A MULTI-PANEL QOS CONTROL COMMUNICATIONS FRAMEWORK IN HETEROGENEOUS NETWORKS

YAN CHENG

A THESIS

IN

THE DEPARTMENT

OF

COMPUTER SCIENCE AND SOFTWARE ENGINEERING

PRESENTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

CONCORDIA UNIVERSITY

MONTRÉAL, QUÉBEC, CANADA

DECEMBER 2008

Your file Votre référence
ISBN: 978-0-494-63364-9
Our file Notre référence
ISBN: 978-0-494-63364-9

Abstract

A Multi-panel QoS Control Communications Framework in Heterogeneous Networks

Yan Cheng, Ph.D.

Concordia University, 2009

The Mobile Communication environment has changed dramatically from providing pure voice service exclusive to cellular users to providing multimedia services to users with access to an all-IP infrastructure via heterogeneous access technologies. IP in next generation networks will provide efficient and cost-effective interworking between different systems. Therefore, IP QoS provisioning will play a key role in satisfying mobile users' expectations on seamless access to multimedia services of desired quality anywhere and anytime. However, heterogeneity of the systems and diversity of the services challenge IP QoS provisioning schemes in multiple aspects. It is one of the most critical challenges to maintain QoS level to mobile users when handoff occurs. QoS control communications must interwork with various management entities to ensure the minimization of service disruption during handoff.

In this thesis, we propose a generic multi-panel QoS control communications framework to provide a guideline for optimizing control communications between various functionalities and entities. The framework is an effective platform to analyze control communications from two perspectives, intra-panel and inter-panel. Specifically, we propose a novel MIPv6 handoff scheme in the QoS panel to smooth the procedure of QoS updates during a MIPv6 handoff. Vertical handoff is the most challenging event in the framework. The two key problems of QoS control communications are QoS domain switch and automatic service adjustment between different access networks. We propose a QoS control architecture in the QoS panel to address

these issues. Both analytical results and simulation results have demonstrated that the proposed schemes for different events can promote the efficiency of intra-panel control communications in the QoS Panel. The direct benefits are greatly-reduced QoS configuration delay during handoff and smooth end-to-end QoS provisioning in terms of different measurements.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Dr. J. W. Atwood. This thesis could not have been possible without his encouragement, great insight, sound advice and remarkable patience throughout the thesis-writing procedure. His dedication to research inspires my life.

I would like to thank Dr. Samuel Pierre for being the external examiner of my thesis. The valuable input from my committee members, Dr. Anjali Agarwal, Dr. Peter Grogono, and Dr. Rajagopalan Jayakumar, has enlightened my research in different ways.

I would like to thank Ms. Halina Monkiewicz and Ms. Pauline Dubois for giving me assistance and encouragement during my long-term course of study.

Lastly, I am so indebted to my family. I would like to thank my parents for their unconditional love and support. I dedicate this thesis to my husband, Yunyu, and my dear son, Leyang, whose smiles have led me through all the difficulties.

Contents

List of Figures

List of Tables

List of Acronyms

2G	Second Generation mobile systems
3G	Third Generation mobile systems
3GPP	The Third Generation Partnership Project
3GPP2	The Third Generation Partnership Project 2
AAA	Authentication, Authorization and Accounting
AAAC	AAA and Charging
AC	Access Category
AF	Assured Forwarding
AN	Access Network
ANG	Access Network Gateway
AR	Access Router
ATM	Asynchronous Transfer Mode
AU	Agent Update
BB	Bandwidth Broker
BS	Base Station
BU	Binding Update
CBR	Constant Bit Rate
CDMA	Code Division Multiple Access

CN	Core Network
CN	Corresponding Node
CoA	Care-of-Address
CoS	Class of Service
DiffServ	Differentiated Services
DSCP	Differentiated Services Code Point
DSL	Digital Subscriber Line
EDCA	Enhanced Distribution Coordination Access
EDGE	Enhanced Data rates for GSM Evolution
EF	Expedited Forwarding
ER	Edge Router
ETSI	European Telecommunications Standards Institute
FA	Foreign Agent
FMIPv6	Fast Mobile IPv6
FT	Flow Transparency
GGSN	Gateway GPRS Support Node
GIST	Generic Internet Signaling Transport
GPRS	General Packet Radio Service
GSM	Global System for Mobile communications
HA	Home Address
HC	Hybrid Coordinator
HCCA	HCF Controlled Channel Access
HCF	Hybrid Coordination Function
HMIPv6	Hierarchical Mobile IPv6

HSCSD	High-Speed Circuit-Switched Data
IETF	Internet Engineering Task Force
IMT-2000	International Mobile Telecommunications 2000
IntServ	Integrated Services
ISP	Internet Service Provider
MA	Mobility management Agent
MAC	Media Access Control
MAD	Maximum Allowed Delay
MIPv4	Mobile IPv4
MIPv6	Mobile IPv6
MN	Mobile Node
MPLS	MultiProtocol Label Switching
NCoA	New CoA of MN
NCR	Nearest Common Router
NSIS	Next Steps in Signaling
NSLP	NSIS Signaling Layer Protocol
NTLP	NSIS Transport Layer Protocol
PCoA	Previous CoA of MN
PDP	Packet Data Protocol
PDP	Policy Decision Point
PEP	Policy Enforcement Point
PHB	Per-Hop forwarding Behavior
PPP	Point-to-Point Protocol
QA	QoS Agent

QAA	QA-Assisted scheme
QCD	QoS Configuration Delay
QH	QoS Handler
QoS	Quality of Service
QSB	QoS State Block
QSTA	QoS station
RAN	Radio Access Network
RSVP	Resource reSerVation Protocol
RTCP	Real-Time Transport Control Protocol
SDU	Session Data Unit
SIP	Session Initiation Protocol
SLA	Service Level Agreement
TXOP	Transmission Opportunity
UE	User Equipment
UMTS	Universal Mobile Telecommunications System
UTRAN	UMTS terrestrial RAN
WiMax	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network
WMAN	Wireless Metropolitan Area Network

Chapter 1

Introduction

The explosive demand for Internet access and packet data applications is changing the nature of future communications from circuit-switching-based architecture to packet-data-focused architecture. Concurrent voice, data, and multimedia applications also blur the lines between telecommunications and data services, and therefore drive future architectures to multi-service platforms. Two main features will strongly characterize the future telecommunications environment: heterogeneity of telecommunications platforms and customization of multimedia services.

On the other hand, all wireless data-applications are IP-based. IP provides a unified platform to support the major features of future generation networks, i.e., supporting different access technologies and providing diverse network services to users. Therefore, we vision an all-IP wireless network in the future. The focus of our research is IP QoS provisioning in wireless networks. When QoS is mentioned in this thesis, it always refers to IP QoS unless specified otherwise.

1.1 Motivation

QoS is an important issue in IP-based future generation wireless networks. End users expect to receive real-time application data of high quality while holding mobile terminals and roaming all over the world. Yet the characteristics of wireless communication

channels limit the bandwidth and signal quality on many occasions.

Mobility causes another critical issue regarding to maintaining high quality service to mobile users, namely handoff performance. Moving from one network region to another, the mobile user inevitably suffers from a short period of service degradation or even disruption.

Quoting from RFC 2990 [Hus00], "It is extremely improbable that any single form of service differentiation technology will be rolled out across the Internet and across all enterprise networks". This precisely describes the diverse nature of QoS deployment. Moreover, the integration of wireless networks with the legacy Internet has imposed new challenges in terms of heterogeneous service responses from the network providers. To expect all applications, mobile and fixed end systems, routers and base stations, network policies, and inter-provider arrangements to coalesce into a homogeneous service environment that can support an agreed range of services is unrealistic. It is more practical to expect a number of small scale deployments of service differentiation mechanisms and put efforts on bridging these environments together in some way.

IP-based end-to-end QoS was initially developed for the legacy Internet. It must be enhanced to meet new challenges in mobile/wireless environments. There have been numerous efforts to this purpose. However, we have not seen a control communications framework designed for heterogeneous QoS models and wireless technologies, which is the essence of future telecommunication environments.

Based on the observation of the above situations, we argue that an adequate set of signaling mechanisms to enable effective interworking and communications among various mobility protocols, QoS models, and wireless services is extremely important and essential in the future generation networks.

1.2 Goal of the Thesis

The control communications framework presented in this thesis aims at enabling optimizations of the complicated control communications between various functionalities and entities, so that the overall performance of QoS provisioning in different scenarios can be improved with a guide. Our way is to group all the control communications into multiple panels. Currently we have User Panel, Access Panel, and QoS Panel. Inter-panel control communications and intra-panel communications are analyzed according to each target scenario. Hence, optimizations can be conducted based on the analysis results.

Limited by the time and resources available for the thesis, the QoS Panel is the major focus of our research. Enabling fast QoS updates during handoff is the most essential goal in the QoS Panel, because this feature is highly desired by both end users and service providers.

Heterogeneity of wireless technologies fosters roaming between different types of wireless networks, which is referred to as vertical handoff. End users can select the most suitable network for their high demand data sessions and opt to another network for economic reasons when low capacity is sufficient. Under this circumstance, automatic service adjustment is the most desired feature of the QoS Panel during vertical handoff.

1.3 Major Innovations of this Work

We present our innovations in this thesis in the following aspects.

- **QoS control communications are generalized into multiple panels, shielding the diversity of QoS models and wireless technologies.** The multi-panel control communications framework is independent of any QoS model, mobility model, or wireless technology. The functionalities of the agents in the

control panels are triggered by common events, such as session initiation, vertical handoff, and session termination. Therefore, heterogeneous QoS models can be easily accomodated into the control communication framework. Furthermore, the analysis for performance optimization is suitable for any QoS model that fits into the framework.

This is the most significant contribution of this thesis because such a generalization of control communications has not been seen in the literature. The other two contributions are instantiations of the framework. Following the requirements and the design of the framework, they provide novel approaches to reduce QoS update latency and adjust service provisioning during different scenarios of handoff.

- **Carrying QoS information via mobility management messages enables fast QoS updates.** We propose the concept of critical information and critical node for QoS updates during handoff. Fast delivery of critical information to critical nodes is a key step to achieve fast re-configuration of QoS parameters and QoS states during handoff. We propose to utilize the mobility management messages to carry critical information, which normally includes the address information and the generalized QoS profiles of the mobile node. Assisted by QoS Agents located in critical nodes along the data path, QoS updates can be completed before the arrival of the packets destined to new locations of the mobile node in most cases. Thus, QoS configuration latency can be largely reduced and end-to-end performance can be maintained smoothly during handoff.

- **Two levels of service adjustment can meet different situations of vertical handoff.** We propose two levels of service adjustment during handoff, QoS granularity adjustment and service parameter adjustment. The diversity of networks and service models determines the uncertainty of QoS provisioning capability change and network capacity change during vertical handoff. QoS granularity adjustment meets the challenges of QoS provisioning capability change,

and service parameter adjustment meets the challenges of network capacity change. In both situations adjustment is automatically made, with assistance of the handoff scheme and the novel service adjustment algorithm proposed in Chapter 6.

1.4 Thesis Organization

The layout of the thesis is displayed in the paragraphs below.

Chapter One presents the general background, motivation, and goal of our research. Major innovations of the thesis are also presented.

Chapter Two introduces the research literature. Multiple research areas are involved as the background of our research, given the complex nature of control communications in heterogeneous network environments. Specifically, we look into state-of-the-art of QoS models, wireless QoS technologies and mobility management. Moreover, we discuss various efforts on QoS provisioning in wireless/mobile networks.

Chapter Three gives a comprehensive analysis of the challenges faced by QoS provisioning in future networks. Based on the analysis, we propose the overall requirements and functional requirements of the control communications framework.

Chapter Four proposes the multi-panel control communications framework. Both intra-panel communications and inter-panel communications are introduced. The control events are analyzed for each major scenario.

Chapter Five proposes the QoS-Agent assisted handoff scheme, which is the essential part of the QoS Panel. The concept and roles of QoS Agent are introduced and the interworking of the QoS Agents and other entities in the QoS Panel is discussed. Extensive simulations have been done to evaluate the performance of the handoff scheme. The simulation results are presented and discussed in the last part of this chapter.

Chapter Six extends the QoS-Agent assisted handoff scheme to the vertical handoff scenario. In fact, a complete control communications architecture in the QoS

Panel is proposed for vertical handoff. The service adaptor and service mapper work with QoS Agents to enable fast service adjustment during vertical handoff. Analytical evaluation and simulation results demonstrate the performance of the proposed architecture.

Chapter Seven concludes the thesis by summarizing the major contributions and discussing the potential directions of future research.

Chapter 2

Literature

High-speed wireline, mobile, and satellite network segments will converge towards a universal personal communications platform in which users can freely and dynamically decide on the best access to multimedia services and customized configuration through suitable service level agreements. In this environment, where access points of multimedia customers will vary from session to session and network conditions will change during the same communication session, the presence of quality of service (QoS) support is crucial. QoS support will need to include mechanisms, services, and protocols for service differentiation, management, and continuous provision to multimedia traffic over unreliable, unpredictable, and intrinsically unstable wireless and mobile communications systems.

Without specifying any application domain, QoS can be chiefly used to "measure a specified set of performance attributes typically associated with a service" [Net]. QoS can be characterized by a small set of measurable parameters:

- *Service availability*: the reliability of the user's connection to the Internet service.
- *Delay*: also known as *latency*, refers to the interval between transmitting and receiving packets between two reference points.

- *Delay variation*: also called *jitter*, refers to the variation in time duration between all packets in a stream taking the same route.
- *Throughput*: the rate at which packets are transmitted in a network; can be expressed as an average or peak rate.
- *Packet loss rate*: the maximum rate at which packets can be discarded during transfer through a network; packet loss typically results from congestion or unstable wireless communication conditions.

The term *QoS provisioning* usually indicates the set of technologies for managing delay, jitter, and congestion events throughout a network via traffic policing and resource control.

In this chapter, we review the state-of-art of QoS provisioning in heterogeneous network environments. First of all, it is necessary to clarify the future generation networks that we frequently refer to in the thesis. Therefore, Section 2.1 presents a brief survey on the wireless/mobile systems with their own QoS support mechanisms.

