected since it has not been submitted to admission control. Class 1 throughput shows an insignificant decrease of 0.01% and a variation of 0.02% when the link is more congested.

When speaking of the delay in class 1, it slightly increases (+10ms) when the link is more congested while it remaining unchanged in the other cases.

Therefore, one can conclude that the effect of incoming traffic on the existent priority in this scenario could be considered perfectly negligible.

6.3.4.4 Scenario D

In this scenario, the MN's priority traffic was not affected by incoming MNs as seen in scenario C. The most obvious difference in scenario C is that class 4 had a higher mean throughput reduction (Scenario C: 25% and scenario D: 50%), as well as an increase in standard deviation (Scenario C: 5% and scenario D: 10%). This is a consequence of the fact that accepted flows have increased in priority classes which leads to a traffic increment within priority classes and thereby an elevated reduction in class 4 throughput due to its reallocation to other priority classes.

The results of this scenario for class 1 throughput are very similar to those of scenario C (see Figure 6.31), but with a slight increase in standard deviation (0.04%) in scenario C when the link is congested. In scenario D, even with more accepted flows in priority classes, the flows belonging to priority classes can achieve a similar mean throughput when the link is congested. The average delay in this scenario also slightly increases (+20ms) with the incoming MNs.

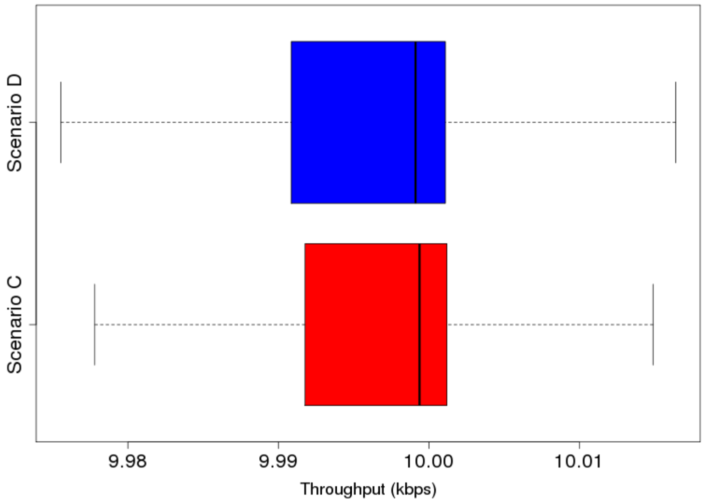


Figure 6.31: Box Plot of MN Throughput(kbps) for Scenario C and D

The t student p-value for a confidence interval of 95% was 0.1491 which means that the null hypothesis is true, i.e., mean throughput for scenarios C and D is equal . The calculated throughput quartiles for both scenarios were also almost equal which reveals that even thought there are more flows accepted in the priority classes, the existent priority traffic is not significantly affected by the incoming traffic.

In conclusion, the impact of incoming traffic on the existent priority traffic, in terms of throughput and delay, can be considered meaningless in this scenario.

6.4 Traffic Models and Queueing Disciplines Evaluation.

In this subsection the architecture behavior when it is subjected to both CBR and exponential (EXP) traffic will be analyzed in addition to its behavior with both Priority queue (PRI) and Weighted Round Robin (WRR) scheduling algorithms which are the most common queueing disciplines used in the DiffServ architecture.

Before starting with the analysis of the results, it is important to highlight the essential characteristics of the PRI and WRR queueing disciplines, so that the results can be better understood.

In packet-switched networks, the packets belonging to various flows are queued for transmission at an output buffer. The manner in which they are selected for transmission on the link is known as queueing scheduling discipline. There exist two common queueing disciplines for the DiffServ architecture which are PRI and WRR. In the PRI queueing discipline the packets arriving at a router are classified and forwarded for its output priority queue. The priority queueing discipline always chooses to transmit first the packet that belongs to the highest priority class.

In the WRR queueing discipline, the scheduler alternates the service among the classes in a circular manner serving the highest priority class first. Each class may have a different amount of service which is assigned by means of a weight. This kind of discipline is called "work-conserving" because it never allows the link to be idle if there are packets queued, to be transmitted, in any class.

This work mainly deals with the class of applications which allows people to use audio/video to communicate with each other in real time. In this class, delays smaller than 150 milliseconds are unperceivable by a human listener and delays between 150 and 400 milliseconds could be acceptable.

In order to transmit voice over Internet, the analog audio signal must be converted or encoded into a digital signal and then compressed to reduce the bit rate of the stream. There are several compression schemes. The most common are GSM (13 kbps), G.729 (8 kbps) and G.723.3 (both 6.4 and 5.3 kbps). In order to simulate traffic voice, the flows generated for class 1 have been modeled with CBR and EXP traffic transmitting at a rate of 5.3 kbps (for EXP traffic when it is in the period ON), which represent voice traffic encoded with G.723.3. The exponential ON-OFF-traffic model has been set up with the periods ON and OFF of 650ms and 350ms, respectively [139].

The simulation scenario is the same that has been used for assessing the dynamic allocator called scenario D which has been described in Sub-section 6.3.1. The weights assigned to the WRR scheduling discipline were 50% for class 1, 35% for class 2, 10% for class 3 and 5% for class 4.

Keeping in mind the simulation scenario used in the previous section, where eight mobile nodes are moving from pAR to nAR randomly in the time range between 50 and 100 seconds, and the other two are always located in the scope of nAR, the tests have been conducted in order to evaluate the voice traffic models and the queueing disciplines in the traffic

PAR (85%)

Class 1	Class 2	Class 3	Class 4	
12	36	30	24	%
42	126	105	84	kbps
5.25	15.75	13.3	10.5	(kbps)×8 MN

nAR (15%)

Class 1	Class 2	Class 3	Class 4	
6	12	18	18	%
21	42	63	63	kbps
10+11	20+22	30+33	30+33	(kbps)$(1+1)$ MN

(a) Generated Traffic in the pAR and nAR

By applying the AD Policies to new flows one gets:

pAR

Class 1	Class 2	Class 3	Class 4	
10	20	30	24	%
31.5	63	93.1	84	kbps

nAR

Class 1	Class 2	Class 3	Class 4	
6	12	18	18	%
21	42	63	63	kbps

(b) Envisioned Traffic Load After Applying the Admission Control to New Flows

After the handover of $8 \times$ MN in pAR to the nAR one gets:

pAR

Class 1	Class 2	Class 3	Class 4	
9	18	26.6	24	%
31.5	63	93.1	84	kbps

nAR

Class 1	Class 2	Class 3	Class 4	
15	30	44.6	42	%
131.6% of overload				

(c) Envisioned Traffic Overload at nAR After MNs Making the Handover

Table 6.2: Generated Traffic, Admission Control and Traffic Load

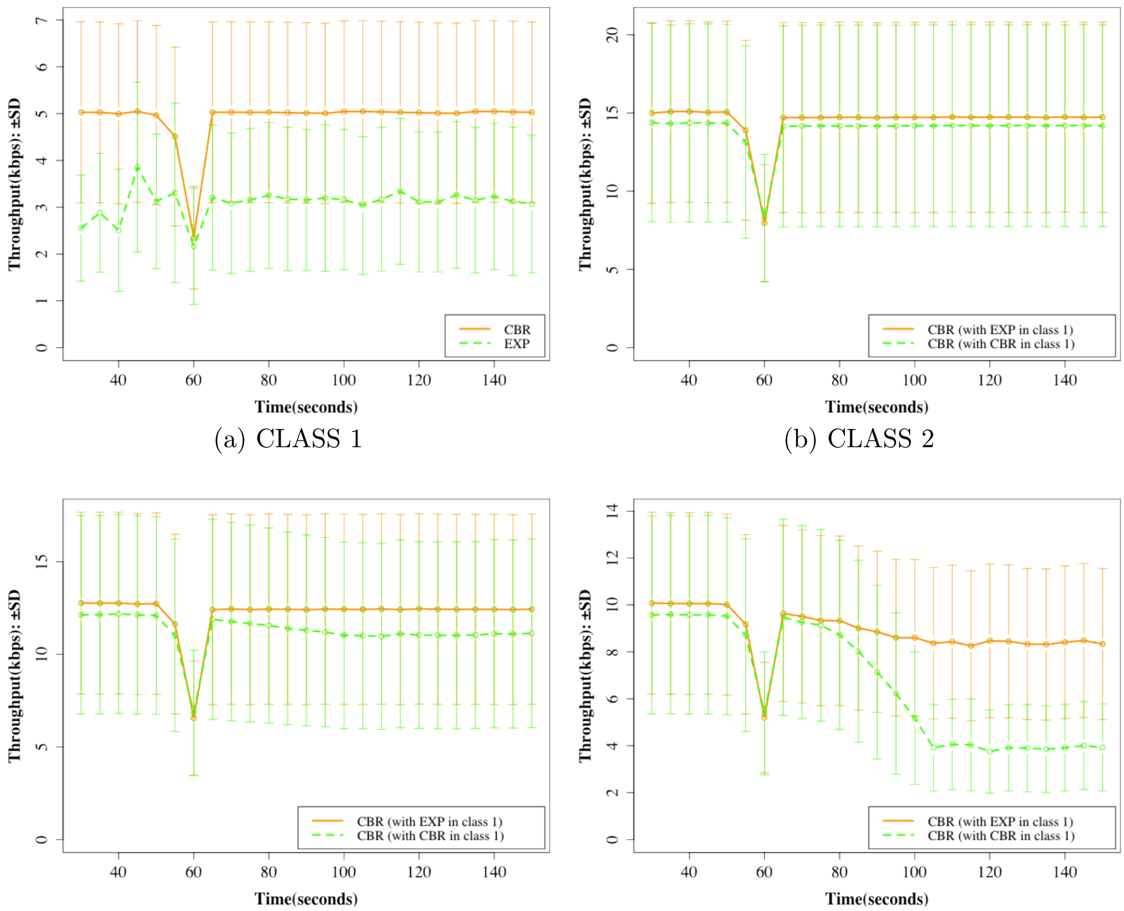
belonging to a mobile node, in movement to nAR and in a stationary mobile node at nAR. The mobile nodes have been denominated as MN_1 and MN_10, respectively.

This scenario has been configured in four different manners: 1) with CBR traffic and PRI queueing discipline; 2) with CBR traffic and WRR queueing discipline; 3) with EXP traffic in class 1 and WRR queueing discipline; 4) and with EXP traffic in class 1 and PRI queueing discipline. The traffic generated in the four classes of each mobile node are described in Table 6.2.

Figures 6.32 and 6.33 show the traffic behavior in each class, respectively for MN_1 and MN_10, when the scenario is configured with the CBR traffic and PRI queueing discipline and with the EXP traffic and PRI queueing discipline, respectively.

As expected, figures show that the influence of EXP traffic on the transmission rate of the other flows is insignificant and shows only small differences which eventually should be due to the EXP bursty nature which implies additional retransmissions at the MAC level.

Nevertheless, the MN_10' traffic (Fig. 6.33) can maintain its transmission rate in the two higher priority classes even in the presence of EXP traffic, showing only a slight decrease



(a) CLASS 1
(b) CLASS 2
(c) CLASS 3
(d) CLASS 4

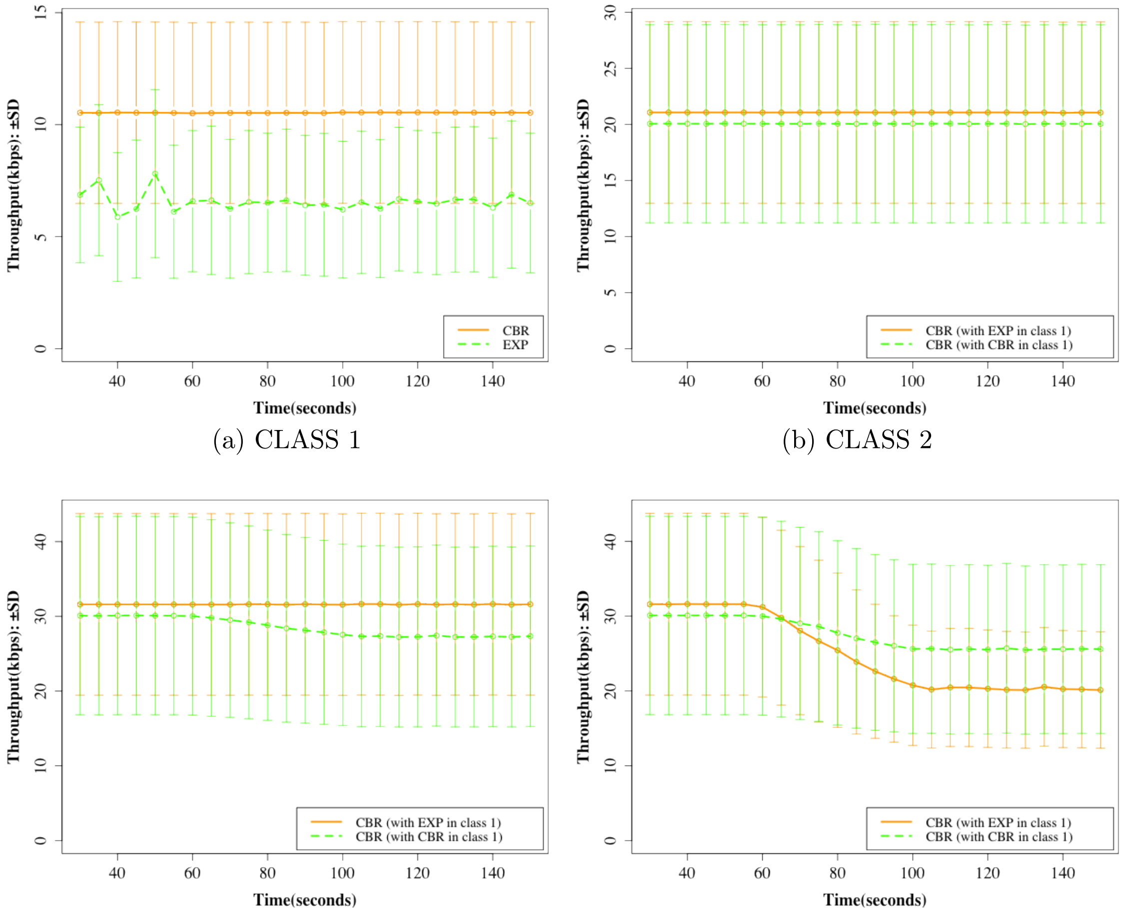
Figure 6.32: MN_1 traffic with PRI scheduling

in the lowest priority class.