2.1 Wireless/Mobile Systems

2.1.1 Future Generation Cellular Networks

The wide success of Second-Generation (2G) mobile systems prompted the development of Third-Generation (3G) mobile systems. Instead of carrying speech and low-bit-rate data as in 2G networks, 3G Systems are intended to provide a global mobility with a wide range of services including telephony, paging, messaging, Internet and broadband data. The 3rd Generation Partnership Project (3GPP), initiated by the European Telecommunications Standards Institute (ETSI) in December 1998, has been producing the standard for 3G systems based on the evolved GSM core network and the radio access technologies that they support. Meanwhile, the Third Generation Partnership Project 2 (3GPP2) has been comprising North American

and Asian interests developing global specifications for 3G networks. Both the 3GPP specification (i.e., UMTS) and 3GPP2 specification (i.e., cdma2000) are 3GPP standards under the International Telecommunication Union's generic name of IMT-2000 (International Mobile Telecommunications 2000).

The evolution towards 3G from 2G results in a diversity of 2.5G solutions, such as GPRS, EDGE, HSCSD, etc. Moreover, the Fourth Generation (4G) networks refer to a further expectation of telecommunications systems, which will have broader bandwidth, higher data rate, and smoother and quicker handoff and will focus on ensuring seamless services across a multitude of wireless systems and networks.

In summary, wireless communications systems in different generations will coexist for quite a long time. We call these systems future generation cellular networks, among which UMTS and cdma2000 are the two major standards that we are interested in for this thesis.

As specified in [3GP03], the UMTS network architecture consists of three interacting domains: Core Network (CN), UMTS terrestrial RAN (UTRAN), and User Equipment (UE). The Core Network is divided into circuit switched (CS) and packet switched (PS) domains. UTRAN provides an access platform for mobile terminals to all core networks and network services. It hides all radio-access-technology-dependent and mobility functions from the core network.

To provide data delivery with appropriate end-to-end QoS guarantees, the UMTS QoS architecture ([3GP05]) defines QoS parameters, traffic classes, the end-to-end data delivery model, and the mapping of end-to-end services to the services provided by the network elements of the UMTS. IP connectivity is provided through the PS domain as a pure network layer service between a UMTS mobile service and an Internet host.

To realize a certain network QoS, a Bearer Service with clearly defined characteristics and functionality is to be set up from the source to the destination of a service. The End-to-End Service used by the TE will be realized using a TE/MT Local Bearer

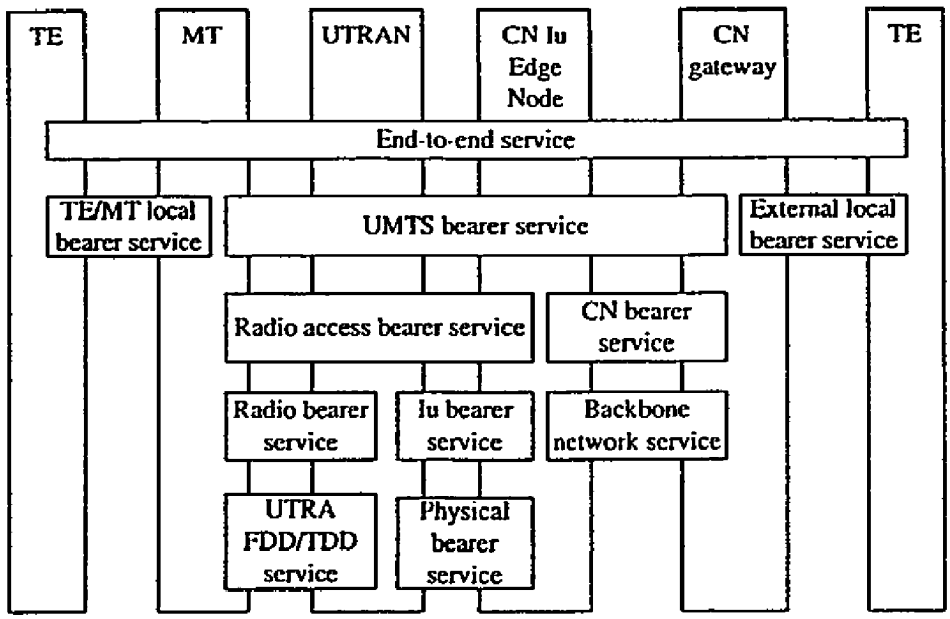


Figure 1: The layered architecture of end-to-end bearer service: UMTS

Service, a UMTS Bearer Service, and an External Bearer Service. This layered architecture, illustrated in Figure 1, requires the definition of QoS attributes for each bearer service. Such attributes serve to map the end-to-end QoS requirements to appropriate requirements for each bearer service used by a data connection. Among the most important attributes are traffic class, maximum bit rate, guaranteed bit rate, delivery order, maximum SDU size, SDU format information, SDU error ratio, transfer delay, traffic handling priority, allocation/retention priority, etc. Four traffic classes are currently defined in the UMTS QoS architecture, namely *conversational, streaming, interactive,* and *background.*

Figure 2 shows the layered architecture of the end-to-end user bearer service of cdma2000. The wireless bearer service is built on top of a PPP bearer service, which consists of the radio access bearer service and the RP bearer service. The Wireless IP Network Standard [3GP04] describes the QoS support at the point-to-point protocol (PPP) level and on the R-P interface. The same four traffic classes are supported in cdma2000 networks as in UMTS networks. The cdma2000 service options allow various voice and non-voice services to be defined and specified independently within

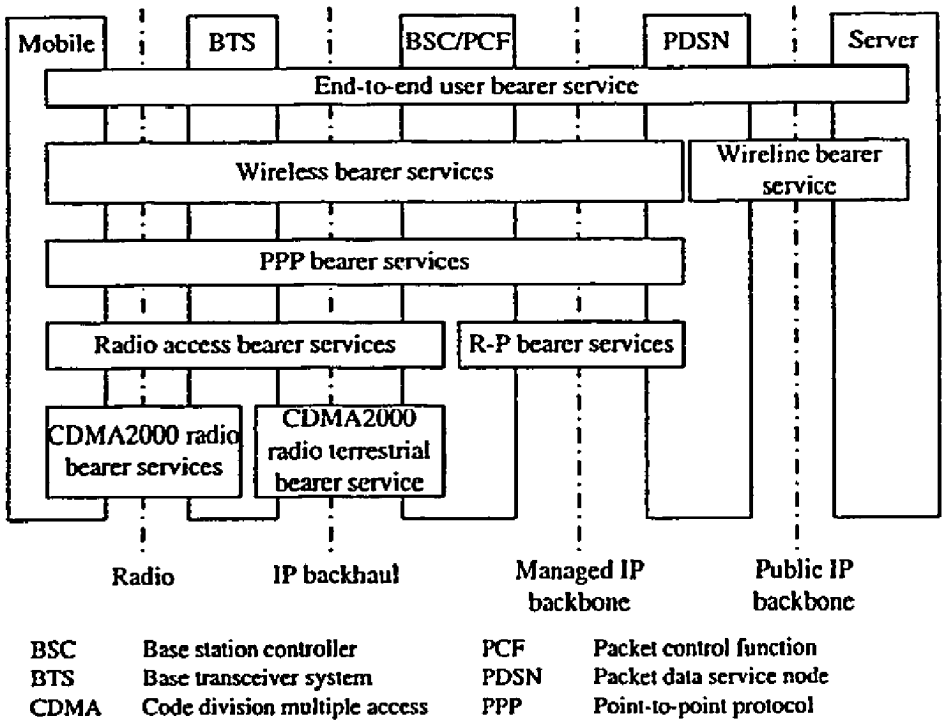


Figure 2: The layered architecture of end-to-end user bearer service: cdma2000

the confines of the physical layer and the multiplex sub-layer interface. The standard also proposes using DiffServ over the A10 interface.

For the air-link segment of the radio access bearer service, [3GP00] allows two QoS modes: assured mode and nonassured mode. A mobile host subscribed to the assured mode has the option of sending a set of QoS parameters, referred to as the *QoS BLOB*. A mobile host subscribed to the nonassured mode has the option of assigning different priority levels. The MAC layer multiplexes traffic from packet data service instances onto physical channels.

2.1.2 WLAN

The IEEE 802.11 Wireless Local Area Network (WLAN) has been gaining great popularity for data applications in university campuses, enterprise networks and hotspots since the baseline standard of it was approved in 1997. During the same time, the maximum bandwidth of the 802.11 system has been increased greatly, from 2Mbps

to 54Mbps. The legacy 802.11 standard was unable to provide adequate QoS for applications that require tight QoS support, such as VoIP and video transmission.

The 802.11e standard introduces Hybrid Coordination Function (HCF) to provide QoS enhancements to WLAN ([NRT04]). Enhanced Distribution Coordination Access (EDCA), the contention-based access mode of HCF, defines four access categories (AC) that support eight priorities for each QoS station (QSTA). Each frame arriving at the MAC with a priority is mapped into an AC and put into a queue. There is a set of EDCA parameters associated to a certain AC, in order to differentiate the channel access among traffic with different AC values. While EDCA is regarded as the major mode of 802.11 QoS support, 802.11e also defines a polling-based channel access, called HCF Controlled Channel Access (HCCA). A QoS-aware point coordinator, called a hybrid coordinator (HC), maintains a polling list and schedules the polled-TXOP for the QSTAs on the list. The TXOP (Transmission Opportunity) is defined as an interval of time when a particular QSTA has the right to initate transmissions onto the wireless medium. A QSTA can formally signal its desire of the TXOP to HC using TSPEC, which contains the set of parameters that define the characteristics and QoS expectations of a unidirectional traffic stream.

2.1.3 WiMax

Based on the IEEE 802.16 Wireless Metropolitan Area Network (WMAN) air interface standard ([IEE04]) and the IEEE 802.16e amendment ([IEE06]), WiMax (Worldwide Interoperability for Microwave Access) is a standards-based technology enabling the delivery of last mile wireless broadband access as an alternative to cable and DSL.

In contrast to WLAN, which uses contention access in the MAC layer, the 802.16 MAC uses a scheduling algorithm, where the subscriber station only has to compete once (for initial entry into the network). After that it is allocated a time slot by the base station, and it can keep the time slot even when the channel is overloaded. The scheduling algorithm also allows the base station to control QoS parameters by balancing the time-slot assignements among application needs of subscriber stations.

Generally, the scheduling algorithm decides which user traffic to map into a frame from queues according to their service class. The scheduling service type is one of the following: *Unsolicited Grant Service (UGS)*, *real-time Polling Service (rtPS)*, *non-real-time Polling Service (nrtPS)*, and *Best Effort (BE) service.*

2.1.4 Big Convergence

The integrations of different wireless technologies have been more and more embraced. Instead of being alternatives, WLAN and WiMax are becoming more important as complementary technologies of 3G and beyond 3G cellular networks. For example, the authors of [NVFA06] proposed a possible UMTS-WiMax interworking architecture based on 3GPP standards and the seamless inter-system handover scheme, which enables the service continuity with low handover latency and packet loss. M. Jaseemuddin ([Jas03]) proposed a possible architecture of integrating UMTS and 802.11 WLAN, which allows a mobile node to maintain data connection through WLAN and voice connection through UMTS in parallel.

A big picture of hybrid mobile/wireless environments is illustrated in Figure 3, which serves as a general platform for the research in this thesis.

To enable smooth communications of multiple types of applications over such heterogeneous networks, IP becomes the unified platform that shields the user from the diversity of access technologies. IP-based future generation networks promise support of mobility, security, and quality of service to end users and applications. Given the achievements on IP QoS provisioning in fixed networks, it is expected that end-to-end QoS support can be obtained by extending IP QoS into wireless segments. In the next section, we give a brief review of major approaches in IP QoS provisioning.

2.2 IP QoS Provisioning

In the IP network environment, QoS refers to the performance of IP packets flowing through one or more networks. There have been considerable efforts on delivering QoS

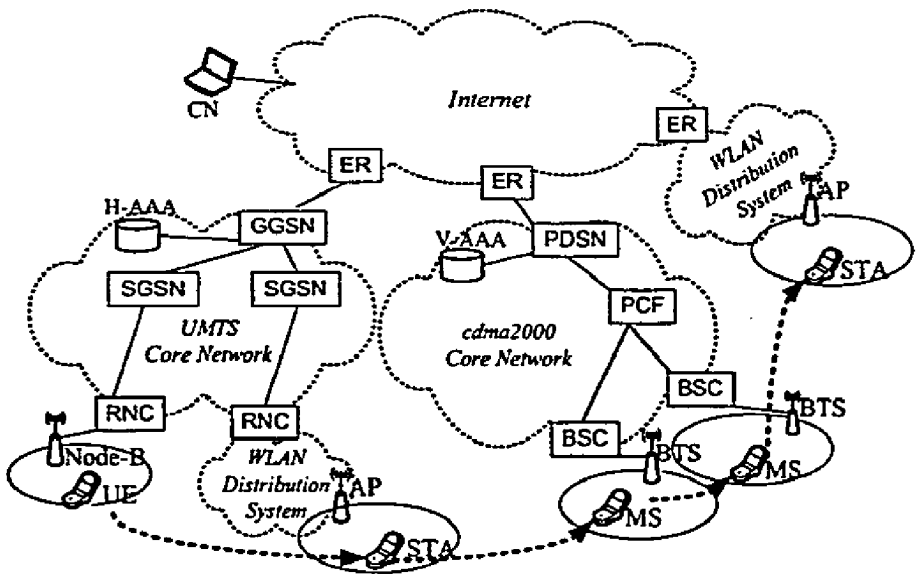


Figure 3: A big picture of mobile computing environments

into the Internet in the past decade. Some attempts aim to build QoS architectures, which suggest the interaction and cooperation between different components to provide QoS guarantees as a whole, while others focus on QoS signaling protocols, which deliver and negotiate QoS configuration parameters between network devices. These first generation attempts had no intention to accommodate situations of terminal mobility.