Figures 6.34 and 6.35 show the traffic behavior in each class for MN_1 and MN_10 when the scenario is configured with CBR traffic and WRR queueing discipline, and with EXP traffic and WRR queueing discipline, respectively.

The Figures show that with the WRR queueing discipline, the mean rate oscillates a little more for both types of traffic however, with CBR traffic in class 1, the mean rate has slightly reduced in the higher priority classes while with the EXP traffic in class 1, the mean rate in all classes is similar that obtained in the Figs 6.32 and 6.33.

As the WRR scheduler alternates the service among the classes in a circular manner, providing each class a certain amount of service, the CBR traffic can only transmit during its assigned amount of service. After this has occurred, it must wait for its turn to transmit again. This causes a slight reduction in the mean rate, mainly for lower priority classes because they have a smaller weight.

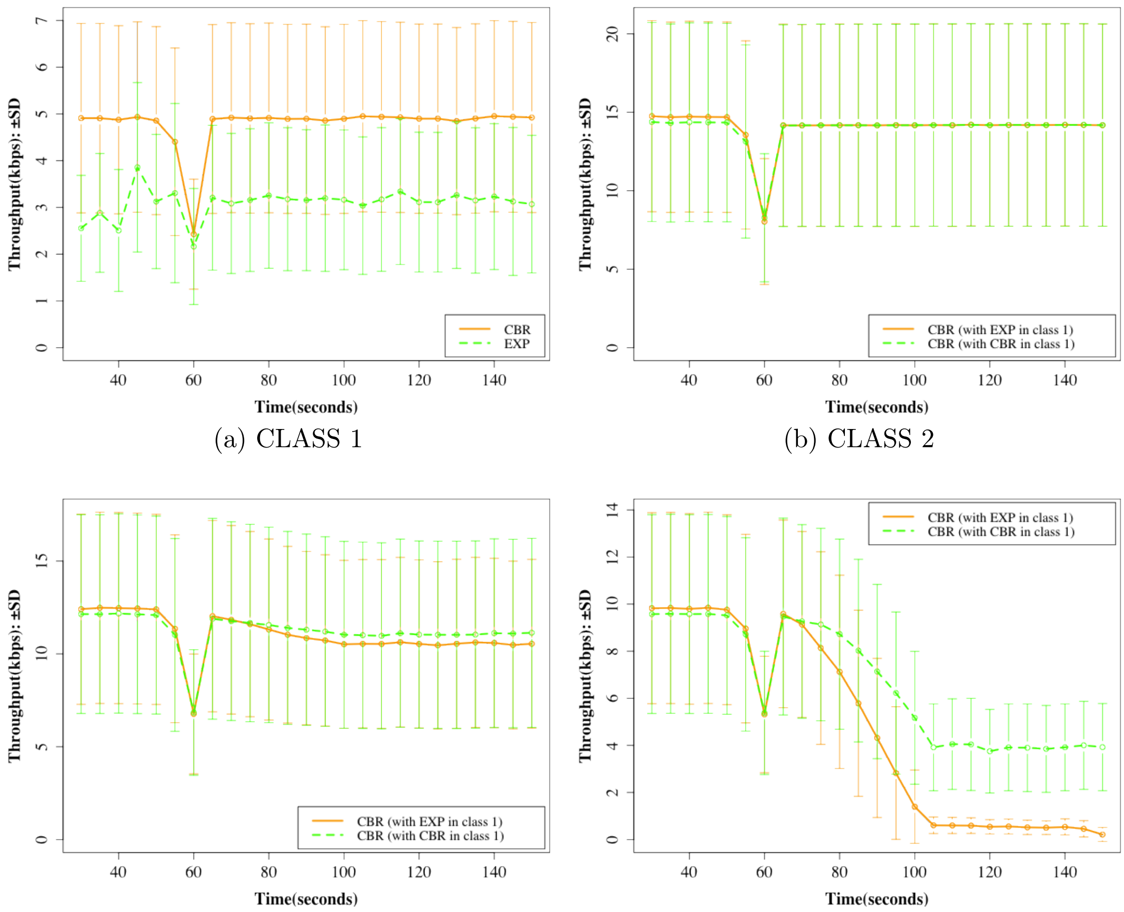


(a) CLASS 1 (b) CLASS 2

(c) CLASS 3 (d) CLASS 4

Figure 6.33: MN_10 traffic with PRI scheduling

In conclusion, the results are very similar for both types of traffic, as expected because the QoS mechanisms of the proposed model protect traffic belonging to the priority classes regardless of the traffic type. The small differences between CBR and EXP traffic can eventually derive from the fact of the 802.11 error control, the ARQ error control mechanism, makes additional retransmissions with EXP traffic. The PRI queueing discipline has shown to be more efficient for CBR traffic belonging to a priority class because on the contrary to the WRR queueing discipline, the packet is served according to is priority and do not need to wait for its queue turn to be served when a packet arrives at a priority queue.



(a) CLASS 1 (b) CLASS 2

(c) CLASS 3 (d) CLASS 4

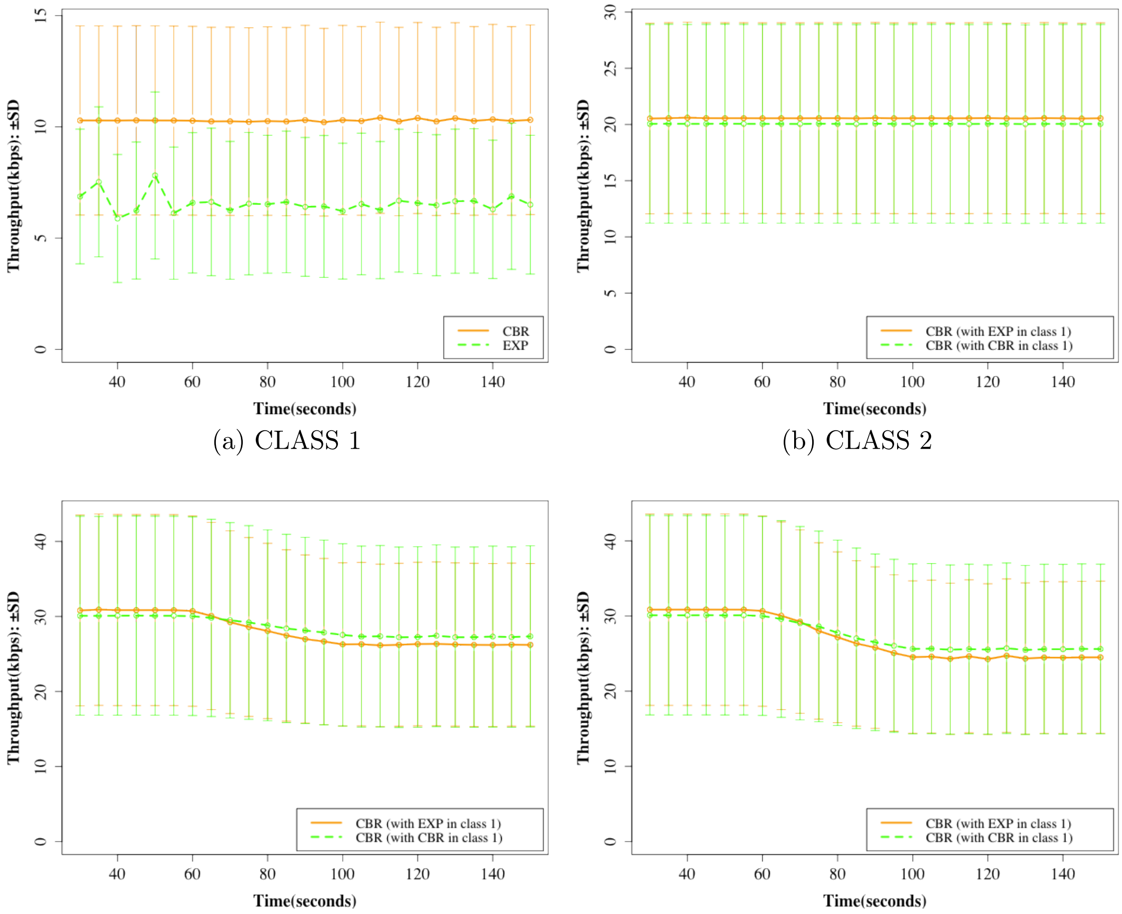
Figure 6.34: MN_1 traffic with WRR scheduling

6.5 Model Parametrization

The model parametrization is made by establishing the following parameters: 1) $ClassBW_i$: the bandwidth initially allocated for class i; 2) $\triangle max_i$: the maximum bandwidth variation of class i; 3) $\triangle min_i$: the step unit size.

The first two parameter values should be chosen by a network administrator based on the Internet Service Provider (ISP) policies and the knowledge of his network traffic. Here, the most appropriate values for his domain are assigned. The last parameter ($\triangle min$) determines the number of steps needed to achieve the $\triangle max$. The $\triangle min$ value infers in the QoS provided by the dynamic allocator and in the network stability being as the frequent reallocations in a class can cause instability..

Considering T_{BW} the total wireless link bandwidth, the first parameter $ClassBW_i$ which



(a) CLASS 1 (b) CLASS 2

(c) CLASS 3 (d) CLASS 4

Figure 6.35: MN_10 traffic with WRR scheduling

is the allocated bandwidth for each DiffServ class, has been set up with: 10% for class 1, 20% for class 2, 30% for class 3 and 40% for class 4.

The second parameter which is the maximum bandwidth variation of the class has been established with: 50% for class 1, 40% for class 2 and 30% for class 3, the sum of these variations corresponds to 22% ($0.1T_{BW} \times 50\% + 0.2T_{BW} \times 30\% + 0.3T_{BW} \times 20\% = 0.22T_{BW}$) which is the maximum negative variation of class 4 (the class with BE traffic). Figure6.36 shows a representation of the defined parameters.

The $\triangle min$ value determines the number of steps necessary to achieve the $\triangle max$ (see table 6.3), to have a more consistent network the number of steps within each class should be the lowest possible.

In order to evaluate the $\triangle min$ parameter's influence on the model architecture, the network stability, maximum bandwidth utilization and maximum class variation have been used as criteria.

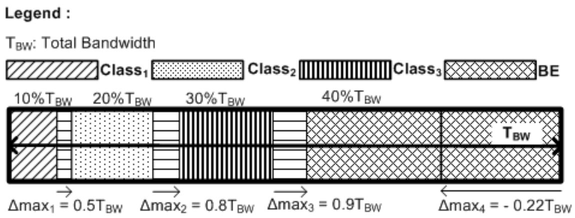


Figure 6.36: Defined Parameters

$\Delta Class_{min}$	$\#steps_i = \frac{\triangle Class}{\triangle Class_{min}}$
1%	100
2%	50
5%	20
10%	10
15%	7
20%	5
50%	2

Table 6.3: Relation between $\Delta Class_{min}$ and $\#steps$

In order to calculate the maximum class variation in each class, an optimization algorithm has been used [140–142].

The main objective of the optimization algorithm is to maximize the link rate utilization on the access router according to the size of flows generated by the sources. One method commonly used to measure the link utility is the sum of the utilities of all sources.

The utility function is referred to as $U(x)$. This utility is the allocation rate obtained by the source x. It has been assumed that the utility function is continuously differentiable, non-decreasing and strictly concave.

The concavity assumption is based on idea that a customer feels the effect of rate increase from 1 to 100kbps much more than one from 1 to 1.1Mbps although the increase is identical in both cases.