2.2.1 IP QoS Architectures

There are two major flavors of QoS architectures in the literature: *reservation-based* and *prioritization-based.* Reservation-based QoS architectures are usually *stateful,* i.e., intermediate routers maintain per-flow states, making admission control decisions and exchanging information via signaling protocols. Prioritization-based QoS

architectures are usually *stateless*, i.e., once the packet treatments are differentiated by the classifier residing at border routers, the intermediate nodes do not keep per-flow information or have any control over admission. They simply forward the packets according to the service levels marked by the border routers.

Two end-to-end IP QoS architectures, each representing one flavor, have been standardized by the IETF and widely deployed.

- *The Integrated Services (IntServ) architecture*

  The philosophy of this model is that "there is an inescapable requirement for routers to be able to reserve resources in order to provide special QoS for specific user packet streams or flows. This in turn requires flow-specific state in the routers" [BCS94]. Three classes of such flow-based service are defined in RFC 1633 [BCS94].

  - Guaranteed Service: providing strict guarantees on QoS requirements, such as bandwidth, bounded delay, and packet loss;
  - Controlled-load Service: approximating best-effort service in a lightly loaded network;
  - Best-effort Service: similar to what the Internet currently provides under a variety of load conditions, from light to heavy.

  The main components required to implement IntServ include the signaling protocol, the admission control routine, the classifier and the packet scheduler. Being a stateful QoS model, IntServ uses the Resource reSerVation Protocol (RSVP) [BZ+97] between senders and receivers for per-flow signaling. RSVP messages traverse the network to distribute traffic characteristics and request resources. Routers along the path including core routers must maintain soft states for RSVP flows.

  Scalability is a key architectural concern, since IntServ requires end-to-end signaling and must maintain per-flow classification and scheduling at every router

along the path. Other concerns include how to authorize and prioritize reservation requests and what happens when signaling is not deployed end-to-end.

- *The Differentiated Service (DiffServ) architecture*

  The DiffServ architecture [BB+98] intends to provide a "looser" or "simpler" service, compared with that provided in IntServ. It minimizes signaling and concentrates on aggregated flows and per hop behavior applied to a network-wide set of traffic classes.

  An end-to-end differentiated service is obtained by concatenation of per-domain services and Service Level Agreements (SLAs) between adjoining domains along the path that the traffic crosses in going from source to destination. Per domain services are realized by traffic conditioning at the edge and simple differentiated forwarding mechanisms at the core of the network. Therefore, it is a typical stateless QoS model. The DS byte in each IP packet is used to define per-hop forwarding behavior (PHB). Two of the most popular proposed forwarding schemes are:

  - Expedited Forwarding [J+99](EF, also called Premium Service)- A forwarding treatment for a particular DiffServ aggregate where the departure rate of the aggregate's packets from any DiffServ node must equal or exceed a configurable rate.
  - Assured Forwarding [H+99](AF, also called Assured Service)- A means for a provider DS domain to offer different levels of forwarding assurances for IP packets received from a customer DS domain.

  One problem with the DiffServ architecture is the lack of explicit signaling protocols to dynamically negotiate QoS requirements and accordingly configure network resources, which triggers the proposal of the ISSLL architecture, as introduced below. Besides, hard QoS guarantees cannot be provided by DiffServ, which is not in accordance with the requirements of some time critical

applications.

- *Other architectures*

  IntServ/RSVP and DiffServ can be seen as complementary technologies in the pursuit of end-to-end QoS [Ber00]. The IETF ISSLL working group proposed architecture for *Integrated Services over DiffServ Networks*, which provides a reservation-based QoS architecture with some feedback signaling about the state of the network. The architecture uses RSVP to signal resource needs but uses DiffServ as the technology to do the actual resource sharing among flows. The reference architecture includes a DiffServ region in the middle of two IntServ regions. Basically, the more DiffServ routers we have, the more scalable the service is. The basic requirements and assumptions are that the resource signaling is done with RSVP and that we have a mapping at the border nodes for RSVP-based reservations to DSCP values.

  Resembling the two-tier routing hierarchy in the Internet, Terzis, et al. [TW+99] proposed a *Two-Tier resource management model* for the Internet. Individual administrative domains independently make their own decisions on strategies and protocols for internal resource management and QoS control. The aggregate traffic crossing domain borders is served according to relatively stable bilateral agreements. End-to-end QoS support is achieved through the concatenation of such bilateral agreements. A Bandwidth Broker (BB) acts as the resource manager for each administrative domain. On the inter-domain level BB is responsible for negotiating QoS parameters and setting up bilateral agreements with neighboring domains. This two-tier model pushes the unavoidable complexity in resource management to the edge of the network, which provides substantial scaling characteristics by decoupling inter-domain allocations from individual end-to-end flows. However, BB is a very complex entity. Moreover, additional cost has to be added to protect the authentication and confidentiality of messages exchanges between BBs, maintain the robustness of BBs and

manage the state sharing between neighboring BBs in a soft-state manner.

2.2.2 QoS Signaling Protocols

A reservation-based QoS model relies on signaling protocol messages to deliver traffic characteristics and QoS requirements between end users, so that routers can collect necessary information and perform admission control and resource allocation accordingly. In a prioritization-based QoS model, signaling protocols may also play a role in indicating changes to edge routers and adjusting service level agreements dynamically. The importance of signaling protocols in QoS provisioning triggers our interest of research in this thesis. In this section, we investigate on various IP QoS signaling protocols. X. Fu gives a comprehensive analysis of IP QoS signaling protocols in RFC 4094 [MF05]. Among all of the protocols, RSVP is the most significant one.

Resource reSerVation Protocol (RSVP)

The protocol was designed to provide end-to-end signaling services for application data streams. Hosts use RSVP to request a specific quality of service from the network for particular application flows. Routers use RSVP to deliver QoS requests to all routers along the data path. RSVP maintains and refreshes per-flow soft states for requested QoS. The sender advertises the traffic characteristics periodically to the receivers via *Path* reserve resources along the flow path from the sender. messages. On receipt of an advertisement, a receiver may generate a *Resv* message to RSVP scales in that it supports large multicast groups, at the cost of high complexity in dealing with multicast in its basic protocol. Since the protocol was designed for multicast applications, a full-fledged set of features for supporting multicast is needed even for unicast applications, which results in huge handling overhead and inefficient resource reservation.

The standard RSVP has some problems dealing with mobility support. There are many discussions and proposals on the interactions of RSVP and mobility management schemes, which will be discussed in Section 2.4.3.

NSIS Signaling Suite

The existence of signaling protocols with various intentions in heterogeneous environments results in problems such as parameter mapping and function conversion between protocols, and local complexity with multiple implementations of protocols. To develop a universal, flexible, efficient, and reusable signaling framework, the IETF Next Steps in Signaling (NSIS) working group concentrates on a two-layer signaling paradigm. The intention is to re-use, where appropriate, the protocol mechanisms of RSVP, while at the same time simplifying it and applying a more general signaling model [Bru04] [HK+05].

In order to achieve a modular solution for the NSIS requirements, the NSIS protocol suite is structured in two layers:

- NSIS Transport Layer Protocol (NTLP) - a "signaling transport" layer, responsible for moving signaling messages around, which should be independent of any particular signaling application. The basic NTLP functionality is essentially just efficient upstream and downstream peer-to-peer message delivery, in a wide variety of network scenarios.

- NSIS Signaling Layer Protocol (NSLP) - a "signaling application" layer, which contains functionality such as message formats and sequences, specific to a particular signaling application. Different signaling applications will require specific protocol behavior in their NSLP. The most common options include sender or receiver orientation, uni-/bi-directional operation, heterogeneous operation, peer-to-peer and end-to-end relationships, etc.

QoS-NSLP ([MKM08]), designed with QoS as the signaling application, is conceptually similar to RSVP, and uses soft-state peer-to-peer refresh messages as the primary state management mechanism. However, QoS-NSLP does not mandate any specific QoS architecture. The Generic Internet Signaling Transport (GIST, [SH08]) provides a possible solution for NTLP. GIST manages its own internal state and the configuration of the underlying transport and security protocols to ensure the transfer

of signaling messages on behalf of signaling applications in both directions along the flow path.

Other Protocols

There are also many other signaling protocols proposed to deliver QoS into the legacy Internet. ST-II [DB95] is an experimental reservation protocol designed for point-to-multipoint communication. However, since it is sender-initiated, it does not scale with the number of receivers in a multicast group. It is also complex, and the total amount of reservation state is very large, which to some degree triggers the design of RSVP.

YESSIR [PS98] is another resource reservation protocol that seeks to simplify the process of establishing reserved flows while preserving many unique features introduced in RSVP. The proposed mechanism generates reservation requests by senders to reduce the processing overhead. However, it is built as an extension to the Real-Time Transport Control Protocol (RTCP); therefore it requires support in applications, which puts certain limitations over its wide deployment.

[Rob05] provides an "in-band" QoS signaling proposal for use within IPv4 and IPv6 protocols. It allows the necessary resources to be allocated to a flow (or group of flows) as they traverse the network. Thus, the QoS is set up in real time across the network without a separate, out-of-band, software signaling structure such as RSVP. This signaling scheme can be used to set the rate, burst tolerance, preemption priority, delay priority and charging direction for a flow. However, it requires hardware or microcode support in the participating routers.

2.3 IP QoS in Wireless Networks

The IP-based convergence of different wireless networks has promoted a lot of work on proposing IP-based QoS mechanisms in both cellular systems and the integrated networks.

2.3.1 IP QoS in 3G networks

The authors in [CC+03] constructed a framework that aims at achieving end-to-end wired-wireless QoS control with effective QoS translation, proper control management and dynamic SLA-based resource provisioning. The framework makes use of dynamic QoS arbitration, by using PDP context Activation/Modify messages, which can be changed during a real-time session. The philosophy of this approach is to interwork QoS of both wired and wireless (UMTS) network, using a mix of dynamic SLA-based and policy control schemes. It pays little attention to the impact of mobility on QoS provisioning. Besides, it does not address the problem of QoS signaling exchanges, since the QoS control mechanism has been fixed to the DiffServ approach.

2.3.2 IP QoS for Integrated 3G and WLAN Networks

Xiao, et al. in [XL+05] provided a comprehensive survey on integration of 3G and WLAN. They particularly studied handoff QoS mapping and guarantees between 3G and WLAN, as well as how seamless voice/multimedia/data handoff becomes possible.

[SP04] proposed a framework for a unified QoS support mechanism for IP traffic over UMTS and Wireless LANs. The approach utilizes RSVP for negotiating QoS parameters and reserving resources in an end-to-end basis. The authors presented a method to map RSVP parameters onto UMTS and 802.11e HCCA QoS parameters, as well as required signaling exchanges.

A generic reservation-based model was proposed by X. Wang, et al. in [WMAB03] and [WMM04] for end-to-end QoS support in the integrated WLAN and cellular networks. The adaptive QoS architecture consists of components of QoS policy provisioning, degradation profile, QoS connection admission control, QoS mobility management module, and QoS monitoring.

2.4 QoS Provisioning in Mobile Computing Environments

The major problem caused by terminal mobility arises when handoff happens. When the mobile node switches its point of attachment, the changes of address and routing often result in inaccessibility of previously established resource allocation status along the data path before handoff. Three procedures account for the potential service disruption during handoff.

- Authorization and authentication: to ensure the mobile has eligible data service access at the new location;
- Address and routing updates: to ensure packets are delivered to and from the new location;
- QoS updates: to ensure the desired service level is maintained after the handoff.

While the first aspect is extremely important and necessary for both mobile users and service providers, we focus on minimizing service degradation after the mobile obtains a new address. Since most approaches to this end interact or cooperate with mobility management schemes, we briefly review different levels of mobility and corresponding mobility management protocols in the next section, before current proposals for QoS provisioning in mobile environments are discussed.

2.4.1 Mobility Management

Not all the movements bring changes to IP address of the mobile node and the entire route of data delivery. Three distinct levels of mobility support can be identified in accordance with the overall network model presented in Figure 3.

Access mobility support refers to the methods and protocols that ensure uninterrupted communication as a host changes position between the APs within the scope

of a single RNC or an AR. This level of mobility management is tightly coupled with the specific wireless technology, and therefore is out of the scope of our research.

An IP-enabled mobile host communicates with its correspondents via a globally reachable CN node, e.g., GGSN in UMTS. Changing positions between different such nodes requires wide-area mobility support, or *macro-mobility support*. This level of mobility management is usually based on the family of Mobile IP (MIP) protocols, either MIPv4 or MIPv6. Every time a host moves beyond the limits of link layer connectivity, a registration or a binding update message needs to propagate all the way to the host's Home Address (HA). We will give a few more details below.

Generally, there are two types of addresses associated with a mobile host, its identifying address (e.g., the MIP static HA) that uniquely distinguishes the host from its peers, and the routable address that is used to reach the mobile from elsewhere in the network (e.g., a temporary Care-of-Address in MIP). The term *micro-mobility* refers to any host movement outside the scope of a single AR that does not require a change of its routable address. If a mobile host experiencing micro-mobility changes its point of attachment so frequently, the MIP tunneling mechanism introduces network overhead in terms of increased delay, packet loss and signaling.

Mobility management schemes dealing with different levels of mobility are presented below. The difficulty brought by mobility to QoS control lies in providing the requested service or maintaining the service level of ongoing sessions when the mobile node changes its point of attachment to the network, no matter in which level the change takes place.