For instance, assuming a logarithmic sum-utility as in Figure 6.37 (Figure reprinted from [142], pp. 7), with the utility functions of sources 1 and 2 being $2log(x_1)$ and $log(x_2)$ it can be obtained the desired effect.

Considering y_1, y_2, y_3 and y_4 the aggregate rates of classes 1, 2, 3 and 4 in the access router, with c capacity and S sources, the network utility maximization formulation problem is the maximization of the sum-utility function $\Sigma_s U_s(x_s)$, with the source rates x and subject to following linear constraints:

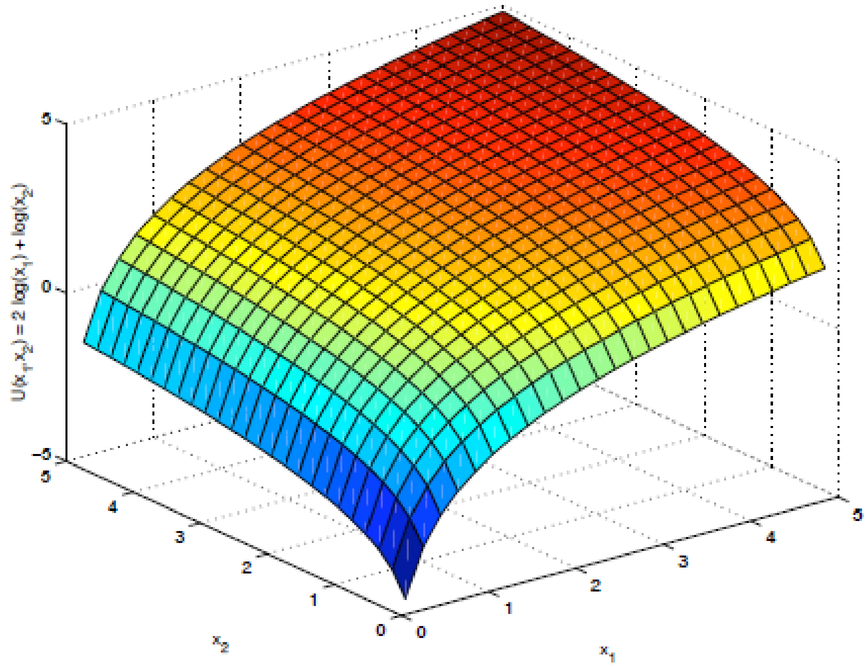


Figure 6.37: Sum Utility of Two Strictly Concave Functions

$$
\begin{aligned}
&\underset{x,y_1,y_2,y_3,y_4 \geq 0}{maximize} && \Sigma U_s(x_s) \\
&subject\,to && \sum_{s\epsilon S} x_s \leq y_i \qquad \forall i = 1,2,3,4 \\
& && y_1 + y_2 + y_3 + y_4 \leq c \\
& && c_{min_i} \leq y_i \leq c_{max_i}
\end{aligned}
\tag{6.1}
$$

In order to decompose the problem into two independent sub-problems, the aggregate rates y_1, y_2, y_3 and y_4., have been fixed. In this situation, the problem is turned into two basic network utility maximization

$$
\begin{aligned}
&\underset{x \geq 0}{maximize} && \Sigma U_s(x_s) \\
&subject\,to && \sum_{s\epsilon S} x_s \leq y_i \quad \forall i = 1,2,3,4
\end{aligned}
\tag{6.2}
$$

Where the master problem is

$$
\begin{aligned}
&\underset{y_1,y_2,y_3,y_4 \geq 0}{maximize} && U_1^*(y_1) + U_2^*(y_2) + U_3^*(y_3) + U_4^*(y_4) \\
&subject\,to && y_1 + y_2 + y_3 + y_4 \leq c \\
& && c_{min_i} \leq y_i \leq c_{max_i} \qquad \forall i = 1,2,3,4
\end{aligned}
\tag{6.3}
$$

Where $U_i^*(y_i)$ is the optimal maximum variation value of 6.2for a given y_i. The results obtained from this optimization algorithm are showed in the next subsection.

The choice of the $\triangle min$ value should be a commitment between network stability, the maximum bandwidth utilization and the optimal maximum class i variation value.

6.5.1 $\triangle min$ Analysis

To analyze the influence that the choice of the $\triangle min$ has on the efficiency of the model architecture, some graphics and results reagrding the throughput for different $\triangle min$ values are presented and discussed. The chosen values for $\triangle min$ have been 10% (or, bandwidth variation in 10 steps), 25% (or, 4 steps) and 50% (or, 2 steps), denominated as Low, Middle and High. The $\triangle min$ has been evaluated under three different scenarios of network load (see Table 6.4)

	Class 1	*Class 2*	*Class 3*	*Class 4*	*Total*
S1	*15.0%*	*31.0%*	*45.0%*	*30.0%*	*121.0%*
S2	*15.0%*	*30.0%*	*48.0%*	*36.0%*	*129.0%*
S3	*15.0%*	*30.0%*	*45.0%*	*42.0%*	*132.0%*

Table 6.4: The Three Scenarios of Network Load in nAR

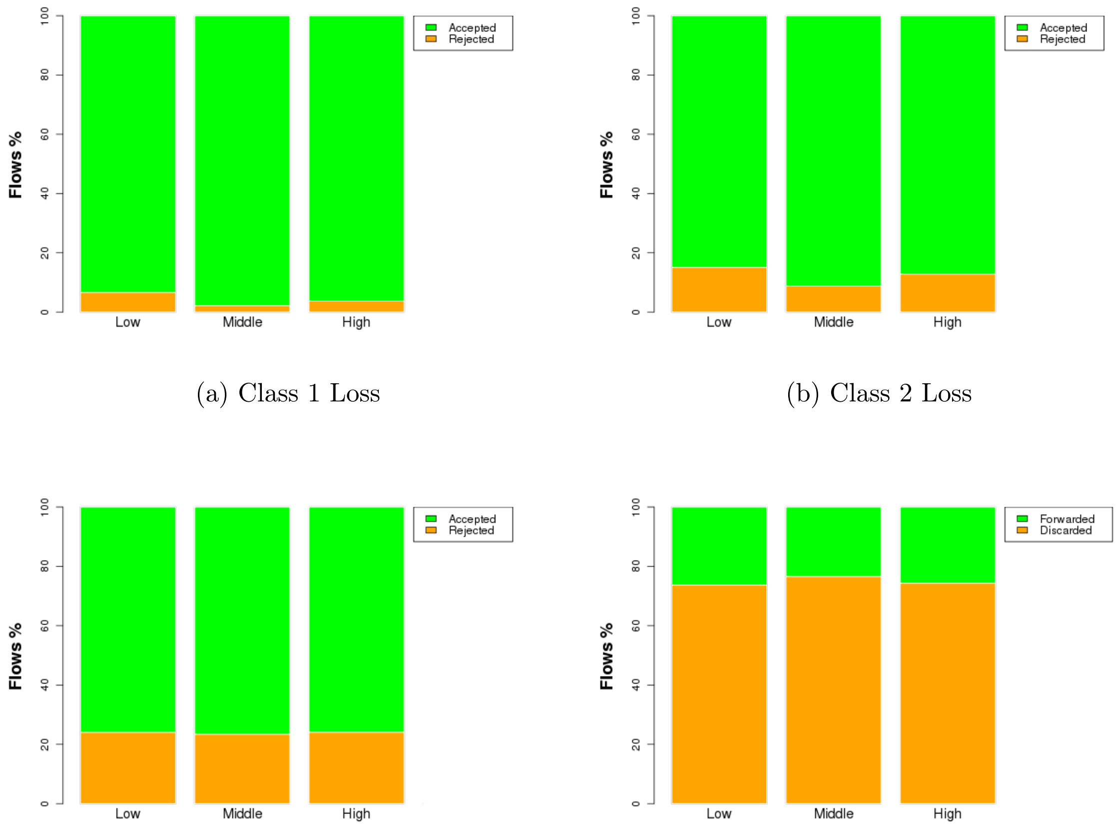
The simulation topology, network configurations and the third scenario (S3) of simulation are the same as those of the previous section.

Figure 6.38 shows the rejected and accepted traffic for priority classes 1, 2 and 3 as well as the forwarded and discarded traffic for class 4. Being as class 4 is not subject to admission control, it is only subject to DiffServ mechanisms, for the three $\triangle min$ values in the scenario S3.

Figure 6.38a shows the percentage of flows lost when subject to a handover process within class 1 for $\triangle min$ values, where Low $\triangle min$ presents losses of 6.6% during handover period, the Middle $\triangle min$ losses of 2.2% and High $\triangle min$ losses of 3.7%.

Figure 6.38d shows the traffic forwarded and discarded in class 4 by DiffServ mechanisms during handover period, where Low $\triangle min$ had 73.7% of discarded packets, Middle $\triangle min$ in turn had 76.6% of discarded packets and High $\triangle min$ had 74.4% of discarded packets.

Therefore, through the interpretation of these results one can conclude that for the present simulation, the $\triangle min$ parameter configured with the Middle value shows less traffic losses in the priority classes 1,2 and 3 . Whereas in class 4, which has the best-effort



(a) Class 1 Loss

(b) Class 2 Loss

(c) Class 3 Loss

(d) Traffic Forwarded and Discarded by DiffServ in the Class 4

Figure 6.38: S3 - Histogram of Class Traffic Behavior During Handover For Distinct Values of $\triangle min$ (Low, Middle and High)

traffic, the Middle $\triangle min$ had less traffic forwarded and more discarded.

Tables 6.5, 6.6 and 6.7 display the reallocated bandwidth for classes 1, 2, 3 and 4, as well as the results of the optimization algorithm in scenarios S1, S2 e S3, respectively.

The tables show that the Middle $\triangle min$ has achieved better bandwidth utilization for the priority class 1 in the tested scenarios, and one can observe that the $\triangle min$ has a considerable impact in the bandwidth distribution among classes. It can also be observed that the relation between data flow rate and $\triangle min$ influences the amount of reallocated bandwidth, i.e. if the flow rate and the $\triangle min$ step of a given class are closer, the reallocation mechanism achieves higher values of bandwidth utilization.

For instance, in scenario S2, the flow rate in class 1 is $0.03T_{BW}$(kbps) which represents a percentage utilization of 14.2% for a Middle $\triangle min$ with a step size of $0.012.T_{BW}$, whereas

	Low	*Middle*	*High*	*Optim*	*Load(nAR)*
Class 1	14.1%	14.2%	14,2%	*15%*	16%
Class 2	25.3%	27,6%	27,1%	*28%*	32%
Class 3	38.7%	38,8%	38.1%	*39%*	48%
Class 4	21.8%	19.7%	20.5%	*18%*	28%

Table 6.5: S1 - Bandwidth Utilization Percentage, Optimization value and Traffic Load in nAR

	Low	*Middle*	*High*	*Optim*	*Load(nAR)*
Class 1	*14.0%*	*14.2%*	*14.2%*	15%	*15%*
Class 2	*26,6%*	*27,5%*	*27,2%*	28%	*30%*
Class 3	*38,2%*	*39.0%*	*38,5%*	39%	*45%*
Class 4	*21,0%*	*19,2%*	*20,0%*	18%	*36%*

Table 6.6: S2 - Bandwidth Utilization Percentage, Optimization value and Traffic Load in nAR

in scenario S3 with a flow rate of $0.015T_{BW}$(kbps) a percentage utilization of 14.7% in class 1 has been achieved.

	Low	*Middle*	*High*	*Optim*	*Load(nAR)*
Class 1	*14,4%*	*14,7%*	*14,5%*	15%	*16%*
Class 2	*25,2%*	*26,0%*	*26,0%*	28%	*32%*
Class 3	*38.1%*	*38,4%*	*38,4%*	39%	*48%*
Class 4	*22,3%*	*20,9%*	*21,1%*	18%	*42%*

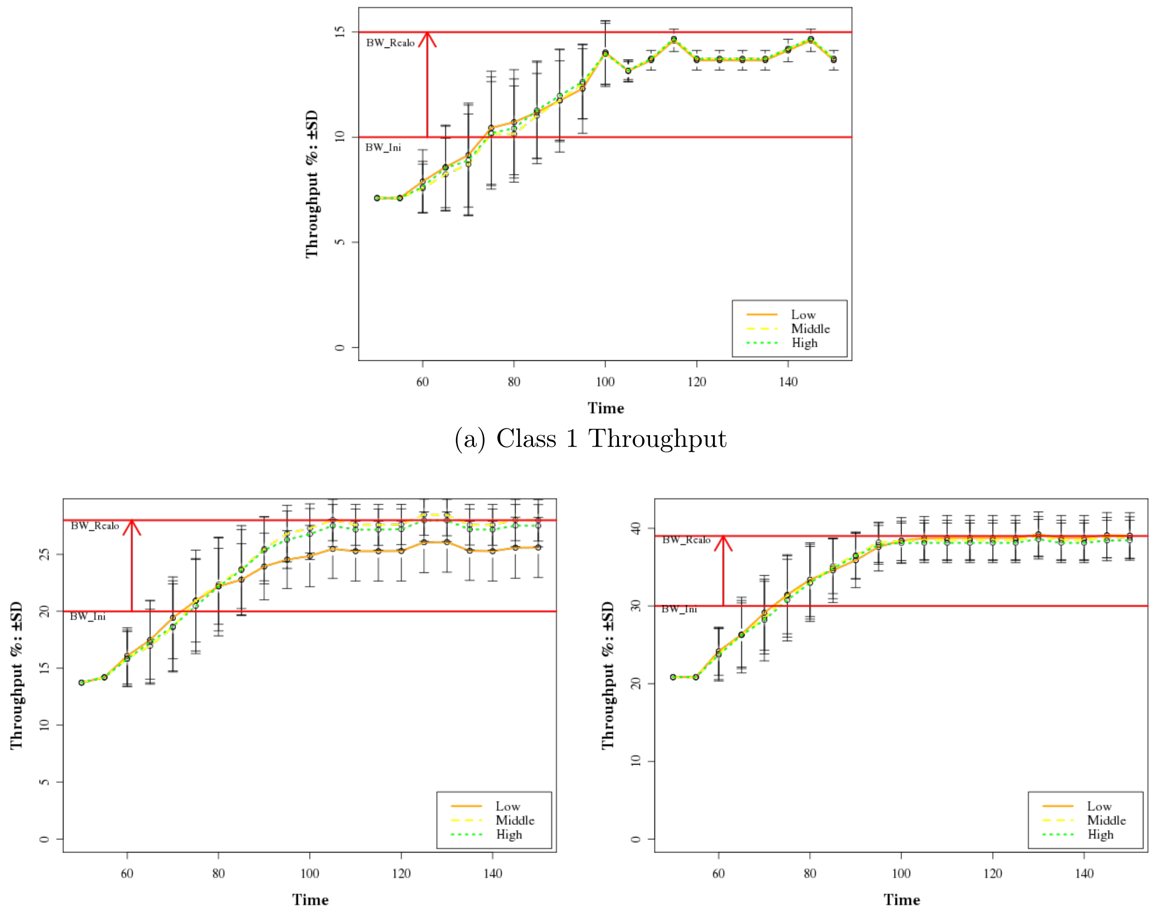
Table 6.7: S3 - Bandwidth Utilization Percentage, Optimization value and Traffic Load in nAR

Equally important is the fact that despite in scenario S2, the traffic generated for class 4 (S1:36%, Tab. 6.4) did not totally fill the allocated bandwidth for this class (40% of allocated bandwidth, Fig. 6.36) the reallocation mechanism takes advantage of the available bandwidth in the class 4 to increase the allocated bandwidth for priority classes, thus increasing the bandwidth utilization to approximately its maximum capacity. Obviously, according the AC algorithm policies, this improvement may also imply the decrease of BE throughput if the allocated bandwidth for this class is totally occupied.

Furthermore, Figures 6.40a and 6.41a also show that in this case, the reallocated bandwidth converges quickly to the maximum variation value.

Figures 6.39, 6.40 and 6.41 show the ongoing throughput during the handover period for each DiffServ class when subject to the three $\triangle min$ values on the three scenarios.

With regards to the optimization values also illustrated in the tables 6.5,6.6 and 6.7,



(a) Class 1 Throughput

(b) Class 2 Throughput

(c) Class 3 Throughput

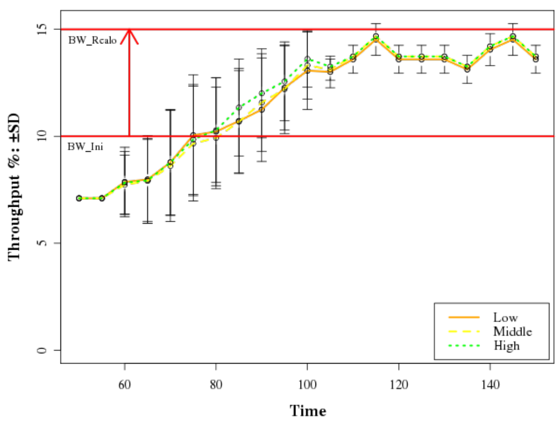
Figure 6.39: S1 - Class Throughput For Distinct Values of $\triangle min$ (Low, Middle and High)

it can be stated that the optimization values are equal to the $\triangle max$ plus the initially allocated bandwidth on a given class ($ClassBW_i$). This is due to the fact that it has applied a more aggressive behavior to priority classes by providing higher weights[1] to the logarithmic utility function on those classes.

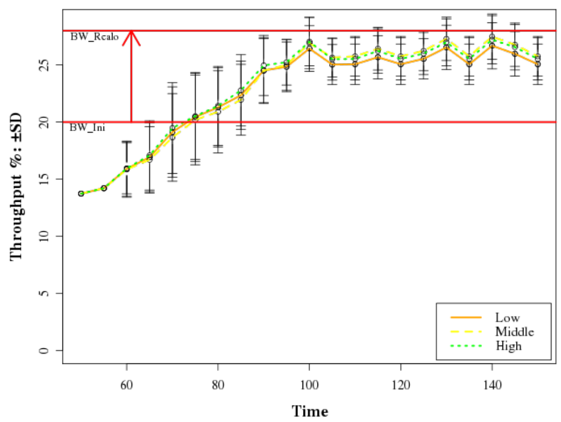
In summary, based on the results obtained for the three scenarios, one can conclude that the Middle $\triangle min$ achieves a better bandwidth utilization percentage for the priority classes when compared to the other two $\triangle min$ values, being Low $\triangle min$ the poorer.

In this sense one can say that the better $\triangle min$ for the proposed model is the one that achieves a bandwidth utilization percentage closest to the $Optim$ value. Thus, by analyzing the results presented and by taking into account the criteria of network stability and the simulation results, one can verify that a $\triangle min = 25\%$ was showed and that in the given simulated scenarios, is the best choice to parametrize the proposed architecture.

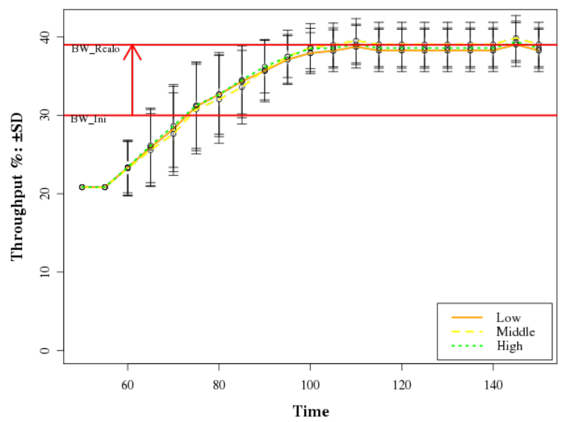
[1]The values assigned respective to class 1,2 and 3 were 4, 3 and, 2.



(a) Class 1 Throughput



(b) Class 2 Throughput



(c) Class 3 Throughput

Figure 6.40: S2 - Class Throughput For Distinct Values of $\triangle min$ (Low, Middle and High)

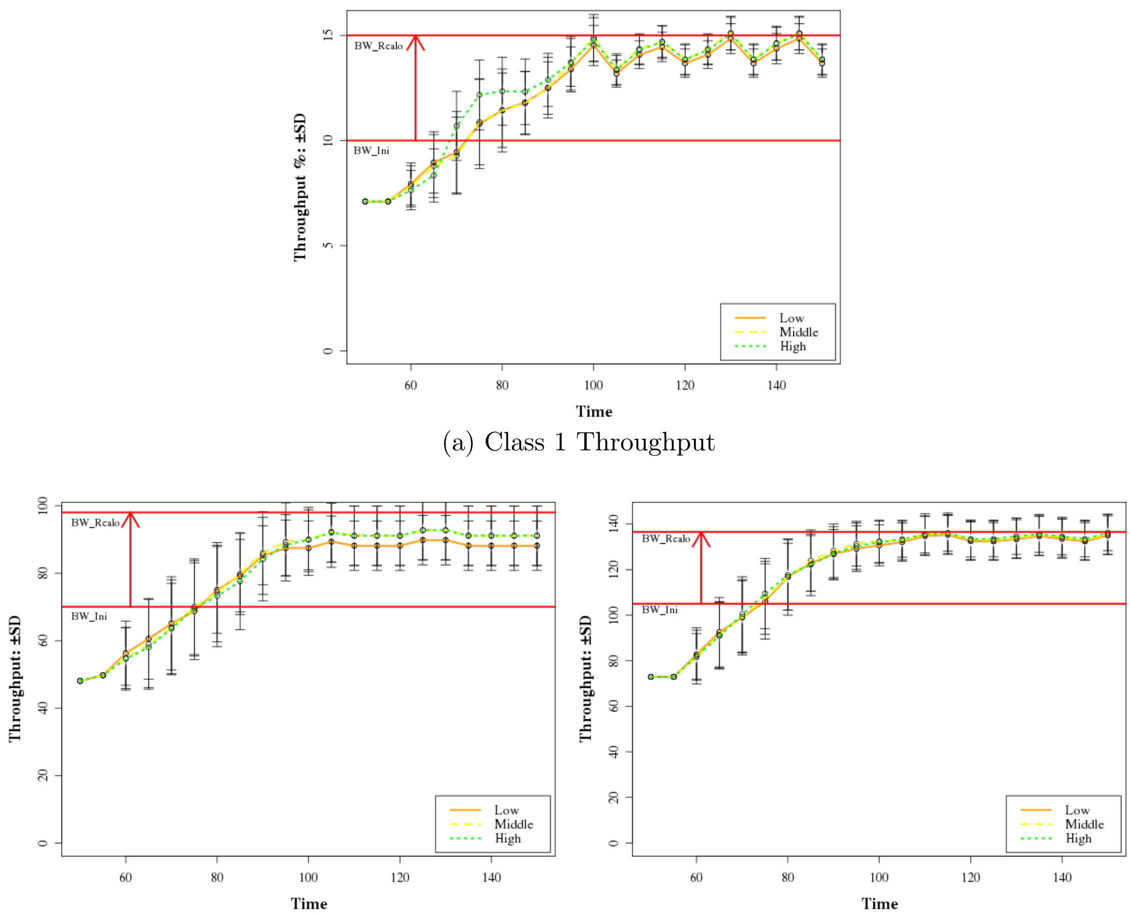
The $\triangle min = 50\%$ could also be a good choice if the option is to have a more stable network in detriment of bandwidth utilization.

On the other hand, if the option is to have a more flexible reallocation the choice should be a smaller step size such as $\triangle min = 10\%$ but this option will generate more instability in the system.

6.6 Conclusion

In this chapter the rate estimators TSW and EMA have been analysed in the perspective of selecting which is the most suitable for global architecture. The results indicate that the TSW rate estimator presents a better time responsiveness behavior to mobile node handovers than the EMA. Therefore, it is the most suitable to mimic the dynamic behavior of mobile environments.

After the proposed model was analysed along with its dynamic allocator element. The



(a) Class 1 Throughput

(b) Class 2 Throughput

(c) Class 3 Throughput

Figure 6.41: S3 - Class Throughput For Distinct Values of $\triangle min$ (Low, Middle and High)

QoS mechanisms for model architecture used, proved to be able to avoid link congestion and misbehave classes limiting their throughput to the allocated bandwidth. Consequently, significant improvements in loss, jitter and delay can be achieved when compared to the other scenarios.

In addition to this, has been added to the the dynamic allocator element has been added to the QoS framework of model architecture . The dynamic allocator permitted a significant enhancement in the accepted handover flows belonging to priority classes.

Furthermore, it also provides an improved bandwidth network utilization by reallocating bandwidth to those in need, thus improving resource utilization efficiency in a environment where the resources are scarce. Furthermore, results show that the model is capable of providing the desirable QoS for the handover flows and for the existing flows.

Finally, in what respects to the model parametrization, the $\triangle min$ value with the best commitment between the criteria of network stability and maximum bandwidth utilization was the Middle $\triangle min = 25\%$, which means that for the studied scenarios the reallocation mechanism should have four steps. It has also been observed that the relation between the flow rate size and the step unit size influences the quantity of the reallocated bandwidth. If one has the approximate values, the reallocation mechanism can achieve higher bandwidth utilization values.

Chapter 7

Conclusions and Future Work

The massive expansion of wireless networks represents a new world when it comes to application's services which now require new quality demands. This uncontestable reality results in the necessity to find new QoS models or adjust existing ones to the specific characteristics of mobile environments.

The question now becomes: what is the best solution capable of satisfying these new quality demands for mobile networks? This research work presents, discusses and analyzes one solution that will maintain acceptable QoS levels to mobile terminals running real-time applications.

The maintenance of QoS levels in mobility are extremely important in video (e.g. NetMeeting) and audio (e.g. Skype) applications because they contain stricter quality of service restrictions in order to function correctly and in an appropriate manner therefore, the degradation of quality of service because of mobility could prevent a satisfactory use of these multimedia applications in mobile environments.

This Chapter presents the concluding remarks regarding the developed research work, a synthesis of the research objectives covered, a discussion concerning the main research contributions, and finally a few guidelines for future research work.