Macro-mobility Management

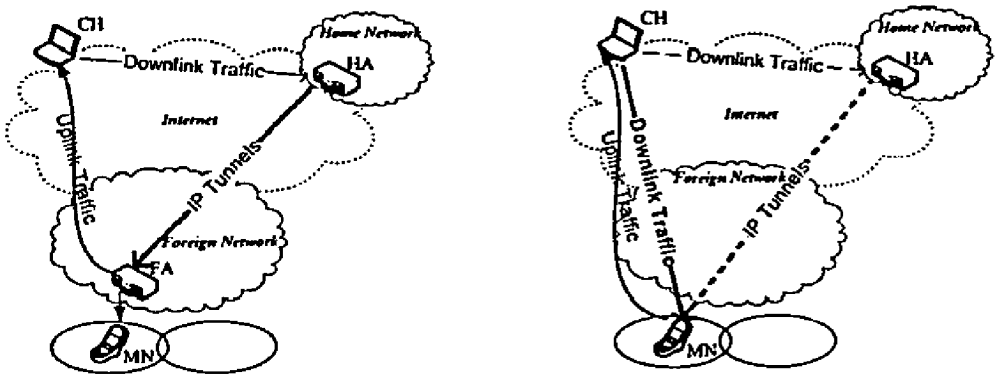
The original Mobile IP protocol, also mentioned as MIPv4 ([Per96]), utilizes the forwarding mechanism of IPv4. It enables a mobile node to keep communication with corresponding nodes (CN), after changing its point of attachment from one IP subnet to another. The identifying IP address of the MN, called home address (HA), is used for transport and higher layer sessions. The temporary routable IP address

of the MN, named care-of-address (CoA), is needed to route packets correctly to the actual point of attachment. In this way, the impact of host mobility is reflected only in the routing consideration and is kept transparent to transport and higher layer protocols.

Each time the mobile node (MN) moves from one subnet to another, it obtains a new CoA, which is either the address of the foreign agent (FA), the mobility agent serving in the foreign subnet, or an IP address acquired through some external means. The association between the HA and the CoA of the MN is called a binding. The MN updates its binding with the home agent whenever it obtains a new CoA by sending a Binding Update (BU) message to its home agent, a router in its home subnet that keeps the latest binding in its Binding Cache and serves as a proxy for the MN until the binding entry expires. The home agent intercepts any packets addressed to the MN's HA and tunnels them to the MN's CoA. The FA detunnels the packets and delivers them to the actual location of the MN.

MIPv4 suffers from several drawbacks, among which triangular routing is the most severe one. As shown in Figure 4(a), the path a packet takes on its route to a mobile node is not optimal, since it has to first reach the home agent, and then be tunneled to the final destination. Another problem arises at the ingress filtering routers. Since it is set up as the HA, the source address of the packet from MN may indicate a different network compared to the one from which the packet originated.

The Mobile IPv6 protocol (MIPv6, [JPA04]) retains the basic concepts of its predecessor, but the neighbor discovery feature of IPv6 enables an MN to operate without explicit support of an FA. Besides updating the binding entry list of the home agent, the MN also performs correspondent registration, which enables route optimization and eliminates triangular routing. By sending BU to its correspondent nodes (CN), the subsequent packets can be directly destined to the new CoA (NCoA) of the MN. Figure 4(b) illustrates the route optimization of MIPv6. The source address in the header of an outbound packet is set to the MN's CoA, thus making it recognizable to the source address filtering routers.



(a) Mobile IPv4 (b) Mobile IPv6

Figure 4: Mobility support in IPv4 and IPv6

During MIPv4 or MIPv6 handover, there is a period during which the MN is unable to send or receive packets because of link switching delay and IP protocol operations. This "handover latency" resulting from standard Mobile IP procedures, namely movement detection, new CoA configuration, and Binding Update, is often unacceptable to real-time traffic such as Voice over IP. FMIPv6 [Koo05] speficies a protocol to improve handover latency due to MIPv6 procedures. The protocol reduces the movement detection latency by providing the new access point and the associated subnet prefix information while the MN is still connected to its current subnet. The NCoA can also be formulated ahead of the handover. To reduce the Binding Update latency, FMIPv6 specifies a tunnel between the Previous CoA (PCoA) and the NCoA. Since there have been a number of extensions to MIPv6, it is critical to evaluate the performance and make effective use of different protocols according to traffic characteristics and user mobility models. An analytical framework for performance evaluation of MIPv6 and its extensions has been proposed by Makaya and Pierre in [MP08b]. They investigated the effect of system parameters, such as subnet residence time, packet arrival rate and wireless link delay, and demonstrate a trade-off between performance metrics and network parameters.

With both versions of MIP, the signaling cost during home registration and/or corresponding registration increases proportionally as the distance from the mobile

host to the home network increases. This is one major trigger for developing micro-mobility management schemes.

Micro-mobility Management

Micro-mobility protocols aim to reduce delay and packet loss during handoff and eliminate registration between mobile hosts and possibly distant home agents when mobile hosts remain inside their local coverage areas.

Existing proposals for micro-mobility can be broadly classified into two types [CGC01]: *routing-based* and *tunnel-based* schemes. Routing-based schemes aim to exploit the robustness of conventionail IP forwarding. A distributed mobile host location database is created and maintained within the network domain. The database consists of individual flat mobile-specific address lookup tables and is maintained by all the mobility agents within the domain. These schemes are exemplified by the Cellular IP [VCG98] and HAWAII [RP+02] protocols, which differ from each other in the functionality of the nodes and the construction methods of the lookup tables.

The tunnel-based schemes take advantage of registration and encapsulation in a local hierarchical fashion, thus creating a flexible concatenation of local tunnels. Hierarchical Mobile IPv6 (HMIPv6, [SCEMB05]) falls into this category. It relies on a tree-like structure of foreign agents. Encapsulated traffic from the home agent is delivered to the root foreign agent. Each foreign agent on the tree decapsulates and then re-encapsulates data packets, as the packets are forwarded down the tree of foreign agents towards the mobile host's point of attachment.

Mobility support may take advantage of the capabilities of MPLS in terms of support of QoS, traffic engineering, advanced IP services, and fast restoration. The authors of [CKK02] proposed the MPLS-based transport concept in UTRAN. A hierarchical mobility management scheme based on MPLS was proposed in [GAD02].

2.4.2 QoS Architectures Supporting Mobility

Since mobility support imposes many new requirements on QoS provisioning, some researchers proposed new QoS architectures to embrace the challenges of mobility support, while others proposed extensions to the existing QoS schemes in order t adapt to the new situation.

- *QoS architectures dedicated to mobility support*

  Although earlier proposals are mainly based on ATM as the transport technology, several useful features would still enlighten us. The AQuaFWiN framework ([VJFD99]) used a feedback mechanism to support adaptability at all layers of the wireless network, from adaptive applications to adaptive medium access control. Moreover, the authors proposed flexible QoS parameters to specify the QoS requirements of adaptive applications. Instead of specifying an average bandwidth, the application can specify minimum and average bandwidths, so that at least the minimum bandwidth could always be provided.

  Proposing additional QoS parameters for the mobile computing environment is one of the major features of [Sin96]. The graceful degradation of service parameter characterizes the degree of degradation that occurs due to mobility. A loss profile specifies the preferred way in which data can be discarded when adequate bandwidth is not available. The probability of seamless communication parameter specifies in probability terms the requirement for seamlessness of the connection. Besides, the nodes in mobile network are organized in a three-layer hierarchy to provide efficient mobility management. The QoS associated with the wireless parties were suggested to be negotiated separately from the fixed part.

  Gao, et al. in [GWM04] proposed a new QoS framework integrating a three-

plane network infrastructure and a unified terminal cross-layer adaptation platform to provide seamless support for future applications in heterogeneous networks. After defining hyper handovers and addressing QoS issues in hyper handovers, the authors presented the data plane, the control plane, and the management plane of the new framework. A unified cross-layer adaption platform (CLAP) is also presented to hide the complexity of the underlying heterogeneity from mobile applications.

- *DiffServ-based architectures*

  A global QoS architecture for multimedia traffic in mobile heterogeneous environments is described in [MA+03]. An integrated management approach to service and network management is presented based on cooperative association between QoS brokers and AAAC systems. Since the QoS architecture relies on a DiffServ approach, this approach suffers from intrinsic DiffServ problems, as each domain has to implement a correspondence between each network service and a DSCP. Furthermore, only soft QoS can be provided with this approach, which is not sufficient for the applications that impose stringent delay or other time-related requirements to the network providers.

  Focus has also been put on extending specific types of services designed for mobility scenarios. J. Diederich, et al. [DWZ03] proposed the Mobile Differentiated Services QoS Model (MoDiQ), a DiffServ-based framework to provide QoS in wireless mobile networks. The MoDiQ service model comprises separate services for mobile terminals and non-mobile terminals, in order to enhance the efficiency of resource utilization in scenarios where not all mobile terminals are actually moving.

  Regarding the interworking of Mobile IP and DiffServ, Yoon, et al. in [YL+00] proposed schemes with hierarchical agent architecture for supporting QoS in mobile/wireless IP networks. To support DiffServ for mobile hosts, they configure a foreign agent to have pre-allocated resources in advance for the static

SLA case. In case the foreign network has no pre-configured SLA or enough resources for the mobile host, a new SLA is to be set up among bandwidth brokers using dynamic SLA.

2.4.3 QoS Signaling Protocols for Mobile Networks

Since QoS signaling protocols play a key role in delivering QoS information and negotiating service level, it is natural to make certain enhancements to existing protocols to accomodate multiple levels of mobility.

- *QoS signaling framework for next generation networks*

  As we mentioned in Section 2.1, heterogeneity is one of the most significant features of the future generation networks. The authors in [P+05] took the diversity of QoS models chosen by the network operators into consideration. They proposed a 4G network architecture and evaluated possible associated QoS signaling strategies. They presented message sequence charts associated with the scenarios covering terminal-initiated signaling, network-controlled signaling, and application-provider controlled signaling. The conclusion is that the optimum QoS signaling solution depends on the QoS model being used, which is directly related to the business model chosen by the operators: application-oriented, user-oriented, or service-oriented.

  The authors in [H+04] tackled another aspect of heterogeneous future generation networks, i.e., seamless handover over different access networks. Their QoS signaling architecture integrates resource management with mobility management, based on a Domain Resource Manager concept, which is essentially a centralized QoS control architecture. The approach aimed at supporting various handover types, but in particular supported anticipated handover with pre-reservation of resources over the old network before the mobile node is attached to the new access point.

- *Proposed mobility support in the NSIS signaling suite*

Mobility support is considered as one of the desired features of the NTLP, which was identified in [FST03]. The basic function of NTLP mobility support is to handle the route changes and notify the associated NSLP of the related information to trigger fast QoS re-establishment in the level of QoS-NSLP, which is the most critical function in QoS-NSLP to provide mobility support in the NSIS suite.

There was also a proposal for identifying mobility functions in the QoS-NSLP, named mQoS-NSLP ([MKM08]). One of its key features is fast QoS re-establishment. It also should support resource reservation in both sender- and receiver-oriented modes where a mobile node may act as a sender or receiver. The desired mobility functions for the mQoS-NSLP include make-before-break, mobility event detection, localized path repair, state management, interaction with mobility protocols, and so on.

All of the above mobility functions in the NSIS signaling protocols are basically intended to work in the mobile environments managed by Mobile IP. Detection of route changes caused by mobility is based on the change of IP address of the related node. Frequent handovers will result in a high number of interactions between NTLP and NSLP, which might cause processing overhead at the involved nodes and reduce the efficiency of fast QoS re-establishment.

- *RSVP Extensions*

Many researchers proposed to extend RSVP to provide mobility support, because RSVP is the most important QoS signaling protocol in use. B. Moon and H. Aghvami summarized most of the RSVP extensions for real-time services in wireless mobile networks in [MA01].

Mobile RSVP (MRSVP, [T+01]) proposed by Talukdar is a typical *advance reservation* solution, which provides resource reservation in advance in some potential new cells. MRSVP extended the reservation model of the original RSVP. It proposed a network architecture in which a mobile host can make "advance

resource reservation" along the data flow paths to and from the locations it may visit during the lifetime of the connection. It is based on the assumption that the set of locations the mobile host will visit, recorded in MSPEC, can be acquired, either from the network or from its mobility profile. Although it ensures that there is enough resource at the next cell, MRSVP generates too much control traffic because each user flow needs to periodically refresh its reservation states in multiple cells. Furthermore, it does not utilize the wireless bandwidth efficiently. If the mobility profile could not be obtained precisely, the advance resource reservation would cause a huge amount of overhead due to the state maintenance in the cells waiting for a potential visit.

A similar approach has been proposed for micro-mobility handoff. QoS negotiation in micro-mobility is performed within an access network using RSVP in [PC02]. In the proposed scheme, when a Path message arrives at an Access Network Gateway (ANG), the ANG reserves one distinct path to the Access Router (AR) in the current location of the destination MH and multiple shared paths to the ARs of adjacent cells. By reserving resources in multiple paths in not only the specified cell but also adjacent cells in advance, the authors aim to provide the seamless QoS handling for micro mobility management. However, the signaling overhead in multiple paths may take over use of the limited resources in the wireless links.

The second type of RSVP extensions is based on *interworking of RSVP and Mobile IP family.* Terzis, et al. ([TSZ99]) proposed a simple QoS signaling protocol by combining pre-provisioned RSVP tunnels with Mobile IPv4. When the home agent is informed about the MN's new location, it sets up a tunnel RSVP session between itself and the foreign agent if one does not exist. It then encapsulates Path messages from the sender and sends them through the tunnel towards the MN's new location. When the FA receives a Resv message from the MN, it sends a Resv message for the corresponding RSVP tunnel session between itself and the home agent of the MN. This approach can be

easily implemented with minimal changes to other components of the Internet architecture. However, the triangle routing problem occurs when the MN is far away from the home network.

G. Chiruvolu, et al. ([CAV97]) proposed a MIPv6 and RSVP integration model. The main idea is to use RSVP to reserve resources along the direct path between the CN and the MN without going through their home agents. Flow identification (source and destination addresses) is based on the MN's CoA. Whenever the MN performs a handoff, which incurs a path change, it sends a binding update to the CN. The CN then sends a Path message associated with the new flow from CN to MN. Upon receiving this message, the MN replies with a Resv message to complete the resource reservation procedure. All the RSVP re-negotiations are conducted end-to-end even when the path change affects only a small portion of the entire route. Hence, the handoff resource reservation delays are long and the signaling overheads are large, which leads to notable service degradation to real-time applications.

To solve the problems in the previous model, Q. Shen, et al. ([SLSK00]) proposed a Flow Transparency (FT) concept, which keeps flow identity unchanged during handoff in order to keep host mobility transparent to transport layer protocols. This method limits the handoff RSVP re-negotiation process within the newly added portion of the path between CN and MN. However, implementation of this model requires some modifications to MIPv6. LePaja, et al. ([LF+02]) extended this FT model to meshed access network topologies, by using the previous access router(s) as a nearest common router (NCR) for the old and the new added flow paths.