7.1 Summary

The future architecture of the Internet, in addition to supporting global mobility will also enable multimedia applications to explicit their requirements of network service with the quality desired by the respective applications and consequently the network instead of solely providing a unique service, the best service (best- effort), will support multiple service classes that provide probabilistic performance guarantees to multimedia applications.

This work analyzes and proposes a solution for improving the resource management system which supports the mobility of terminal equipment (such as the iPad, PDAs, iPhones and other cell phones) by properly and dynamically managing the available resources and endowing the network service with multi-class probabilistic quality guarantees to the future multimedia applications.

This new solution for managing mobile resources is well adapted to mobile environments with high dynamics thus, it contains adaptive features which are achieved through the exchange of signaling messages and a distributed resources management.

The maintenance of the quality of service is achieved by transferring the prior QoS context of the mobile device to the new Internet access router before the mobile device is established in the new location. The mechanisms implemented in the access routers, based on context information received and in its current state, automatically reconfigure themselves in order to provide an equivalent quality of service to the mobile device in the new access router. Thus, the quality of service offered to the application running on the mobile device becomes separate from mobility, i.e. separate from the different access conditions that the mobile device can find when in movement.

The evaluation of the proposed model was elaborated using a simulation tool, for this purpose, a simulation model able to represent real situations of user mobility was developed. In the simulations typical speeds for several user mobility scenarios were considered (walk, run, bike, car).

Currently, Internet users want mobility and service quality simultaneously. In this context, one can state that the expansion of multimedia applications in the mobile Internet market depends largely on the quality of service offered. The importance of this solution is related to the possibility of preventing degradation of quality of service offered to applications that run on mobile devices located in a highly dynamic environment where resource availability vary quickly.

In fact, today there is already a quite high resource availability on the existing infrastructure of fixed networks (e.g by using optical fiber) but wireless access is still the place where resources are typically scarce thus requiring good management.

Therefore, this research work proposes a model which provides dynamic QoS provisioning to local mobility that could eventually be easily extended to global mobility. The proposed model aims to enhance global mobility with efficient handovers and quality of service. For this purpose two enhancements were conducted. The first enhancement was the specific integration of FMIPv6 and HMIPv6 (F-HMIPv6) to improve MIPv6 handover latency. The second enhancement was the extension of the standard DiffServ resource management with dynamic and adaptive QoS provisioning. The model uses explicit and implicit setup mechanisms to request resources from the network in order to support admission control and optimize resource allocation. In order to optimize resource allocation, resource and mobility management were coupled resulting in a QoS/Mobility aware network architecture capable of having a proactive behavior with regards to mobility events. In order to avoid signaling overhead and resorting to a complex bandwidth broker, the model offers end-to-end predicted services which provide highly reliable services but without absolute guarantees.

Simulation results indicate that the solution avoids network congestion as well as the starvation of less priority DiffServ classes while increasing resource utilization for priority classes while maintaining the QoS offer to MN's applications by making each DiffServ class unchangeable with MNs mobility.

The proposed mobility management model is simple, easy to implement, and takes into consideration mobile internet requirements. It proved to be capable of providing Internet applications running in mobile devices with acceptable levels of quality of service.

7.2 Covered Objectives

This section intends to make a retrospective of the objectives covered in this research work.

a) ***Investigate the IP mobility and QoS paradigms as well as its mechanisms to control bandwidth access more specifically admission control mechanisms and QoS signaling protocols used in dynamic resources reservations***;

   This objective was addressed in Chapter 2. The first section of the second chapter provides the grounding information concerning IP mobility. Then, the QoS models, IntServ and DiffServ, were discussed and compared using a conceptual and practical perspective. Their applicability to mobile environments was also addressed. After this, the QoS mechanisms used by both QoS models in order to control traffic were examined, more specifically admission control mechanisms and signaling protocols because they are essential elements for providing dynamic QoS support in the Internet networks.

b) ***Essential requirements for a QoS-aware mobile network solution***;

The accomplishment of this objective is very important in order to design an appropriate QoS solution for mobile networks. This issue was addressed in section (2.4 on page 42) and section 3 (2.5 on page 44) of Chapter 2. These discuss the challenges that handover imposes on QoS architectures and provides a few guidelines in the design of a suitable QoS solution for mobile environments.

c) ***Survey of the existing dynamic QoS solutions;***

This objective was addressed in Chapter 3 and intends to act as a survey when it comes to this specific subject. Therefore, a detailed description about the current dynamic QoS solutions for wired and wireless netwoks was provided, among them IntServ over DiffServ, RSVP Aggregation, SCORE, Bandwidth Brokers, Resource Management in DiffServ, Mobile Extensions to RSVP, QoS-Conditionalized Handoff for Mobile IPv6, Hierarchical Mobile RSVP, Policy Based Management System and a few projects concerning QoS in mobile environments were also mentioned briefly.

d) ***Evaluation of the current dynamic QoS solutions in the perspective of their suitability to mobile networks;***

After making the survey of the existing dynamic QoS solutions, the problems of signaling overhead, processing work load and excessive state variables were outlined as problems that have not been completely solved by the current dynamic QoS solutions.

Most of the dynamic QoS solutions are based on the guaranteed service model thus when enforced on mobile wireless networks, they introduce extra signaling overhead due to the QoS renegotiation during handovers in an environment where resources are usually scarce. Consequently, significant scalability problems may arise with these QoS solutions when in wireless networks. Therefore, the current dynamic QoS solutions have difficulties in self-adapting its resource reservations to the dynamic behavior and changeable nature of wireless networks.

e) ***Design a Mobility/QoS-aware network architecture solution;***

The proposed model is defined in Chapter 4. This Chapter includes a detailed description of the architectural components in addition to an explanation of how they are inter-connected. The Mobility/QoS-aware solution is capable of:

I - **support roaming with the same QoS without tearing down the active connections and re-establishing a new session. This should comprise horizontal and vertical handovers;**

By means of handover layer-3 antecipation provided by an optimized mobility management scheme (F-HMIPv6) the construction of a QoS architecture coupled with this mobility management scheme, that is able to efficiently redirected the MN's traffic with the same QoS context for the new location without the need of re-establishing a new session from scratch, has been created.

The QoS architecture was designed at IP level, therefore being a QoS solution for layer-3 it is able to comprises QoS handovers in different access technologies.

II - **provide dynamic QoS functionalities for resource management;**

This feature was accomplished using a signaling protocol for new flows making their QoS requests to the network and using an in-band signaling for handover flows making their Qos requests to the new access router. The QoS signaling, and HI and HAck mobility messages are used as inputs parameters of the resource management function, which after receiving these messages proceeds with the appropriated actions. The actions taken are consistent with the QoS information contained in the messages and in the current network status.

The use of mobility messages for conveying the MN's QoS context allows the resource management function to be aware of mobility events and provides a resource management that is able to proactively react to mobile events.

III - **provide adaptive resource management features;**

The dynamic nature of the mobile environment caused by handovers demands a resource management function with QoS mechanisms able to self-adjust to the constant load variations in the access routers. This objective was achieved by resorting to a reallocation mechanism built on a hysteressis method, which by taking into account the network status and the MN's QoS requirements, is capable of self-adapting its configuration in order to fulfill the QoS requirements of new and existing traffic.

IV - **achieve fast and seamless handovers;**

The proposed model enhances the MIPv6 protocol with a specific integration of FMIPv6 and HMIPv6 (F-HMIPv6). The F-HMIPv6 enhances the MIPv6 mobility with fast and seamless handovers as well as local registrations.

V - **resolve the scalability problems that may arise with MIPv6 protocol and dynamic QoS architectures;**

The Mobile IP was found to not be optimal for supporting local mobility because it generates a significant amount of signaling traffic in the core network with mobile node movement and creates a significant delay in the diffusion of mobile node localization updates. These facts rise scalability problems. Therefore, to shorten handover delay and signaling traffic, the specific integration of FMIPv6 and HMIPv6 (F-HMIPv6) was proposed.

When it comes to state information overhead, signaling overhead and processing load problems caused by the guaranteed service mode,l our approach effort has been to solve these problems with relaxed QoS requirements i.e., the predictive service model of the DiffServ QoS model. Furthermore, as the admission control scheme chosen is based on class traffic measurements, signaling, state information and processing load are minimized.

VI - **fulfill QoS requirements in the new access router for incoming and existing traffic;**

The QoS context received in advance by means of the F-HMIPv6 handover antecipation feature permits the anticipated decision of the admission of new handover flows only if the QoS requirements of the existing and incoming flows are fulfilled.The establishment of the MN's QoS context on nARs before the handover takes place also occurs. This avoids the re-establishment of the MN's QoS context on nAR from scratch and permits the provision of a proactive resource management which is able to behave according to the QoS requirements of existing and incoming traffic.

f) **implementation and evaluation of the Mobility/QoS-aware network architecture solution.**

This objective was addressed in Chapters 5 and 6. The implementation of the proposed model in a simulation prototype provided a flexible test bench which allowed rigorous and extensive evaluations. The selection of NS-2 as a simulation platform conceded the flexibility required to improve and adjust the model functionality to what was expected in conceptual and pratical terms.

The data collected from the simulations was loaded and merged into the R software, whereby a throughly statistical analysis and the generation of graphics was performed in order to make a proof of concept of the research work.

7.3 Work Contributions

The main contributions of this thesis are as follows:

1. ***Evaluation of a Suitable Rate estimator for Mobile Environments*** - Being as the rate estimators are essential mechanisms for the efficiency of the global archi-

tecture, a thorough study was conducted in order to select the best rate estimator for mobile scenarios. For this purpose two rate estimators - Time Sliding Window (TSW) and Exponential Moving Average (EMA) - were studied and evaluated by means of simulations in a QoS-aware wireless mobility scenario. In this study, the TSW estimator indicated a better time-responsiveness than EMA and a reasonable accuracy with regards to the traffic estimation during MN handovers period. This contribution has been published in the 17th International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2009) [13].

2. ***A QoS Framework Encompassing Mobile Requirements*** - The solution comprises enhancements in the mobility management of Mobile IPv6 (MIPv6) and in the resources management of Differentiated Services (DiffServ) QoS model. The mobility management of MIPv6 was extended with fast and local handovers to improve its efficiency in micro-mobility scenarios with frequent handovers. The resource management of DiffServ has been extended with adaptive and dynamic QoS provisioning to improve resources utilization in mobile networks. Furthermore, in order to optimize resource utilization the mobility and QoS messages were coupled for providing resource management able to proactively react and adapt to mobile events. The solution results from a combination of Fast and Hierarchical Handovers, in-band signaling, DiffServ resource management, QoS context transfer and a Measurement-Based Admission Control (MBAC) algorithm. These were integrated into a single and suitable QoS framework solution for mobile environments. This symbiotic combination of components were optimized to work together in order to support seamless handovers with QoS requirements suitable for mobile users running multimedia applications. This solution showed to be effective, simple, scalable andcapable of avoiding the well known traditional resource reservation problems of the guaranteed service model such as state maintenance, signaling overhead and processing load. This contribution has been published in the "Actas da 10a Conferência sobre Redes de Computadores" (CRC 2010) [14].

3. ***Dynamic Allocator*** - The admission control algorithm implemented in access routers

was extended with a reallocation mechanism based on the hysteresis method, called the dynamic allocator. The dynamic allocator's main objective was to achieve a better resource utilization while simultaneously increasing the number of accepted MN's classes meeting the required QoS. The dynamic allocator has proven that it can increase the number of accepted flows during handover through the reduction of the bandwidth allocated for BE in favor of priority DiffServ classes. This will lead to a perceived QoS improvement from the customer's point of view. From the network operator's point of view, this solution may also improve network resources utilization and consequently increase the ISP revenues. This contribution has been published in the International Conference on Ultra Modern Telecommunications (ICUMT 2009) [15].

4. **Model Implementation in the Network Simulator 2** - In order to evaluate the proposed QoS solution, it was deployed in the NS-2 its simulation model. The simulations results showed that the model can provide high reliable and predictive services. However, it cannot assure absolute QoS guarantees to applications however, given the unpredictable nature of wireless links even with the guaranteed service model implemented, it would difficult to provide absolute guarantees in these environments. This contribution has been accepted in April 2011 for publication in the Journal of Telecommunications Systems [16] and it has also been published in a technical report [17].

5. **A Study to Parametrize the Proposed Model** - In order to make an appropriate (not arbitrary) model parametrization a criteria based on network stability has been suggested: maximum bandwith utilization andn an optimization algorithm. This study uses the Δmin as the network stability metric and the optimal maximum class variation value provided by the optimization algorithm as the performance target. Where the Δmin parametrization value should be selected based on the commitment between these three criteria. This contribution has been published in the 18th International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2010) [18].

6. ***A Distributed Resource Management Scheme for Diffserv QoS Model*** - The distributed resource management scheme is a set of resource management functions which are located in the edge routers of the standard DiffServ architecture. The resource management function is comprised by the DiffServ QoS mechanisms (policer, congestion avoidance and scheduling) and a measurement-based admission control mechanism (estimator and admission control algorithm). The resource management function in the Access Routers (ARs) contains an additional element called the dynamic allocator to improve the network utilization with adaptive resource management. This work proposes a QoS micro-mobility solution able to provide QoS support for global mobility in order to enhance global mobility with efficient handovers and QoS. This contribution was accepted in April 2011 for publication in the Journal of Communication Software and Systems [19].

7. ***The Importance of DiffServ Model in a Mobile Environment*** - The importance of the differentiated services in mobile environments was evaluated resorting to a specific implementation of the DiffServ QoS model in IP core of a UMTS network. For this effect, the mapping of DiffServ codepoints into UMTS classes on UMTS core network was applied. Along with this Diffserv implementation was possible to evaluate the importance of Diffserv mechanisms, when facing traffic stressing conditions, in two separate contexts: the purely technical context and the economic context. In the technical context, the implemented DiffServ architecture was proven to meet the ITU QoS recomendations and also increase the bandwidth utilization. In what pertains to the economic context, a revenue function which estimates the profits that ISP could expect with the DiffServ implementation was proposed. The revenue function showed that profits can substantially increase. This contributions have been published in the 2nd International Conference on New Technologies, Mobility and Security (NTMS 2008) [12].

7.4 Future Work

Although a proposal for extending the model for global mobility has been made in this research work, its implementation and evaluation in the NS-2 remains for future work. Another intention for a future work is to support secure end-to-end QoS services for real-time applications accross heterogeneous domains. Therefore, it intends to add a new element to the proposed model with an Authentication, Authorization, Accounting and Charging component (AAAC). Where the AAAC has the role of authenticating a user from a foreign domain, grant a given contracted service and control the payment of used resources.

The global solution would be implemented in the NS-2 libraries in order to be subject to extensive simulations in scenarios with heterogeneous mobile networks domains and with several traffic models.

Furthermore, is also intend to design an analytical model for the proposed model in order to compare the simulation results with the analytical results to substantiate its definition.

Another future development is the analysis of the signaling overhead introduced by inter-domain handovers.

In conclusion, it intends to pursue with this research work by extending its features with a new AAAC element and with the capacity of the network self-adapts its configuration to the different network environments requirements by means of optimization algorithms. This way, the wireless network should be capable of optimizing its traffic control mechanisms, i.e., congestion control, routing and scheduling, in order to fairly allocate the resources for the users while maximizing its utilization via optimization algorithms.

The tremendous increase of mobile traffic places a great demand on network capacity and the quality of service of mobile networks. The radio access networks are evolving to various fourth-generation (4G) systems such as International Mobile Telecommunications (IMT) - Advances systems (e.g. Long Term Evolution (LTE) in the 3GPP and WiMAX). The mobile nodes of the future mobile network will have to support vertical handovers between heteregeneous IMT -advanced systems for allowing a huge bandwith capacity, where the mobile node can select the radio access technology that offers the better service quality.

Therefore, another possibility for a future work would be to extend the model architecture so that mobile nodes could discover the radio access network that offers better quality of service to priority classes and select it by making vertical handovers.

Bibliography

[1] A.-E. Taha, H. Hassanein, and H. Mouftah, "Extensions for Internet QoS paradigms to mobile IP: a survey," *Communications Magazine, IEEE*, vol. 43, no. 5, pp. 132–139, May 2005.

[2] D. Johnson, C. Perkins, and J. Arkko, "Mobility Support in IPv6," RFC 3775 (Proposed Standard), Internet Engineering Task Force, Jun. 2004. [Online]. Available: http://www.ietf.org/rfc/rfc3775.txt

[3] H. Soliman, C. Castelluccia, K. ElMalki, and L. Bellier, "Hierarchical Mobile IPv6 (HMIPv6) Mobility Management," RFC 5380 (Proposed Standard), Internet Engineering Task Force, 2008. [Online]. Available: http://www.ietf.org/rfc/rfc5380.txt

[4] E. A R. Koodli, *Mobile IPv6 Fast Handovers*, rfc 5568 ed., IETF, July 2009. [Online]. Available: http://www.ietf.org/rfc/rfc5568.txt

[5] A. Campbell, J. Gomez, S. Kim, A. Valko, C.-Y. Wan, and Z. Turanyi, "Design, implementation, and evaluation of cellular IP," *Personal Communications, IEEE*, vol. 7, no. 4, pp. 42 –49, aug 2000.

[6] A. Wali, M. Ahmad Ghazali, and S. R. Ayyubi, "Comparative Analysis of Mobile IP and HAWAII," in *Wireless Networks, Information Processing and Systems*, ser. Communications in Computer and Information Science, D. M. A. Hussain, A. Q. K. Rajput, B. S. Chowdhry, and Q. Gee, Eds. Springer Berlin Heidelberg, 2009, vol. 20, pp. 169–179.

[7] R. Braden, D. Clark, and S. Shenker, "Integrated Services in the Internet Architecture: an Overview," RFC 1633 (Informational), Internet Engineering Task Force, Jun. 1994. [Online]. Available: http://www.ietf.org/rfc/rfc1633.txt

[8] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Service," RFC 2475 (Informational), Internet Engineering Task Force, Dec. 1998, updated by RFC 3260. [Online]. Available: http://www.ietf.org/rfc/rfc2475.txt

[9] I. Mahadevan and K. Sivalingam, "Quality of Service architectures for wireless networks: IntServ and DiffServ models," in *Fourth International Symposium on Parallel Architectures, Algorithms, and Networks, 1999. (I-SPAN '99) Proceedings.*, 1999, pp. 420 –425.

[10] A. Jamalipourn and J. Kim, "Measurement-based admission control scheme with priority and service classes for application in wireless IP networks," *International Journal of Communication Systems*, vol. 16, pp. 535–551, May 2003.

[11] D. Black, "Differentiated Services and Tunnels," RFC 2983 (Informational), Internet Engineering Task Force, Oct. 2000. [Online]. Available: http://www.ietf.org/rfc/rfc2983.txt

[12] N. Lopes, M. Nicolau, and A. Santos, "Efficiency of PRI and WRR DiffServ Scheduling Mechanisms for Real-Time Services on UMTS Environment," in *2nd IFIP International Conference on New Technologies, Mobility and Security, 2008. NTMS '08*, Nov 5-7 2008, pp. 173–178.

[13] N. V. Lopes, M. J. Nicolau, and A. Santos, "Evaluating rate-estimation for a mobility and QoS-aware network architecture," in *17th International Conference onSoftware, Telecommunications Computer Networks (SoftCOM 2009)*, Sept. 2009, pp. 348 –352.

[14] N. V. Lopes, M. Nicolau, and A. Santos, "A QoS-Enable Solution for Mobile Environments," in *Actas da 10a Conferencia sobre Redes de Computadores*, M. N. e. a. A. Santos, P. Carvalho, Ed., no. ISBN: 978-989-96929-1-6. Universidade do Minho, 2010, pp. 43–48.

[15] N. Lopes, M. Nicolau, and A. Santos, "QoS-aware architecture for FHMIP micromobility," in *International Conference on Ultra Modern Telecommunications Workshops, 2009. ICUMT '09.*, Oct. 2009, pp. 1 –8.

[16] N. V. Lopes, M. J. Nicolau, and A. Santos, “A QoS-Enabled Resource Management Scheme for F-HMIPv6 Micro Mobility Approach,” *Journal of Telecommunications Systems*, 2011, accepted in April 2011.

[17] N. Lopes, M. Nicolau, and A. Santos, “A Novel Resource Management scheme for FHMIP Micromobility Protocol,” DI-CCTC Universidade do Minho, Tech. Rep. DI-CCTC-10-03, 2010.

[18] N. V. Lopes, M. J. Nicolau, and A. Santos, “A Micro-Mobility Solution for Supporting QoS in Global Mobility,” *18th International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2010)*, 23-25 September 2010, (Best Student Paper Award).

[19] N. V. lopes, M. J. Nicolau, and A. Santos, “A QoS/Mobility-Aware Model for Mobile Internet,” *Journal of Communication Software and Systems*, 2011, accepted in April 2011.

[20] R. Zhang and K. Long, “QoS issues in mobile IP: challenges, requirements and solutions,” in *ICCC ’02: Proceedings of the 15th international conference on Computer communication*, Washington, DC, USA, 2002, pp. 802–812.

[21] D. O. Awduche and E. Agu, “Mobile extensions to RSVP,” in *IC3N ’97: Proceedings of the Sixth International Conference on Computer Communications and Networks*. Washington, DC, USA: IEEE Computer Society, 1997, p. 132.

[22] S.-J. Leu and R.-S. Chang, “Integrated service mobile internet: RSVP over mobile IPv4/6,” *Mob. Netw. Appl.*, vol. 8, no. 6, pp. 635–642, 2003.

[23] N.-F. Huang and W.-E. Chen, “RSVP extensions for real-time services in hierarchical mobile IPv6,” *Mob. Netw. Appl.*, vol. 8, no. 6, pp. 625–634, 2003.

[24] G. Fankhauser, S. Hadjiefthymiades, N. Nikaein, and L. Stacey, “Interworking RSVP and Mobile IP Version 6 in Wireless Environments,” *Proc. 4th ACTS Mobile Communications Summit*, 1999.

[25] A. Belhoul, A. Y., and N. Mani, “Mobility protocols and RSVP performance in wireless IPv6 networks: shortcomings and solutions,” *Wireless Communications*

and Mobile Computing, vol. 8, no. 9, pp. 1183–1199, 2008. [Online]. Available: http://dx.doi.org/10.1002/wcm.560

[26] C. Perkins, "IP Mobility Support for IPv4," RFC 3344 (Proposed Standard), Internet Engineering Task Force, Aug. 2002, updated by RFC 4721. [Online]. Available: http://www.ietf.org/rfc/rfc3344.txt

[27] W. Simpson, "IP in IP Tunneling," RFC 1853 (Informational), Internet Engineering Task Force, Oct. 1995. [Online]. Available: http://www.ietf.org/rfc/rfc1853.txt

[28] J. Arkko, C. Vogt, and W. Haddad, "Enhanced Route Optimization for Mobile IPv6," RFC 4866 (Proposed Standard), Internet Engineering Task Force, May 2007. [Online]. Available: http://www.ietf.org/rfc/rfc4866.txt

[29] *IETF Mobile IP Working Group*, http://www.ietf.org/html.charters/mobileip-charter.html (Last visited 05/2011).

[30] *IETF Seamoby Working Group*, http://www.ietf.org/html.charters/seamoby-charter.html (Last visited 05/2011).

[31] P. Reinbold and O. Bonaventure, "A Survey of IP micro-mobility protocols," *IEEE Communications Surveys and Tutorials*, vol. 5, pp. 40–57, 2002.

[32] P. M. S. R. V. Estrela, "Transparent and Efficient IP Mobility," Ph.D. dissertation, Instituto Superior Tecnico, December 2007.

[33] A. T. Campbell and J. Gomez-Castellanos, "IP Micro-mobility protocols," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 4, no. 4, pp. 45–53, 2000.

[34] X. Wu and G. Nie, "Comparison of different mobility management schemes for reducing handover latency in mobile ipv6," in *Industrial Mechatronics and Automation, 2009. ICIMA 2009. International Conference on*, may 2009, pp. 256 –259.

[35] C.-H. Shih, J.-L. Kuo, C.-H. Huang, and Y.-C. Chen, "A proxy-based fast handover scheme for hierarchical mobile ipv6," in *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*, ser. ICUIMC '11. New York, NY, USA: ACM, 2011, pp. 21:1–21:10. [Online]. Available: http://doi.acm.org/10.1145/1968613.1968639

[36] R. Koodli, "Mobile IPv6 Fast Handovers," RFC 5568 (Proposed Standard), Internet Engineering Task Force, July 2009. [Online]. Available: http://www.ietf.org/rfc/rfc5568.txt

[37] Y. Gvvon, J. Kempf, and A. Yegin, "Scalability and robustness analysis of mobile IPv6, fast mobile IPv6, hierarchical mobile IPv6, and hybrid IPv6 mobility protocols using a large-scale simulation," in *2004 IEEE International Conference on Communications*, vol. 7, June 2004, pp. 4087–4091.

[38] H. JEONG, S. MAENG, and Y. CHAE, "HIMIPv6: An Efficient IP Mobility Management Protocol for Broadband Wireless Networks," *IEICE Transactions on Information and Systems*, vol. E92.D, no. 10, pp. 1857–1866, 2009.

[39] H. Jung and S. Koh, "Fast handover support in hierarchical mobile IPv6," in *the 6th International Conference on Advanced Communication Technology*, vol. 2, 2004, pp. 551–554.

[40] I.-T. E.800, "Definitions of terms related to quality of service," International Telecommunication Union, Geneva, Recommendation E.800, Sept. 2008.

[41] D. Mustill and P. J. Willis, "Delivering QoS in the next generation network a standards perspective," *BT Technology Journal*, vol. 23, pp. 48–60, 2005.

[42] A. Meddeb, "Internet QoS: Pieces of the puzzle," *Communications Magazine, IEEE*, vol. 48, no. 1, pp. 86 –94, january 2010.