- *Extensions to DiffServ and MPLS*

  [THM05] provides a good survey on extensions that have been proposed to enhance the QoS functionality of Mobile IP. A number of DiffServ extensions and MIP-specific MPLS extensions are presented and discussed with a comparison

of their major features.

Besides, the authors in [JZM02] provided a study of profiled handoff for DiffServ-based mobile nodes, which shows transfering contexts to the new edge routers of wireless subnets helps various marking schemes reach stability earlier.

MPLS-based micro-mobility mechanisms could provide intrinsic QoS support through traffic engineering of the MPLS paths, as proposed in [ZMH04]. Utilizing the Class of Service (CoS) field in the MPLS header, QoS can be supported with mobility. Multiple LSPs can be established for the same destination mobile host in order to accommodate applications with different QoS requirements.

2.4.4 Extending Mobility Management Protocols

In Section 2.4.3, various RSVP extensions have been proposed to interact with MIPv4 or MIPv6 for seamless handoff in the macro-mobility level. The other category of approaches is to modify mobility management protocols to enable proper QoS re-negotiation and re-configuration during handoff.

The approach of QoS Object Option ([CK01]) is based on a new IPv6 extension header option called QOS OBJECT OPTION to forward QoS requirements of mobile nodes to the routers along the connection path. X. Fu, et al. in [FKK02] extended this approach based on HMIPv6 to enable QoS-conditioned handoff.

C. Makaya and S. Pierre in [MP08a] and [MP08c] proposed a mobility management protocol, Handoff Protocol for Integrated Networks (HPIN), based on score function, fast handover and anticipated resource reservation principles, to allievate service disruption during user roaming by allowing selection of the best available network.

Y. Cheng and J. W. Atwood in [CA06a] proposed a hierarchical mobility management architecture providing adaptive QoS control to improve the performance during handoffs. The information for QoS configuration and location management is carried by the same set of messages. The interactions among the hierarchically-organized QoS

Agents (QA) enable smooth handoff. However, the control communications between QAs may bring high overload to the network.

2.5 Chapter Summary

In this chapter, we presented a literature survey of IP QoS provisioning in future generation networks. Given the trend of IP-based convergence of heterogeneous future generation networks, the legacy IP QoS strategies are examined, as they are the base of most later approaches. Terminal mobility in wirelss/mobile networks imposes new requirements on continuous service provisioning. Many approaches have been proposed to reduce service disruption or degradation during handoff. These approaches can be categorized into three groups: QoS architecture-based, QoS signaling protocol-based and mobility managment protocol-based.

The focus of this thesis is on QoS signaling framework and protocols proposed to meet the challenges of wireless/mobile networks. In the next chapter, we will present a detailed analysis on these challenges and validate the need for a novel QoS signaling framework.

Chapter 3

QoS Signaling in Future Networks: Challenges and Requirements

3.1 Challenges to QoS Signaling Protocols

According to [Jai], a QoS architecture is usually composed of the following components:

- Services with different QoS: Service definitions.
- Ways for users to communicate what they need: Signaling or admission control, policy management.
- Ways for providers to ensure that users are following their commitment: Policing or shaping.
- Ways for providers to find the routes: QoS based routing.
- QoS based forwarding: Buffer allocation and drop policy, queuing discipline and service policy, traffic management.

Among these components, QoS signaling mechanisms enable the communications among network entities on different layers. In particular, IP QoS signaling in an

IP QoS architecture enables hosts to request desired level of treatment and routers to reach agreement on service provisioning. Some QoS architectures require explicit signaling, such as RSVP in IntServ. Other QoS architectures imply implicit signaling, such as service level negotiation in DiffServ. In both cases, communications between network entities play an important role in assisting QoS provisioning.

In wireless and mobile environments, the extreme variability of mobile traffic, the need for handoff support and the fact that the radio link is a bandwidth bottleneck suggest that all the participating network systems have to manage resources very carefully. Hence, it is important and natural to develop relatively complicated resource admission, control and handoff procedures in such network systems. With these procedures being added, the control communications among network entities are expected to keep relevant nodes informed of and updated with the constantly changing conditions, such as terminal mobility or shrinking network capacity.

Moreover, the heterogeneous nature of future generation networks invokes more complicated signaling procedures. As we introduced in Section 2.1, each wireless system defines its own QoS architecture or QoS mechanism, resulting in different sets of service definitions, QoS specification parameters and QoS enforcement procedures. Since IP is expected to be a unified platform to hide the heterogeneity of underlying network technologies from the upper layer protocols, IP QoS is expected to provide a unified platform to hide the heterogeneity of wireless and wired QoS mechanisms from end users. This requires additional functions in QoS architectures, such as service mapping, QoS parameter translation, service authorization, and so on. As a result, more sophisticated communications and cooperations between network entities are desired.

Among all the increased complexity of QoS signaling procedures, certain scenarios of handoffs bring up the most challenging situations. Depending on the level of mobility and the type of handoff, the control communications before and after handoffs can be very complicated. Beyond maintaining service availability during any scenario of handoff, QoS control signaling protocols should take the responsibilities of

minimizing service degradation and dealing with issues of service translation, service re-negotiation, and service adaptation.

To better understand the problem, we identify different types of handoff situations in the following subsections. Each type of handoff invokes different control signaling overhead at various network entities. In some occasions, multiple types of handoffs are caused by the MN changing its point of attachment, which further complicates QoS control communications.

3.1.1 Different Levels of Mobility

If a handoff does not cause change of the access router of the mobile host, the handoff control only needs to handle radio resources since the routing paths do not change. As a result, no extra signaling is desired within the network layer QoS mechanisms.

If the access router changes but the gateway of the access network stays the same, the handoff affects both the radio resource availability and the access network resources. Micro-mobility management protocols are responsible for updating terminal location, local address information and triggering the update of packet delivery route inside the access network. On the side of QoS provisioning, the new AR needs to perform admission control upon handoff. Resource reservations need to be refreshed or updated inside the acccess network, if there are any. The additional control communications triggered by the handoff event are usually limited within the access network.

If the gateway changes too, macro-mobility management protocols need to update the terminal location and routable address information of the mobile host. Before these updates are completed, the ongoing traffic flows may experience service degradation or disruption from the network, because the change of routable address usually brings difficulty to multiple QoS components, such as admission control and packet classifying and queuing, in terms of identifying the incoming packets after handoff. The proper functioning of these components relies on the pre-established service agreement or QoS states, which may become obsolete with the change of destination

address or source address. If QoS control communications fail to notify the involved routers about the ongoing changes, the desired service cannot be delivered to the mobile host, which is unbearable to real-time applications and undesirable to other applications. Therefore, QoS signaling procedures have to trigger proper updates on QoS enforcement modules along the new path as soon as possible, in order to minimize the period during which the affected traffic flows cannot receive the subscribed level of treatment.

In summary, QoS signaling protocols need to cooperate with both macro-mobility and micro-mobility management schemes. Since various protocols are designed to handle different levels of mobility, as introduced in Section 2.4.1, there arises a big challenge on how to couple QoS control and mobility management strategies efficiently.

In Section 2.4.3, we have seen various approaches to enable interworking of RSVP and MIPv4/MIPv6, such as [TSZ99], [CAV97], [SLSK00], and [LF+02]. The IETF NSIS working group also addresses the problem of providing mobility functions in both the NSLP layer and the NTLP layer. The issue of QoS negotiation during micro-mobility handoff is studied in [PC02], with solutions being proposed. However, these proposals are specific to a fixed level of mobility management scenario. Assume a mobile host has to experience different levels of handoff while communications are going on. Different micro-mobility management or macro-mobility management schemes are applied during its movement. Thus, various strategies of coupling QoS control and mobility management schemes have to be applied during each handoff in order to achieve optimal overall performance. This results in frequent switches between different schemes and the involved functional entities. Extra control and scheduling overhead will be introduced during these switches, and the overall performance will be degraded.

3.1.2 Inter-QoS-scheme Handoff

A QoS scheme is usually network-provider-dependent. Service definitions, service granularity and service fulfilment schemes are all changed when a mobile host travels across the boundaries of QoS management schemes. Thus, service mappings and QoS parameter translations are usually necessary at the edge routers. This type of handoff may involve a change of IP QoS scheme, such as from a DiffServ domain to an IntServ domain. It could also refer to a transition between wireless QoS schemes, such as from a cellular network QoS provisioning scheme to a WLAN QoS provisioning scheme. Many researchers have paid attention to the issue of QoS parameter mapping between different QoS schemes. For example, [XL+05] suggested QoS mapping schemes between 3G and WLAN.

Furthermore, the strategy of coupling with mobility management may be different for each QoS scheme, as we have seen in Section 2.4.3 and Section 2.4.4. Inter-QoS-scheme handoff raises challenges for mobility management protocols dealing with heterogenous QoS schemes. The framework presented in [CK01] proposed to carry QoS information within a hop-by-hop option header in BU messages to trigger certain QoS procedures at the intermediate network domains. The framework described the QoS procedures for multiple QoS models, but it lacks a fast decision mechanism in the case where the new path segment is able to provide sufficient resources for the data flow, and it also lacks a QoS adaptation mechanism in the case where the resources along the new path segment are insufficient for the QoS requirements.

3.1.3 Vertical Handoff

A handoff is known as a *vertical handoff* if the mobile host moves between access points of different type, in contrast with a handoff caused by the mobile host moving between access points of the same type, namely a *horizontal handoff* ([MK04]). Different access networks may apply their own QoS schemes, which makes a vertical handoff also an inter-QoS-scheme handoff. Meanwhile, a vertical handoff may require different levels

of mobility management depending on the situation of IP address change. Therefore, QoS control communications may face all the challenges analyzed in the previous two sections.

There are some new challenges brought by vertical handoffs. Data service rate may change dramatically when the mobile host is attaching to a different type of access point. It is highly possible that network congestion and service degradation occur when traffic flows move from a high data rate network, such as a WLAN, to a low or medium data rate network, such as UMTS. When the mobile host is moving from a lower data rate network to a high data rate network, service upgrade may be desired by the users. In these circumstances, QoS control communications are expected to play significant roles in the procedures of QoS-based access point selection, fast service re-negotiation, and service adaptation, so that seamless and continuous service provisioning is possible.

M. Stemm and R. H. Katz ([SK98]) implemented a vertical handoff system that allows users to roam between cells in *wireless overlay networks*, a hierarchical structure of room-size, building-size, and wide area data networks, where different wireless technologies are applied in each layer. While the goal of this system is to provide best possible connectivity for as long as possible with a minimum disruption during handoff, the other QoS issues are not taken into account.

To enable seamless WLAN-to-cdma2000 handoff, H. Parikh, H. Chaskar, et al. ([PC+03]) proposed to improve the handoff performance using proactive handoff techniques, which include fast handoff signaling, authentication with the cellular network, informing it about QoS aspects and so on. Their guideline to reduce handoff latency, including connnection establishment, authentication and service authorization, and QoS setup, is to avoid the message exchange on a relatively low speed cellular wireless link to the extent possible. HPIN is an inter-technology mobility management protcol proposed by C. Makaya and S. Pierre ([MP08a]). It is based on score function, fast handover (with FMIPv6) and anticipated resource reservation (with RSVP)

principles. With handoff score function, the target network that results in the highest computed score function value is the network that would provide most significant benefits to the user. L. Zhang, S. Pierre and L. Marchand ([ZPM06]) proposed Access Routers Tunneling Protocol, which defines a novel approach called seamless vertical handover for Mobile IPv6 (SVHO) to allow a mobile node to resume real-time sessions immediately after being attached to the new link.

The above proposals rely on specific mobility management protocols and QoS signaling protocols in specific vertical handoff scenarios. They lack adaptability in heterogeneous network environments. For example, in order to adapt the proposed basic scheme in [PC+03] in a Fast Mobile IPv6 (FMIPv6, [Koo05]) environment, new messages need to be defined in at least three steps.

3.1.4 Inter-administrative-domain Handoff

When a mobile host moves from the administration domain of one network provider to another, the user has to get authenticated and authorized to access the service in the new domain, unless a trust relationship already exists between the two network domains. Also, the charging policies may differ depending on the network providers. This requires the QoS management schemes to interact with policy control entities, Authentication, Authorization, Accounting, and Charging (AAAC) system and other security management entities. This type of handoff often happens together with vertical handoff. For example, in the basic WLAN-to-cdma2000 handoff scheme presented in [PC+03], the home AAA, the visited AAA, and possibly middle AAA are involved in the first phase of the handoff procedure to complete authentication and authorization steps.

In summary, to provide end-to-end QoS in a heterogeneous network environment, QoS control communications are facing great challenges when dealing with different scenarios of handoff. They have to interact with various functionalities and different technologies. As we have seen, there are quite a few proposals to address the problems

with a specifc scenario. However, it is very likely that multiple types of handoff mentioned above occur at the same time. Yet there is no explicit control framework that is able to guide the coordinations of various entities in all the situations. Therefore, we argue that it is necessary and highly desired to design a QoS control framework that takes into consideration all the challenges raised by different types of handoff in future generation networks.

3.2 Requirements

The proposed QoS control communications framework aims to provide a unified guideline to dynamically negotiate and re-negotiate QoS requests between communication peers; facilitate the resource configurations and re-configurations in heterogeneous wired and wireless networks; maintain and quickly adjust the requested service level in different situations of terminal mobility. Aided by this framework, various functional components should be able to cooperate with each other to achieve smooth end-to-end QoS provisioning in dynamic network environments.

To fulfill the above objectives, a set of overall performance requirements have been imposed to the framework. These requirements are ordered by priority. When some of them conflict with each other, the requirements on the top of the list are preferable to those on the bottom.

- *Reduce service interruption or service disruption at the time of handoff.* While a certain degree of service degradation is unavoidable in many cases of handoff, we require that the period of service disruption and the variations of service level be reduced to a minimum.