[43] K. Kilkki, "Quality of Experience in Communications Ecosystem," *Journal of Universal Computer Science*, vol. 14, no. 5, pp. 615–624, 2008.

[44] C. Aurrecoechea, A. T. Campbell, and L. Hauw, "A survey of QoS architectures," *Multimedia Syst.*, vol. 6, no. 3, pp. 138–151, 1998.

[45] S. Shenker, C. Partridge, and R. Guerin, "Specification of Guaranteed Quality of Service," RFC 2212 (Proposed Standard), Internet Engineering Task Force, Sep. 1997. [Online]. Available: http://www.ietf.org/rfc/rfc2212.txt

[46] J. Wroclawski, "Specification of the Controlled-Load Network Element Service," RFC 2211 (Proposed Standard), Internet Engineering Task Force, Sep. 1997. [Online]. Available: http://www.ietf.org/rfc/rfc2211.txt

[47] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, "Assured Forwarding PHB Group," RFC 2597 (Proposed Standard), Internet Engineering Task Force, Jun. 1999, updated by RFC 3260. [Online]. Available: http://www.ietf.org/rfc/rfc2597.txt

[48] B. Davie, A. Charny, J. Bennet, K. Benson, J. L. Boudec, W. Courtney, S. Davari, V. Firoiu, and D. Stiliadis, "An Expedited Forwarding PHB (Per-Hop Behavior)," RFC 3246 (Proposed Standard), Internet Engineering Task Force, Mar. 2002. [Online]. Available: http://www.ietf.org/rfc/rfc3246.txt

[49] A. Charny, J. Bennet, K. Benson, J. Boudec, A. Chiu, W. Courtney, S. Davari, V. Firoiu, C. Kalmanek, and K. Ramakrishnan, "Supplemental Information for the New Definition of the EF PHB (Expedited Forwarding Per-Hop Behavior)," RFC 3247 (Informational), Internet Engineering Task Force, March 2002. [Online]. Available: http://www.ietf.org/rfc/rfc3247.txt

[50] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification," RFC 2205 (Proposed Standard), Internet Engineering Task Force, Sep. 1997, updated by RFCs 2750, 3936, 4495. [Online]. Available: http://www.ietf.org/rfc/rfc2205.txt

[51] J. Crowcroft, S. Hand, R. Mortier, T. Roscoe, and A. Warfield, "QoS's downfall: at the bottom, or not at all!" in *RIPQoS '03: Proceedings of the ACM SIGCOMM workshop on Revisiting IP QoS*. New York, NY, USA: ACM, 2003, pp. 109–114.

[52] S. Tripathi, N. Droux, T. Srinivasan, K. Belgaied, and V. Iyer, "Crossbow: a vertically integrated QoS stack," in *WREN '09: Proceedings of the 1st ACM workshop on Research on enterprise networking*. New York, NY, USA: ACM, 2009, pp. 45–54.

[53] K. Nichols, S. Blake, F. Baker, and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers," RFC 2474 (Proposed Standard), Internet Engineering Task Force, Dec. 1998, updated by RFCs 3168, 3260. [Online]. Available: http://www.ietf.org/rfc/rfc2474.txt

[54] K. Nichols, V. Jacobson, and L. Zhang, “A Two-bit Differentiated Services Architecture for the Internet,” RFC 2638 (Informational), Internet Engineering Task Force, Jul. 1999. [Online]. Available: http://www.ietf.org/rfc/rfc2638.txt

[55] B. Braden, D. Clark, J. Crowcroft, B. Davie, S. Deering, D. Estrin, S. Floyd, V. Jacobson, G. Minshall, C. Partridge, L. Peterson, K. Ramakrishnan, S. Shenker, J. Wroclawski, and L. Zhang, “Recommendations on Queue Management and Congestion Avoidance in the Internet,” RFC 2309 (Informational), Internet Engineering Task Force, Apr. 1998. [Online]. Available: http://www.ietf.org/rfc/rfc2309.txt

[56] L. Delgrossi and L. Berger, “Internet Stream Protocol Version 2 (ST2) Protocol Specification - Version ST2+,” RFC 1819 (Experimental), Internet Engineering Task Force, Aug. 1995. [Online]. Available: http://www.ietf.org/rfc/rfc1819.txt

[57] R. Hancock, G. Karagiannis, J. Loughney, and S. V. den Bosch, “Next Steps in Signaling (NSIS): Framework,” RFC 4080 (Informational), Internet Engineering Task Force, Jun. 2005. [Online]. Available: http://www.ietf.org/rfc/rfc4080.txt

[58] J. Manner, G. Karagianni, and A. McDonald, “NSIS Signaling Layer Protocol (NSLP) for Quality-of-Service Signaling,” RFC 5974 (Proposed Standard), Network Working Group, October 2010.

[59] M. Brunner, “Requirements for Signaling Protocols,” RFC 3726 (Informational), Internet Engineering Task Force, Apr. 2004. [Online]. Available: http://www.ietf.org/rfc/rfc3726.txt

[60] I. Más and G. Karlsson, “A model for endpoint admission control based on packet loss,” in *Proceedings of the 7th international IFIP-TC6 networking conference on AdHoc and sensor networks, wireless networks, next generation internet*, ser. NETWORKING’08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 506–517. [Online]. Available: http://portal.acm.org/citation.cfm?id=1792514.1792571

[61] M. Grossglauser and D. N. C. Tse, “A time-scale decomposition approach to measurement-based admission control,” in *Proc. IEEE Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies INFOCOM ’99*, vol. 3, 1999, pp. 1539–1547.

[62] S. Jamin, S. J. Shenker, and P. B. Danzig, "Comparison of measurement-based admission control algorithms for controlled-load service," in *Proc. IEEE Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies INFOCOM '97*, vol. 3, 1997, pp. 973–980.

[63] Y. Jiang, P. J. Emstad, V. Nicola, and A. Nevin, "Measurement-based admission control: A revisit," *Seventeenth Nordic Teletraffic Seminar*, pp. 25–27, 2004.

[64] S. Jamin, P. B. Danzig, S. J. Shenker, and L. Zhang, "A measurement-based admission control algorithm for integrated service packet networks," *IEEE/ACM Trans. Netw.*, vol. 5, no. 1, pp. 56–70, 1997.

[65] H. El Allali and G. Heijenk, "Resource management in IP-based radio access networks," February 2001, pp. 19–24.

[66] J. Kim and A. Jamalipour, "Traffic management and QoS provisioning in future wireless IP networks," *IEEE Personal Communications*, vol. 8, no. 5, pp. 46–55, 2001.

[67] D. K. S. Seung Yeob Nam, Sunggon Kim, "Measurement-based admission control at edge routers," *IEEE/ACM TRANSACTIONS ON NETWORKING*, vol. 16, no. 2, April 2008.

[68] L. Breslau, S. Jamin, and S. Shenker, "Comments on the performance of measurement-based admission control algorithms," in *Proc. IEEE Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies - INFOCOM 2000*, vol. 3, 2000, pp. 1233–1242.

[69] R. J. Gibbens and F. P. Kellya, "Measurement-based connection admission control," *In 15th International Teletraffic Congress Proceedings*, June 1997.

[70] E. Knightly and N. Shroff, "Admission control for statistical QoS: theory and practice," *Network, IEEE*, vol. 13, no. 2, pp. 20 –29, Mar/Apr 1999.

[71] E. Tragos, G. Tsiropoulos, G. Karetsos, and S. Kyriazakos, "Admission control for QoS support in heterogeneous 4G wireless networks," *Network, IEEE*, vol. 22, no. 3, pp. 30 –37, may-june 2008.

[72] S. Floyd, "Comments on Measurement-based Admissions Control for Controlled-Load Services," *Lawrence Berkeley Laboratory*, July 1996, http://www.icir.org/floyd/admit.html (Last visited 06/2010). [Online]. Available: http://www.icir.org/floyd/admit.html (Last visited 06/2010)

[73] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *Selected Areas in Communications, IEEE Journal on*, vol. 9, no. 7, pp. 968 –981, September 1991.

[74] R. Gibbens, F. Kelly, and P. Key, "A decision-theoretic approach to call admission control in ATM networks," *IEEE Journal on Selected Areas in Communications,*, vol. 13, no. 6, pp. 1101 –1114, Aug 1995.

[75] P. Guang, "Admission control algorithms: A survey," University of California, White paper, 2001.

[76] C. Partridge, "A proposed flow specification," RFC 1363 (Informational), Internet Engineering Task Force, Sep. 1992. [Online]. Available: http://www.ietf.org/rfc/rfc1363.txt

[77] M. S. P. F. R. Lima, "A distributed admission control model for class-based ip networks," Ph.D. dissertation, University of Minho, September 2005.

[78] C. Casetti, J. Kurose, and D. Towsley, "An Adaptive Algorithm for Measurement-based Admission Control in Integrated Services Packet Networks," in *Proceedings of the International Workshop on Protocols for High-Speed Networks*, October 1996, pp. 13–28.

[79] S. Georgoulas, P. Trimintzios, G. Pavlou, and K. Ho, "An integrated bandwidth allocation and admission control framework for the support of heterogeneous real-time traffic in class-based IP networks," *Comput. Commun.*, vol. 31, no. 1, pp. 129–152, 2008.

[80] IEEE, "IEEE 802.1q: VLAN," Tech. Rep., November 2008, http://www.ieee802.org/1/pages/802.1Q.html (last visited 05/2011). [Online]. Available: http://www.ieee802.org/1/pages/802.1Q.html

[81] I. of Electrical Electronics Engineers, “IEEE Standard for Local and Metropolitan Area Networks,” Tech. Rep., 2009, http://standards.ieee.org/getieee802/download/802.16-2009.pdf (last visited 05/2011). [Online]. Available: http://standards.ieee.org/getieee802/download/802.16-2009.pdf

[82] J. Kempf, “Problem Description: Reasons For Performing Context Transfers Between Nodes in an IP Access Network,” RFC 3374 (Informational), Internet Engineering Task Force, Sep. 2002. [Online]. Available: http://www.ietf.org/rfc/rfc3374.txt

[83] J. Loughney, M. Nakhjiri, C. Perkins, and R. Koodli, “Context Transfer Protocol (CXTP),” RFC 4067 (Experimental), Internet Engineering Task Force, Jul. 2005. [Online]. Available: http://www.ietf.org/rfc/rfc4067.txt

[84] Y. Bernet, P. Ford, R. Yavatkar, F. Baker, L. Zhang, M. Speer, R. Braden, B. Davie, J. Wroclawski, and E. Felstaine, “A Framework for Integrated Services Operation over Diffserv Networks,” RFC 2998 (Informational), Internet Engineering Task Force, Nov. 2000. [Online]. Available: http://www.ietf.org/rfc/rfc2998.txt

[85] F. Baker, C. Iturralde, F. L. Faucheur, and B. Davie, “Aggregation of RSVP for IPv4 and IPv6 Reservations,” RFC 3175 (Proposed Standard), Internet Engineering Task Force, Sep. 2001, updated by RFC 5350. [Online]. Available: http://www.ietf.org/rfc/rfc3175.txt

[86] I. Stoica and H. Zhang, “Providing guaranteed services without per flow management,” *SIGCOMM Comput. Commun. Rev.*, vol. 29, no. 4, pp. 81–94, 1999.

[87] E. Addisu and J. Yuming, “Flow Aggregation Using Dynamic Packet State, booktitle = Networked Services and Applications - Engineering, Control and Management,” ser. Lecture Notes in Computer Science, F. Aagesen and S. Knapskog, Eds. Springer Berlin / Heidelberg, 2010, vol. 6164, pp. 263–265.

[88] L. Zhang, “Virtual clock: a new traffic control algorithm for packet switching networks,” *SIGCOMM Comput. Commun. Rev.*, vol. 20, no. 4, pp. 19–29, 1990.

[89] T. Braun and G. Stattenberger, "Providing differentiated services to Mobile IP users," in *Proc. 26th Annual IEEE Conference on Local Computer Networks LCN 2001*, November 2001, pp. 89–90.

[90] T. Braun, L. Ru, and G. Stattenberger, "An AAA architecture extension for providing differentiated services to mobile IP users," in *Proc. Sixth IEEE Symposium on Computers and Communications*, 3–5 July 2001, pp. 472–478.

[91] G. Stattenberger and T. Braun, "QoS Provisioning for Mobile IP users," *1st Int'l.Wksp. Svcs. and Apps. in Wireless Public Infrastructure*, July 2001.

[92] A. Bader, L. Westberg, G. Karagiannis, C. Kapplerand, and T. Phelan, "RMD-QOSM: The NSIS Quality-of-Service Model for Resource Management in Diffserv," RFC 5977 (Experimental), Internet Engineering Task Force, October 2010.

[93] *IETF NSIS Working Group - Next Steps in Signaling (NSIS)*, http://www.ietf.org/dyn/wg/charter/nsis-charter.html (Last visited 05/2011). [Online]. Available: http://www.ietf.org/dyn/wg/charter/nsis-charter.html (Last visited 05/2011)

[94] H. Schulzrinne and R. Hancock, "GIST: General Internet Signalling Transport," RFC 5971 (Experimental), Internet Engineering Task Force, 2010.

[95] G. Ash, A. Bader, C. Kappler, and D. Oran, "QSPEC Template for the Quality-of-Service NSIS Signaling Layer Protocol (NSLP)," RFC 5975 (Experimental), Internet Engineering Task Force, 2010.

[96] J. Wroclawski, *The Use of RSVP with IETF Integrated Services*, RFC 2210 (Proposed Standard), Internet Engineering Task Force Std. 2210, Sep. 1997. [Online]. Available: http://www.ietf.org/rfc/rfc2210.txt

[97] T. Sanda, X. Fu, S. Jeong, J. Manner, and H. Tschofenig, "NSIS Protocol operation in Mobile Environments," RFC 5980 (Proposed Standard), Network Working Group, March 2011. [Online]. Available: http://www.rfc-editor.org/rfc/rfc5980.txt

[98] A. Terzis, J. Krawczyk, J. Wroclawski, and L. Zhang, "RSVP Operation Over IP Tunnels," RFC 2746 (Proposed Standard), Internet Engineering Task Force, Jan. 2000. [Online]. Available: http://www.ietf.org/rfc/rfc2746.txt

[99] C.-C. Tseng, G.-C. Lee, R.-S. Liu, and T.-P. Wang, "HMRSVP: A Hierarchical Mobile RSVP Protocol," *Wireless Networks*, vol. 9, pp. 95–102, March 2003.

[100] A.-H. A. Hashim, W. H. Hassan, S. Syed, and M. A. F. Mudathir, "A HMRSVP Approach to Support QoS Challenges in Mobile Environment," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 9, no. 1, pp. 69–73, 2009.

[101] X. Fu, H. Karl, and C. Kappler, "QoS-Conditionalized Handoff for Mobile IPv6," in *NETWORKING 2002: Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications*, ser. Lecture Notes in Computer Science, E. Gregori, M. Conti, A. Campbell, G. Omidyar, and M. Zukerman, Eds. Springer Berlin / Heidelberg, 2006, vol. 2345, pp. 721–730.

[102] R. Yavatkar, D. Pendarakis, and R. Guerin, "A Framework for Policy-based Admission Control," RFC 2753 (Informational), Internet Engineering Task Force, Jan. 2000. [Online]. Available: http://www.ietf.org/rfc/rfc2753.txt

[103] D. Durham, J. Boyle, R. Cohen, S. Herzog, R. Rajan, and A. Sastry, "The COPS (Common Open Policy Service) Protocol," RFC 2748 (Proposed Standard), Internet Engineering Task Force, Jan. 2000, updated by RFC 4261. [Online]. Available: http://www.ietf.org/rfc/rfc2748.txt

[104] C. Liu, Y. Liu, D. Qian, and M. Li, "An Approach of End-to-End DiffServ/MPLS QoS Context Transfer in HMIPv6 Net," in *Autonomous Decentralized Systems, 2007. ISADS '07. Eighth International Symposium on*, march 2007, pp. 245 –254.

[105] R. Mameli and S. Salsano, "Use of COPS for Intserv operations over Diffserv: architectural issues, protocol design and test-bed implementation," in *IEEE International Conference on Communications, 2001 (ICC 2001).*, vol. 10, 2001, pp. 3265–3270.

[106] L. Westberg, A. Eriksson, G. Karagiannis, G. Heijenk, Rexhepi, and D. Partain, "Dynamic QoS Management In Differentiated Services Using Bandwidth Brokers, RSVP Aggregation And Load Control Protocols," Patent EP1 312 226, May, 2009. [Online]. Available: http://www.freepatentsonline.com/EP1312226B1.html

[107] T. Robles, A. Kadelka, H. Velayos, A. Lappetelainen, A. Kassler, H. Li, D. Mandato, J. Ojala, and B. Wegmann, “QoS support for an all IP system beyond 3G,” *Communications Magazine, IEEE*, vol. 39, no. 8, pp. 64 –72, Aug. 2001.

[108] *IST - Moby Dick Project - Mobility and Differentiated Services in a Future IP Network*, http://www.tik.ee.ethz.ch/ mobydick/ (last visited 05/2011).

[109] H. Einsiedler, R. Aguiar, and et al., “The Moby Dick Project: A Mobile Heterogeneous ALL-IP Architecture,” in *In: Proc. of the ATAMS 2001. Krakow*, 2001, pp. 17–20.

[110] N. Taylor, P. Robertson, B. Farshchian, K. Doolin, I. Roussaki, L. Marshall, R. Mullins, S. Dru andsedow, and K. Dolinar, “Pervasive Computing in Daidalos,” *Pervasive Computing, IEEE*, vol. 10, no. 1, pp. 74 –81, jan.-march 2011.

[111] S. Sargento, R. Prior, and et al., “End-to-end QoS Architecture for 4G Scenarios,” in *Proceedings of the 14th Wireless and Mobile Communications Summit*, Dresden, Germany, 2005.

[112] G. Carneiro, C. Garcia, and et al., “The DAIDALOS Architecture for QoS over Heterogeneous Wireless Networks,” in *Proceedings of the 14th Wireless and Mobile Communications Summit*, Dresden, Germany, 2005.

[113] E. Mingozzi, G. Stea, M. Callejo-Rodriguez, J. Enriquez-Gabeiras, G. G. de Blas, F. Ramon-Salquero, W. Burakowski, A. Beben, J. Sliwinski, H. Tarasiuk, O. Dugeon, M. Diaz, L. Baresse, and E. Monteiro, “EuQoS: End-to-End Quality of Service over Heterogeneous Networks,” *Computer Communications*, vol. 32, no. 12, pp. 1355 – 1370, 2009, special Issue of Computer Communications on Heterogeneous Networking for Quality, Reliability, Security, and Robustness - Part II. [Online]. Available: http://www.sciencedirect.com/science/article/B6TYP-4V88FPG-3/2/5d98aa7738e49c583d2b4aac62c1cd67

[114] R. Ferrus, X. Reves, A. Umbert, and F. Casadevall, “Real-time emulation of RRM strategies for UMTS bearer services,” in *Vehicular Technology Conference, 2002. Proceedings. VTC 2002-Fall. 2002 IEEE 56th*, vol. 2, 2002, pp. 955 – 959 vol.2.

[115] *QBone Signaling Workgroup*, July 2002, http://qbone.internet2.edu/bb/ (Last visited 05/2011).

[116] S. Sohail and S. Jha, "The Survey of Bandwidth Broker," University of New South Wales, Sydney, Australia, Technical Report UNSQ-CSE-TR-0206, May 2002.

[117] Z. Wan, X. Pan, and L. Ping, "QoS Provisioning in an Enhanced FMIPv6 Architecture," in *International Conference on Computational Science and its Applications (ICCSA 2005)*, vol. 3481, May 2005, pp. 704–713. [Online]. Available: http://www.springerlink.com/content/u2xxkcenqdamf4p5

[118] A. Belhoul, Y. A. Sekercioglu, and N. Mani, "Mobility-Aware RSVP: A framework for improving the performance of multimedia services over wireless IP-based mobile networks," *Computer Communications*, vol. 32, no. 4, pp. 569–582, March 2009. [Online]. Available: http://www.sciencedirect.com/science/article/B6TYP-4V2NP2F-1/2/81e2b340a826d1e4555b79ec01ac6256

[119] H. Soliman, C. Castelluccia, K. E. Malki, and L. Bellier, "Hierarchical Mobile IPv6 Mobility Management (HMIPv6)," RFC 4140 (Experimental), Internet Engineering Task Force, Aug. 2005, obsoleted by RFC 5380. [Online]. Available: http://www.ietf.org/rfc/rfc4140.txt

[120] R. Hsieh, A. Seneviratne, H. Soliman, and K. El-Malki, "Performance analysis on hierarchical Mobile IPv6 with fast-handoff over end-to-end TCP," in *Global Telecommunications Conference, 2002. GLOBECOM '02.*, vol. 3, 2002, pp. 2488–2492.

[121] N. Jordan and P. Reichl, "A Fast Handover System Evaluation in an All-IPv6 Mobility Management - Wireless Broadband Access based Hotspot Network Environment," in *International Conference on Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies, 2006. ICN/ICONS/MCL 2006.*, 2006, pp. 122–122.

[122] Z. Zhang, J. Fang, W. Wang, and S. Zhang, "Performance Comparison of Mobile IPv6 and Its Extensions," in *International Conference on Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007*, 2007, pp. 1805–1808.

[123] X. Pérez-Costa, M. Torrent-Moreno, and H. Hartenstein, "A performance comparison of Mobile IPv6, Hierarchical Mobile IPv6, fast handovers for Mobile IPv6 and their

combination," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 7, no. 4, pp. 5–19, 2003.

[124] V. Padmanabhan, L. Qiu, and H. Wang, "Server-based inference of internet link lossiness," in *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, vol. 1, 2003, pp. 145–155.

[125] S. Georgoulas, G. Pavlou, P. Trimintzios, and K.-H. Ho, "Admission Control for Inter-domain Real-Time Traffic Originating from Differentiated Services Stub Domains," in *WWIC '07: Proceedings of the 5th international conference on Wired/Wireless Internet Communications*. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 115–128.

[126] J. Postel and J. Reynolds, "File Transfer Protocol," RFC 959 (Standard), Internet Engineering Task Force, Oct. 1985, updated by RFCs 2228, 2640, 2773, 3659. [Online]. Available: http://www.ietf.org/rfc/rfc959.txt

[127] F. L. Faucheur, J. Manner, D. Wing, and A. Guillou, "Resource Reservation Protocol (RSVP) Proxy Approaches," RFC 5945 (Proposed Standard), Internet Engineering Task Force, Oct. 2010. [Online]. Available: http://www.rfc-editor.org/rfc/rfc5945.txt

[128] R. Hsieh and A. Seneviratne, "A comparison of mechanisms for improving mobile IP handoff latency for end-to-end TCP," in *MobiCom '03: Proceedings of the 9th annual international conference on Mobile computing and networking*. New York, NY, USA: ACM, 2003, pp. 29–41.

[129] K. Fall and K. Varadhan, *The ns Manual*, UC Berkeley, LBL, USC/ISI, and Xerox PARC., October 2009. [Online]. Available: http://www.isi.edu/nsnam/ns/ns-documentation

[130] P. Pieda, J. Etheridge, M. Baines, and F. Shallwani, *A Network Simulator, Differentiated Services Implementation*, Open IP, Nortel Networks, 2000. [Online]. Available: Open IP, Nortel Networks

[131] C. E. Perkins, "Mobile networking in the Internet," *Mob. Netw. Appl.*, vol. 3, no. 4, pp. 319–334, 1998.

[132] NIST, *The Network Simulator NS-2 NIST add-on (IEEE 802.21 model)*, National Institute of Standards and Technology, January 2007, http://w3.antd.nist.gov/seamlessandsecure/ (last visited 05/2011). [Online]. Available: http://w3.antd.nist.gov/seamlessandsecure/

[133] D. D. Clark and W. Fang, "Explicit allocation of best-effort packet delivery service," *IEEE/ACM Trans. Netw.*, vol. 6, no. 4, pp. 362–373, 1998.

[134] N. S. W. Fang and B. Nandy, "A Time Sliding Window Three Colour Marker (TSWTCM)," RFC 2859, Internet Engineering Task Force, June 2000. [Online]. Available: http://www.ietf.org/rfc/rfc2859.txt

[135] D. Qiao and S. Choi, "Goodput enhancement of IEEE 802.11a wireless LAN via link adaptation," in *IEEE International Conference on Communications, 2001. ICC 2001.*, vol. 7, 2001, pp. 1995–2000.

[136] D. Qiao, S. Choi, and K. G. Shin, "Goodput analysis and link adaptation for IEEE 802.11a wireless LANs," *IEEE Transactions on Mobile Computing.*, vol. 1, no. 4, pp. 278–292, December 2002.

[137] V. Subramanian, K. K. Ramakrishnan, S. Kalyanaraman, and L. Ji, "Impact of interference and capture effects in 802.11 wireless networks on TCP," in *WitMeMo '06: Proceedings of the second international workshop on Wireless traffic measurements and modeling.* New York, NY, USA: ACM, 2006, p. 2.

[138] R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2004, iSBN: 3-900051-07-0. [Online]. Available: http://www.R-project.org

[139] T. D. Dang, B. Sonkoly, and S. Molnar, "Fractal analysis and modeling of VoIP traffic," in *Telecommunications Network Strategy and Planning Symposium. NETWORKS 2004, 11th International*, june 2004, pp. 123 – 130.

[140] D. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE Journal on Selected Areas in Communications,*, vol. 24, no. 8, pp. 1439–1451, August 2006.

[141] D. P. Palomar and M. Chiang, "Alternative decompositions for distributed maximization of network utility: Framework and applications," in *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*, April 2006, pp. 1–13.

[142] S. Shakkottai and R. Srikant, "Network optimization and control," *Found. Trends Netw.*, vol. 2, no. 3, pp. 271–379, 2007.