- *Reduce extra control signaling overhead due to service adjustments and updates.* It is necessary for routers and hosts to exchange any updated information during handoff. However, too many explicit signaling message exchanges during handoff will largely increase the processing overhead at the involved nodes and

delivery overhead in the networks. This is especially important for the wireless network part where processing capabilities and network resources are both precious and limited.

- *Utilize the existing protocols and schemes as much as possible,* including different approaches of QoS provisioning, mobility management, policy management and security management;

- *Accomodate as many different scenarios of handoff as possible.* For example, the framework should consider all the possible changes caused by a veritcal handoff and properly deal with them.

- *Promote resource utilization.* For example, the QoS states along the old path segment should be released as soon as possible after handoff, once they are determined to be obsolete.

In addition, the following functional requirements have been take into consideration:

- Coordination between QoS management schemes and mobility protocols including both macro-mobility and micro-mobility;

- Smooth transitions among heterogeneous QoS mechanisms in both wired and wireless parts;

- Cooperation among QoS control and other functional entities, e.g., application layer signaling, security capabilities and policy control

3.3 Chapter Summary

In this chapter, we analyzed the challenges imposed on QoS control communications in future generation networks. The challenges are mainly caused by various types of handoff events. With each type of handoff, we discussed the changes related to QoS provisioning and the expected functionality of control signaling. We also presented existing proposals to deal with the challenges of each type of handoff. Due to

the limitations of these proposals, we identify the need for a unified framework for control communications in the heterogeneous environments. The overall performance requirements and some functional considerations are presented as guidelines for the design of the framework.

Chapter 4

The QoS Control Communications Framework

4.1 Overview

As we have seen in previous chapters, seamless roaming among heterogeneous networks is highly demanded in the future communications environments. Meanwhile, this feature imposes the greatest challenges to end-to-end QoS control. In this chapter, we propose a QoS control communications framework, which is intended to address the challenges based on the requirement analysis in chapter 3. It is a novel framework in terms of integrating multiple aspects of QoS control communications in mobile/wireless networks. The framework is designed to be event-driven, because the scale and complexity of control communications vary significantly in different situations. Burst number and increased complexity of control communications occur during certain events such as session setup and vertical handoff, while periodic control messages are normally sufficient during stable data transfer periods of a session. Either triggered by a certain event or invoked periodically, a set of involved network nodes communicate with each other via various protocol messages for the purpose of maintaining desired service level to end users. We do not intend to invent new signaling protocols. Instead, existing protocols are integrated and extended if necessary.

In general, the control communications framework plays the same role as the control plane in the QoS framework proposed by Gao, et al. in [GWM04].

4.2 The Multi-panel Control Communications Framework

According to the specific set of tasks to be accomplished, we distribute all the potential QoS control communications into a number of panels. Each panel is responsible for a specific aspect of control communications. Simple interactions between panels enable different function entities to work together, aiming to provide end-to-end QoS control. Since some control functions are more frequently invoked than others, the panels are placed in an order where lower panels are more frequently demanded.

The advantages of grouping control communications into multiple panels lie in the following three aspects:

- Generalizing control functionalities enables heterogeneous approaches.

  We do not and cannot specify protocols or schemes for each panel due to the heterogeneity of future communication networks and numerous proposals on individual functionality related to end-to-end QoS provisioning in such networks. Instead, the framework specifies general functionalities within each panel. Thus, different approaches to implement the functionalities in each panel can be accommodated in the framework.

- Relatively independent groups of functionalities enable efficient system optimization.

  The modifications or enhancements of specific control function within a panel should not affect those in other panels. Functions in different panels are loosely-coupled, which is suitable for quickly applying improvements on specific functions. Furthermore, the panels are placed in such an order that a lower panel is activated more frequently than a higher one during a data session's life time.

This layout of panels assists efficient responses to network events. It is obvious that optimizations in lower panels bring more performance gain to overall control communications.

- Open structure of the framework enables extendability for future networks and technologies

  It is always possible to support new QoS-related mechanisms within a panel as long as the control communications within the panel conform to the requirements imposed by the user or the event and the interactions with entities in other panels do not violate the rules of inter-panel communiations. On the other hand, it is impossible to define all the panels and communications and guarantee their effectiveness with the emergence of new technologies in the future networks. However, the framework can be easily extended by defining new panels, new rules of communications, and new triggering events.

4.2.1 Local Control Communications

The proposed framework can be further regarded as a two dimensional control network. Message exchanges between peer entities in each panel form a distributed control network in the horizontal dimension. The Agents in the same panel of different nodes receive and process messages and generate responses accordingly. In contrast, a local control network is formed in the vertical dimension by interactions between Agents in different panels of the same network node.

End hosts and routers have different types of local control communications. The functionality and complexity of local control communications at different nodes depend on the major triggering events that the nodes constantly deal with. The local control network in an end host is more complex than that in a router or a server, because an end host deals with more types of events than a router or a server does. Figure 5 illustrates different types of triggering events at end hosts and routers with

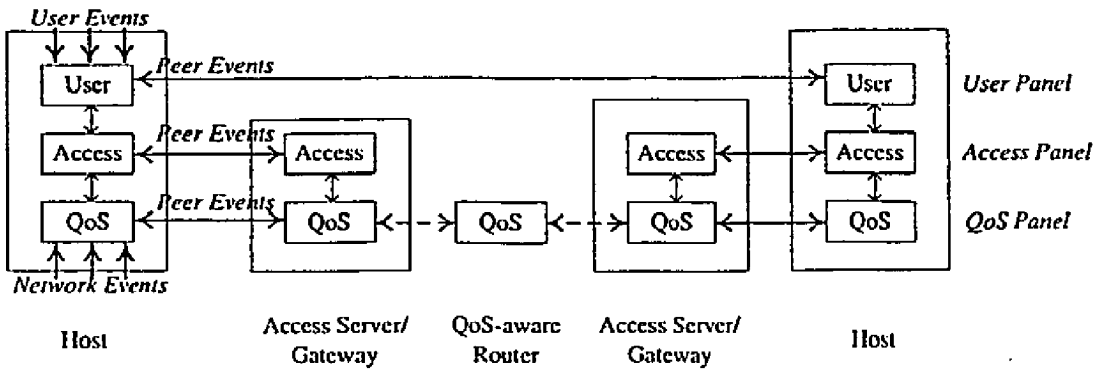


Figure 5: The multi-panel QoS control communications framework

different roles. A *network event*, such as a layer 3 handoff, activates the local control communications at an end host from bottom up. A *user event*, such as session initiation, activates the local control communications at an end host from top down. The local control network at a non-host node is generally activated by a certain *peer event*, which is typically a signaling message from a peer entity of a certain Agent.

4.2.2 Optimization of QoS Control Communications

We can utilize the different types of triggering events described above as a guide to improve the overall performance of QoS control. If network events occur most frequently in a scenario, one should consider improving performances of control functions in lower-positioned panels, because those panels are most frequently involved in dealing with network events. For the same reason, control functions in higher-positioned panels should be enhanced if user events are dominant in another scenario. Peer-event-dominated scenarios desire intra-panel efficiency, which can be achieved by enhancing related signaling protocols.

By identifying all the control communications vertically and horizontally, we provide an express way to improve overall performance of control communications. Once the parts of control that are most relevant to the target scenario are found, it is more efficient to take actions that may enhance the system as much as possible. In Section 4.3 we will see brief examples of analysis targeting at performance optimization.

Chapter 5 shows the optimization of the QoS Panel in response to a general MIPv6 handoff event. Chapter 6 shows further optimization of the QoS control framework in response to a vertical handoff in integrated cellular-WLAN networks.

4.2.3 Three Logical Panels

Currently we have defined three logical panels in the framework: *User Panel, Access Panel,* and *QoS Panel.* The functional entities within each panel are called *User Agent, Access Agent,* and *QoS Agent,* respectively. Figure 5 depicts the logical panels in different network nodes. All of the three panels are active at an end host. The User Panel is not present in routers. Depending on the role that a router plays in a network domain, the Access Panel may be included in a gateway router, a policy enforcement point (PEP), or other network nodes that are decisive to access requests from users. The QoS Panel is present at all the QoS-aware nodes. The functionalities of each panel are described below.

- *User Panel*

  This panel interacts with user applications by dealing with a service request. Some startup functions in this panel include user-level service selection, session initiation, and mobility profile designation. Normally the lifetime of a data session can be divided into three stages: session establishment, data transfer and session termination. In the proposed signaling framework, this panel is generally involved at the stages of session establishment and session termination. We argue that end users should not be expected to respond to every change in network condition, such as network congestion, route change, handoff or variation in wireless channel capacity. In other words, end users are not involved in the QoS control procedure caused by underlying networks during data transfer. However, end users are expected to propose their requirements on data services from the application perspective at the beginning of a session. They may also terminate data sessions at any time, which causes resource release at involved

network nodes. Therefore, peer entities in the User Panel, basically User Agents at the source and the destination of the session, need to communicate via certain signaling protocols during the session establishment stage and the session termination stage. An application layer signaling protocol, such as the Session Initiation Protocol (SIP), can be used for this purpose.

Occasionally, end users may want to change service requests during data transfer stage. For example, a mobile user finds out that the quality of the high-resolution streaming video that he is watching has deteriorated badly. He may opt for the low-resolution choice and continue watching the video. This type of situation will be regarded as terminating the previous session and establishing a new one with different service request parameters.

- *Access Panel*

  This panel decides the accessibility of network services requested by a session. The decision is based on multiple functionalities, including policy control functions, the functions of authentication, authorization and accounting (AAA), charging functions and all the other administrative control functions.

  The communications between Access Agents located at end terminals and access function servers are the major signaling flows in the Access Panel. Besides the stage of session establishment, other events activate control flows in this panel too. For example, an inter-administrative-domain handoff may result in new authentication and authorization requests to the AAA server located in the new domain.

- *QoS Panel*

  The fundamental tasks of QoS configuration and negotiation are assisted by the QoS Agent located in this panel. Depending on the QoS scheme being applied in the node, the entities within the QoS Panel may include:

  - QoS Agent, which takes network events as inputs and notifies other entities

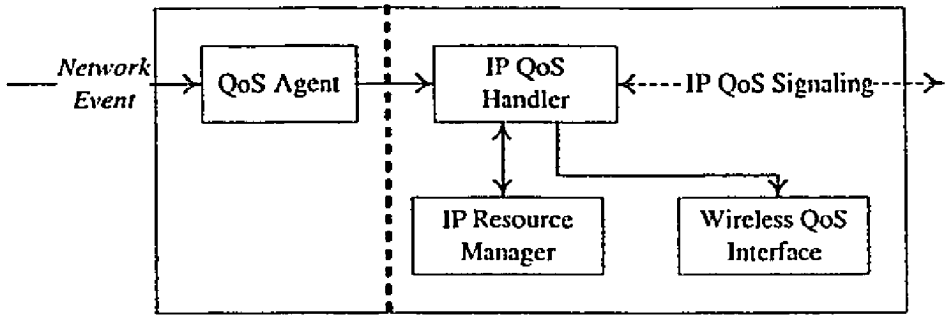


Figure 6: The entities in the QoS Panel

to take proper actions to the events;

- IP QoS signaling protocol handler, which generates and processes IP QoS signaling protocol messages;
- Resource manager, including IP QoS provisioning functions, such as admission control, traffic shaping and queue management; and/or
- Wireless QoS interface, which translates and maps IP QoS parameters to QoS parameters and enables wireless QoS function entities to perform wireless resource allocation.

The interactions between these entities are illustrated in Figure 6. Generally, the QoS Agent receives a network event such as a handoff decision or a network capacity change. It invokes proper functions of the QoS signaling protocol handler, which usually sends messages to peer entities and invokes the resource manager to make resource allocation decisions and corresponding adjustments. In a wireless network segment, the wireless QoS interface is further informed of the ongoing change and service parameters are translated so that the allocation of wireless resources can actually be performed.

This panel is activated whenever IP QoS provisioning functionalities need to be informed of changes and take proper actions. Besides service establishment and service termination, other scenarios that activate the QoS Panel include network capacity change, IP mobility event, vertical handoff, and so on.

4.2.4 Stimulating Events

To better understand how the control framework works, it is useful to identify all the events to which the control framework should react. Since user events and network events are environmental stimuli to the control framework, we focus on these two types of events and ignore the peer events, which are messages exchanged within the framework.

Major user events include:

- Session initiation - Users issue requests for data service. An example set of involved control communications in the framework is presented in the next section.

- Session termination - Users suspend or stop data transfer. This may happen whether data transfer is complete or not. The major control communications are to release occupied resources.

- Change of request - Users change their requests for data service during data transfer. This event actually results in termination of the current session and initiation of a new session.

Major network events include:

- Route change - This may have multiple causes, such as network failure, network congestion, load balancing, and so on. The key point to deal with in this event is to update resource allocation along the new route or the new path segment as soon as possible. In this regard, it is similar to an IP mobility event. The details are out of scope of this thesis.

- IP mobility event - Change of care-of address of mobile users results in an IP mobility event. Interactions between mobility management protocols and the QoS Panel reflect the level of coupling of mobility and QoS management. We take loose-coupling as the primary approach to mobility and QoS coupling.

Instant notifications of location and address updates to the QoS Panel trigger QoS updates at the involved nodes. Further discussion is presented in Chapter 5.

- Inter-administration-domain handoff - Roaming between different administrative domains requires more control communications during handoff. The Access Panel has to make a decision on granting access to handoff sessions based on administration policies and security control functions. We present the triggered control communications in the case study in the next section,.

- Vertical handoff - On most occasions, an inter-access-technology handoff is an IP mobility event and an inter-administration-domain handoff as well. However, transition from one wireless access technology to another requires more control communications, such as QoS mapping and translation. The details are presented in Chapter 6.

4.3 A Case Study

In this section, we describe the work flow of the framework responding to two typical events. Figure 7 illustrates an abstract network model. The types of access networks are not specified because our purpose is to present the control communications in a general situation. A mobile user is first located in its home network, Domain 1. It may later move to Domain 2, which is another administrative network domain. Gateway routers are located at the borders of access networks. The data session will be established between the mobile user and a correspondent node, which is represented by a video server in the figure. There are access servers, such as AAA server and PDP, in each domain.

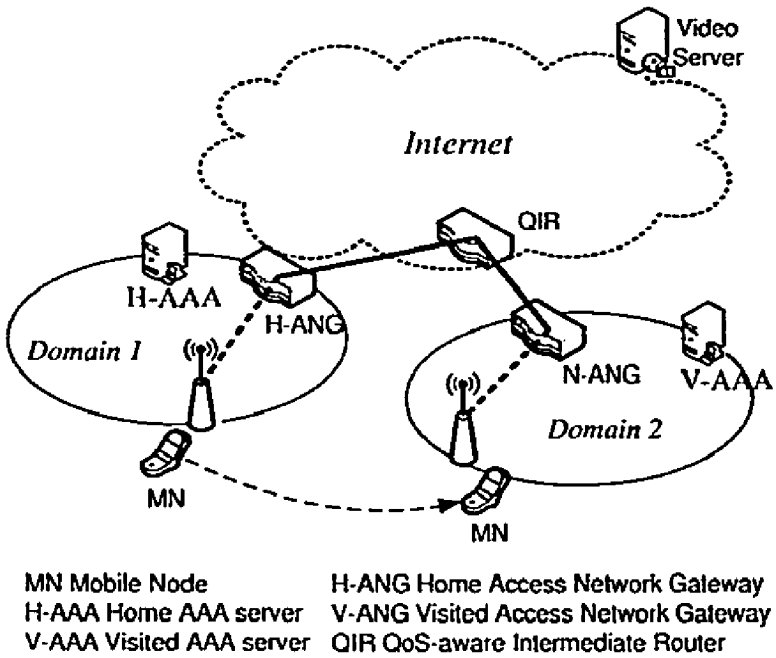


Figure 7: The abstract network model of case study

4.3.1 Scenario One: Session Establishment

The mobile user is initiating a data session with the remote video server. Figure 8 depicts the general control communications triggered by this event. At the end host, all the panels are invoked. The user event (a) of session initiation is first received by User Agent in the User Panel, which sends session request (b) to the remote server. The remote server acknowledges (c) the session request. The service request (d) with QoS requirements is further passed to the Access Agent. Access requests with multiple purposes (e) are sent to different servers within the network. These servers make the decision to deny or accept and return the access decision (f) to the Access Agent of the MN. The access servers could be located at individual server nodes or co-located at gateway nodes of the network, such as the GGSN in UMTS. Upon a positive access request, QoS request (g) is further passed to the QoS Agent, which initiates IP QoS signaling (h) along the route.

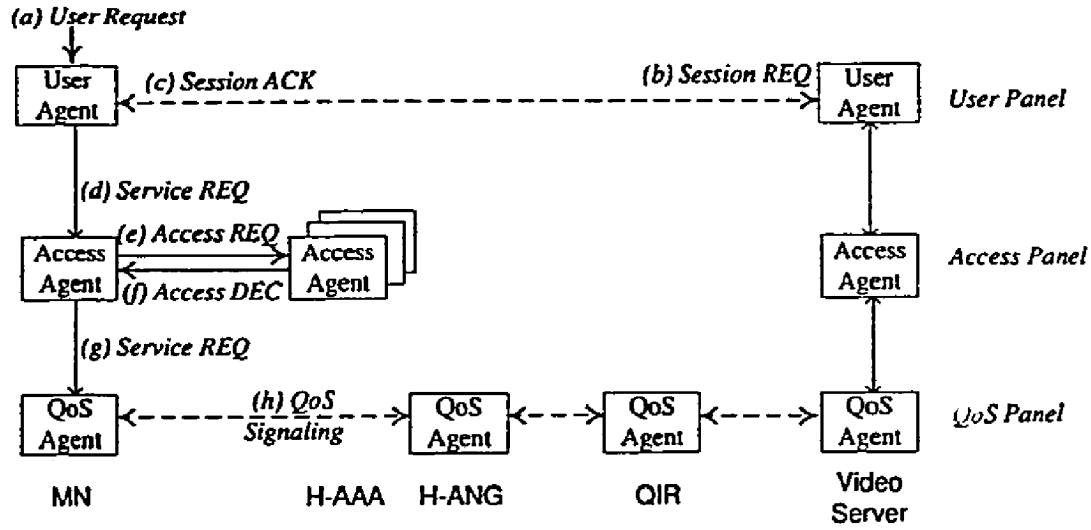


Figure 8: The work flow of the framework on the event of session establishment

4.3.2 Scenario Two: Inter-administrative Handoff

If the mobile user is roaming from Domain 1 to Domain 2 and the two domains are administrated by different ISPs, an inter-administrative handoff initiated by the host occurs. Figure 9 depicts the general control communications triggered by this event. The QoS Agent at the host is notified of this event (a) by the mobility management entity. It asks the Access Agent whether the host is allowed to enter the new network (b). The interactions between Access Agents at the host and the access servers results in an access decision (c,d,e). If the decision is positive, QoS signaling (g) that is necessary to update resource allocation along the new route is triggered by the QoS Agent at the host. If the decision is negative, there are two possibilities. The first one is to select another candidate network and repeat the procedure above. If there is no other candidate network, the alternative action is to notify the end user that the network is inaccessible.

From the above analysis of the work flow of the framework, it is easy to identify the potential points of enhancement on QoS control communications if fast handoff with minimum service is desired.

- Access Agents making the access decision more quickly;

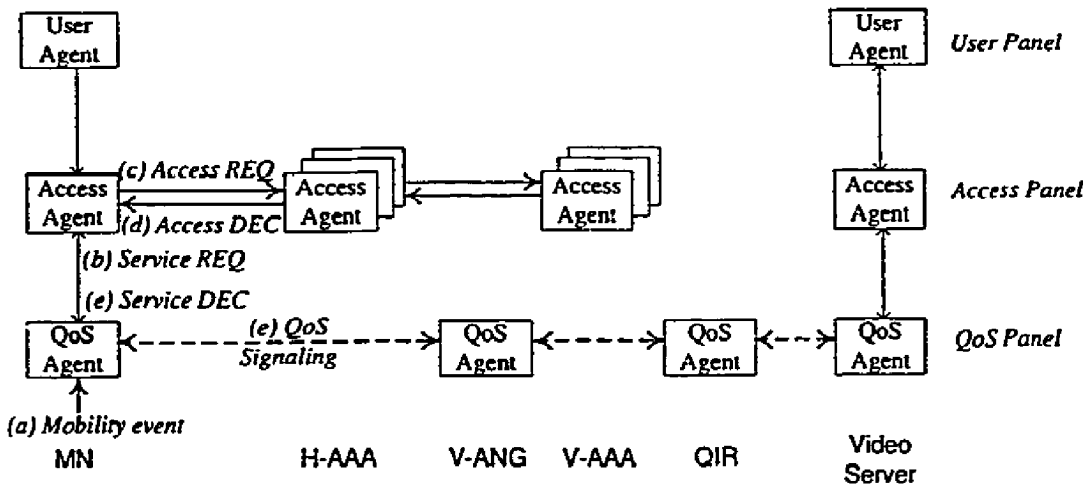


Figure 9: The work flow of the framework in the event of inter-administrative handoff

- QoS Agents updating resource allocation more quickly;
- More efficient local communications between Agents.

4.4 Chapter Summary

In this chapter, we have proposed a multi-panel framework for QoS control communications in heterogeneous network environments that meets the requirements specified in Chapter 3. Three panels are currently defined in the framework. The functionalities of each panel and the interactions among the panels are described. The case study provides the work flow responding to two typical scenarios of this framework. In later chapters, we will see that this thesis has been focused on the functionality of QoS Panel and its reactions to network events.

Chapter 5

Handoff Signaling in the QoS Panel

5.1 Overview

In this chapter, we propose a novel signaling scheme in the QoS Panel to deal with IP mobility events. Our goal is to promote the efficiency of intra-panel communications in the QoS Panel upon MIPv6 handoff, which is a peer-event-dominated scenario. According to Section 4.2.2, an efficient signaling scheme promotes efficiency of intra-panel communications. In a lower-positioned panel such as the QoS Panel, the improvement is particularly critical to the entire control communications framework.

We analyzed the major problems that cause the current QoS signaling mechanims fail to support seamless end-to-end QoS provisioning during a layer 3 handoff. Our solution relies on the functionalities of QoS Agents and their interactions with other entities. The base of the signaling scheme is that critical information for QoS update is delivered by mobility management messages. Combined with our QoS update algorithms, this signaling scheme is able to reduce the QoS configuration delay significantly. A set of comprehensive simulations has been done to evaluate the performance of the proposed scheme.

5.2 Problem Description

No matter which QoS model is being used, a similar situation is encountered when MIPv6 handoff is performed in an ongoing session between communication peers, at least one of which is a mobile node. Service level negotiated and established before handoff cannot be maintained during handoff and even after handoff. The essence of the problem is that MIPv6 handoff brings two critical changes to end-to-end QoS provisioning schemes.

- Change of CoA of the MN. Service differentiation provided by QoS control and resource management modules is normally based on the destination address and source address of an incoming packet. Therefore, when one of the addresses is changed, QoS configuration has to be updated as fast as possible to reduce the number of packets that cannot be properly treated and the period that these packets are affected.

- Change of network resource condition. Moving into a different access network may result in change of network capacity, which directly affects QoS provisioning to the ongoing session during and after handoff.

In this chapter, we focus on the first factor above. The second factor will be studied in the next chapter.

We take the two QoS approaches that have been standardized by the IETF as examples for studying the influences of address changes on QoS configuration update during handoff. For the original RSVP, it is well known that a new session has to be initiated after the MN changes the CoA because the session identifer is changed too. In a Diffserv region, the policy entry at the ingress edge router needs to be updated for the packets forwarded from HA to the new CoA of the MN and the packets sent directly from CN to MN.

The above updates inevitably result in long configuration delay and dramatic service degradation. We identify the most significant factors that affect the QoS

provisioning performance in terms of QoS update latency. In addition, we describe the cause of each factor and the required information to deal with each factor.

- The latency of delivering the new CoA information to the correspondent node and other proper entities.

  This latency is caused by the necessity of delivering the new CoA to the CN so that the subsequent packets will be sent to this new destination address. The required information is the new CoA and the home address of the MN.

- The latency of configuring QoS at the involved routers.

  This latency is caused by notifying the relevant nodes to make proper QoS configuration. The required information includes identification of the ongoing session, the new CoA, and the QoS parameters.

As long as the required information can be obtained by the proper entities as quickly as possible, the involved routers can take proper actions such as admission control and resource allocation. If the desired service level can be maintained along the new path segment, the service disruption due to handoff can be reduced to a minimum. In case the resources at the new places are not sufficient, proper adjustment has to be made, either degrading the service level or allowing the MN to change its attachment point if service degradation is not an option. Therefore, the first principle of our approach for minimizing the MIPv6 handoff impact on QoS provisioning is to provide the required information to the proper entity in a minimal period, so that the correpsonding action results in minimal latency.

Another principle of our approach is to support heterogeneous QoS mechanisms simultaneously while taking maximum advantage of existing QoS mechanims. This is a desired feature for the future generation communications networks, as the MN has a good chance to move between access networks with different QoS strategies being applied. A unified solution will reduce the configuration costs of network providers.

5.3 The QoS Agent-assisted MIPv6 Handoff Scheme

The main idea of the proposed QoS Agent-assisted scheme (QAA scheme, [CA06b], [CA07]) is to integrate the QoS control communications into the process of MIPv6 handover, aiming at a shorter period of service disruption. QoS updates and adaptations are triggered by the proper QoS Agents before the MIPv6 correspondent registration completes.

To support heterogenous QoS models, QoS schemes should not be tightly coupled with mobility management protocols. The extensions made to the individual schemes mentioned in the literature cannot be applied to other schemes. As we introduced in Chapter 2, the dominant macro-mobility management protocol is MIPv6 in the future generation networks. So we present the proposed signaling scheme in the context of MIPv6. However, our scheme can work with other mobility management protocols as well, if the information critical to QoS update can be carried by mobility management messages.

We outlined the entities and their interactions in the QoS Panel in Figure 6 of Chapter 4. In order to optimize the performance of intra-panel communications when a handoff event occurs, we have to address the following issues in the proposed signaling scheme:

- What input does the QA take upon the event?
- How does the QA process the input?
- How does the QA notify other entities to perform QoS updates?

We will discuss these issues in the following text.

5.3.1 QoS Agent

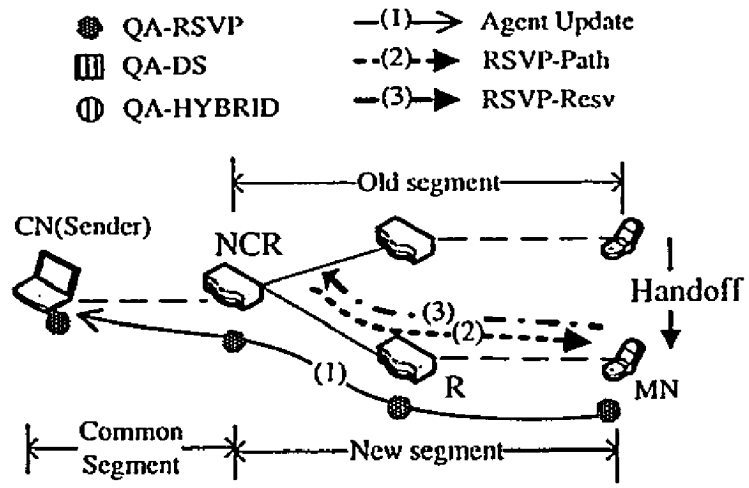
Generally speaking, the QoS Agent acts as a delegate for QoS schemes. The QA is able to extract *critical information* from the update messages of mobility management protocols and trigger local QoS updates immediately. The critical information consists

of two parts. One is the address information of the MN, used to identify the ongoing session and update QoS states accordingly. The other is scheme-dependent QoS information, used to establish QoS states or profiles in the new path segment and make dynamic adjustments in the old path segment.

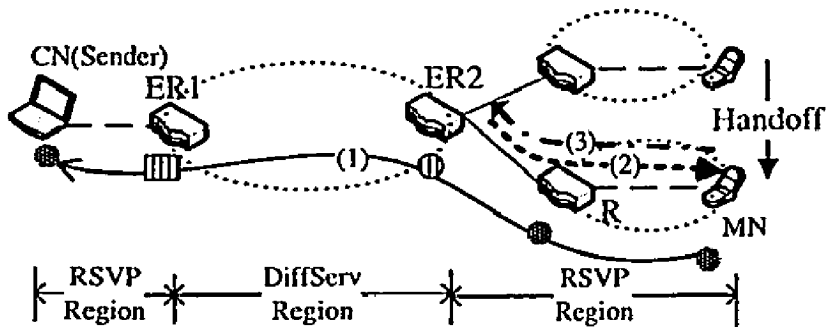
If a node, either a router, or an end host, participates in determining the QoS treatment level of a data session, we call it a critical node for the session. All the RSVP enabled nodes and DiffServ edge routers are typical instances of critical nodes in our approach. Potentially any node could be a critical node in a certain session, so a QA may be associated with each node in the network. However, a QA participates in the process of MIPv6 handoff only if it is activated. In the case of a stateful QoS model, the QAs in all the nodes that maintain QoS states for the ongoing session are activated. For example, if RSVP is used as the QoS signaling protocol in the access network, all the nodes that are RSVP-aware activate the QAs. In the case of a priority-based QoS model, only the nodes that decide the service level that the packets will receive activate the QAs. For example, if DiffServ is used in the core network, only the edge routers activate the QAs.

5.3.2 Features of the QAA Handoff Scheme

We classify QoS models into two categories: stateful and stateless. A stateful QoS model relies on a certain signaling protocol to negotiate service level and the participating nodes need to maintain QoS states for ongoing sessions. A stateless QoS model may or may not use signaling messages to negotiate service level and there are no QoS states maintained at intermediate routers for current sessions. While DiffServ is a typical stateless QoS model, IntServ/RSVP is a classic stateful QoS model because RSVP is mandatory in service negotiation. However, RSVP is not necessarily working with IntServ. In the following text, we use RSVP signaling as the example of a stateful QoS scheme and DiffServ as the example of a stateless QoS scheme, because these two QoS schemes are most extensively studied and widely deployed among others.



(a) Pure RSVP case



(b) DiffServ-RSVP case

Figure 10: Network model of QAA scheme

We introduce the major features of the QAA scheme with the assistance of Figure 10, which illustrates the network model and QA configuration with two cases: pure RSVP case and DiffServ-RSVP interworking case.

- *Support for heterogeneous QoS models.* Each QA is defined with a type, based on the local QoS model. The QA invokes the local QoS handler based on its type. In order to properly trigger actions of the local QoS entity, the QA obtains necessary information from the QoS handler and maintains soft states for involved traffic flows.

For stateful QoS models, the QA maintains a QoS State Block (QSB) list indexed by session identifier. Each QSB entry consists of the most critical information for a session in order to determine further actions, such as sender address, receiver address, flow id, and QoS request. QA-RSVP is a typical example of a QA for stateful QoS models. Case (a) of Figure 10 shows the configuration of QA-RSVP in a pure RSVP-signaled network. RSVP signaling is performed end-to-end without traversing other types of QoS region. All the RSVP-aware nodes are configured with a QA-RSVP.

For stateless QoS models, the QoS Agent maintains a priority list indexed by the same information that would be used to determine the processing priority of a packet. In DiffServ, this information refers to the source address and destination address of the packet because packet classification is based on it. Case (b) of Figure 10 illustrates a network model with interworking of DiffServ and RSVP, where RSVP signaling is transparent in the DiffServ region. The ingress edge router (ER1 in Figure 10) of the DiffServ region in this case is associated with a QA-DS.

If the local node is a transition node between two QoS models, the attached QA is defined as QA-HYBRID, which interacts with both QoS handlers within the same node, and maintains two sets of soft states accordingly. It extracts QoS data for the previous QoS model and fills in the proper QoS data for the next model. In case (b) of Figure 10, there is a QA-HYBRID configured at the egress ER (ER2) of the DiffServ region, which triggers RSVP negotiation in the RSVP region of the new access network.

New types of QoS agents can be easily defined to support other QoS models if desired. The critical information set and the trigger to local QoS updates need to be defined for each type of QA. The locations of QAs, i.e., the critical nodes for the specific QoS model, also need to be defined.

- *Minimum changes to MIPv6 and existing QoS schemes.* We extended the

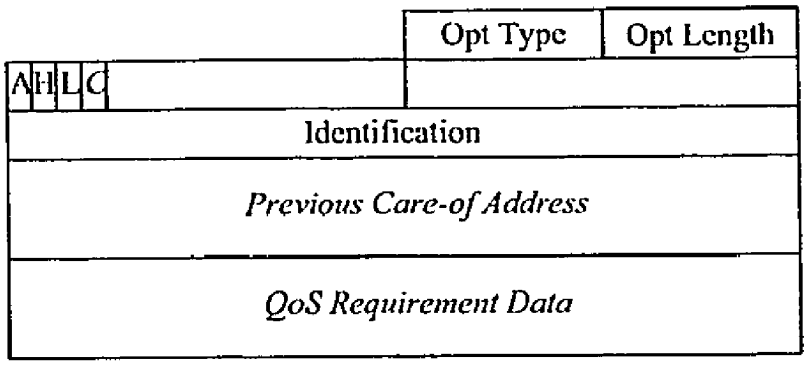


Figure 11: QoS Agent Option of Agent Update message

MIPv6 BU message to include critical information. The new message is named Agent Update (AU) message, in order to leave processing of the original MIPv6 BU messages unchanged. AU messages are essentially the input to the QAs at critical nodes. The AU message format is exactly the same as the BU message format, except that it carries some additional fields in a hop-by-hop QoS Agent Option header, as shown in Figure 11.

The functions of the three new fields in the message are explained below.

- *Flag 'C'.* This flag indicates whether the NCR has already processed this packet. It is initialized to zero. When the NCR identifies itself, the QA will set the flag and return it to the MA. When the later routers see this flag, they will not trigger any QoS updates down the data path. The details are presented in the pseudo code segment.
- *Previous Care-of address.* With this address information and the destination address of the message packet, the QA looks up the corresponding entry in its soft state list.
- *QoS Requirement Data.* Normally different types of QAs need different sets of information.

As to the existing QoS schemes, no explicit extensions are required in the QAA

scheme. It is the QA that triggers the adequate update functions provided by the QoS schemes. The details of how QA and the QoS control module interwork in different nodes are described in Section 5.3.3.

- *Low signaling overhead.* For those QoS models that normally require explicit signaling protocols to exchange QoS data, the QAA handoff scheme localizes QoS updates brought by the handoff. Thus, end-to-end signaling message exchanges are avoided. Taking RSVP as an example, the QAA scheme can localize QoS updates to the same extent as the FT scheme [SLSK00] can. In both schemes, the establishment of new RSVP states depends on the message exchanges along the path segement from the nearest common router to the MN. However, the FT scheme needs extra messages to notify the NCR of the QoS update request, while our QAA scheme does not need any assistant messages.

  For those QoS models that do not depend on signaling messages to establish QoS profiles, the QAA handoff scheme makes use of mobility update messages to carry QoS data. These messages actually perform QoS context transfer to minimize service disruption, as suggested in [JZM02] for DiffServ, without introducing additional signaling overhead.

- *Adaptability to changes of network conditions.*

  QoS Requirement Data carries information for the local QoS entities to perform admission control. The detailed information is QoS scheme dependent. It could be bandwidth requirements, token bucket parameters, or a simple value of a DSCP. If the admission control succeeds, only the session identity of the existing QoS configuration will be updated. All the other parameters will not be modified after handoff. In this case, the mobility event is transparent to the ongoing session. In case the admission control fails based on the information carried in the AU message, the local QoS entities use the indication of the acceptable service level to automatically downgrade the service requirement. This increases the probability of successful handoff. It is especially useful when the

MN moves from a high capacity network to a low capacity network, such as vertical handoff from WLAN to UMTS. The details are given in Chapter 6.

5.3.3 Interworking Between QA and Other Entities

Basically AU messages are sent to the home agent and the correspondent nodes of the MN when it changes its CoA. In the case where packet forwarding is desired, the AU messages are also sent to the node that is going to forward the packets to the MN before binding registration is complete.

The delivery of the AU messages is taken care of by MIPv6. To save the extra latency due to message processing, the AU messages are only processed at a QA-activated node. Other nodes will simply ignore them and forward them to the next hop. The QoS handler (QH) is the actual module that performs QoS configuration at a node. Within a QA-activated node, the QA bridges the MA and the QH so that the two entities can share critical information in order to enable fast QoS handoff. The QA itself does not perform any QoS configuration. Instead, it triggers necessary actions of the QH based on the information obtained during stable communications. Figure 12 illustrates the interworking between QA, QH and MIPv6 entity, which is called Mobility Agent (MA) in the illustration, when the QoS scheme is stateful. For the stateless QoS scheme, the situation is much simpler, as only the edge routers are critical nodes, and there is no end-to-end signaling required to negotiate or update QoS configuration.

The numbers in Figure 12 indicate the order of the events. Upon receiving an AU message, the MA checks if the local QA is activated. If no, the MA simply forwards the message to the next hop. If yes, two operations are performed by the MA. First, it sends the message contents to the local QA, which checks if the node is the nearest common router. If yes, QA sets flag 'C' in the message and returns it to the MA. Then the MA forwards the message to the next hop, and the QA triggers QoS configuration along the new path segment according to the QoS profile extracted from the message contents. If the current node is not the nearest common router, the message will be

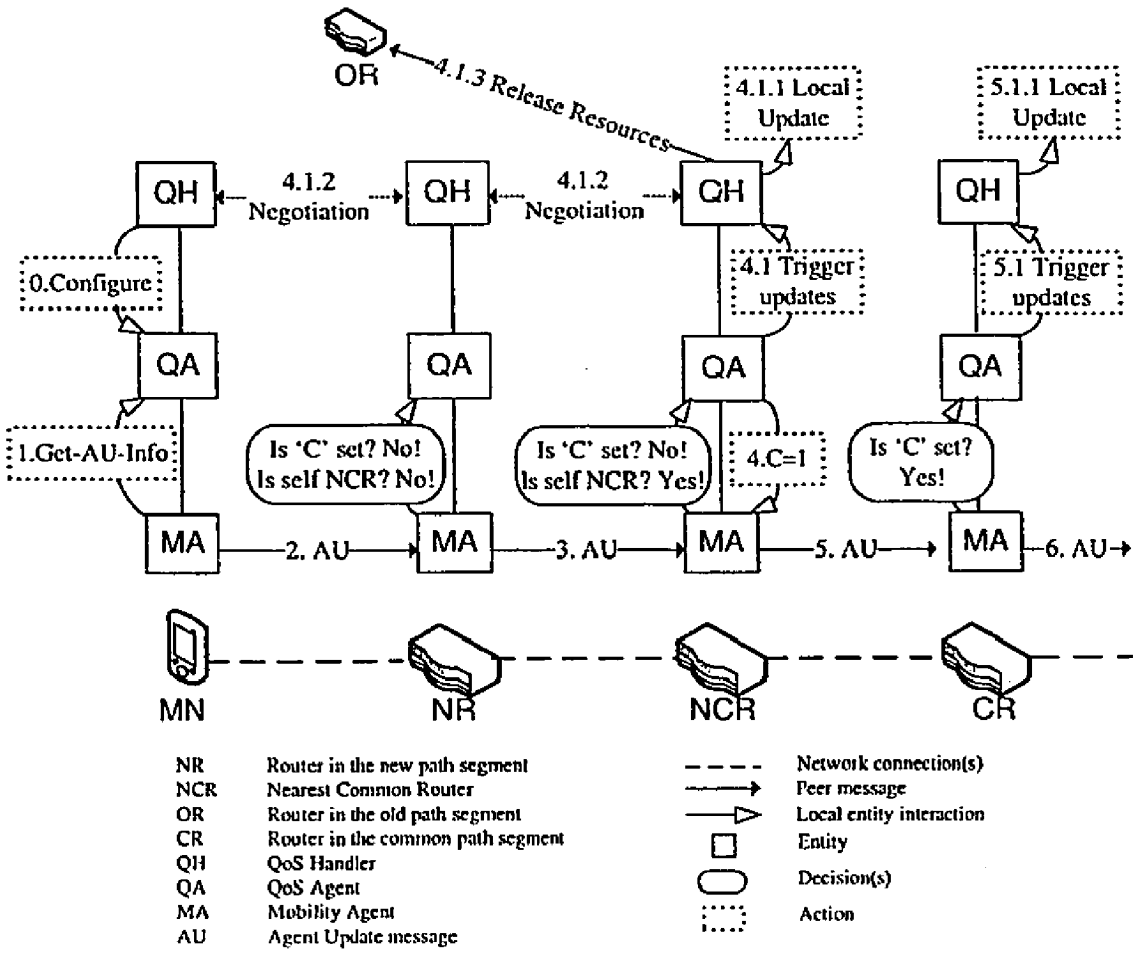


Figure 12: Interworking of QA, MA and QH

forwarded to the next hop by the MA without modification. No QoS configuration will be triggered.

Finally the AU reaches the CN, which encapsulates the subsequent packets and sends them to the new CoA. Thus, the QoS configuration along the route is updated with the location update.

5.3.4 Message Processing

The following segment of pseudocode illustrates the procedures for QA-RSVP and QA-DS to process AU messages. QH refers to QoS Handler, which is the local QoS entity that actually handles QoS configurations. HA, PCoA and NCoA refer to the

home address, the previous CoA, and the new CoA of the MN respectively.

```
QA-DS::recv_AU(msg) {

 //Get address info from msg as search index
 (dst, src) = (PCoA, CN);

 if ((Priority List != NULL) &&
 ((dst,src) in Priority List)) {
   /*Found the entry*/
   QH::update_entry(src, NCoA);
   QH::add_entry(HA, NCoA);
   return;
 }
 else {
   /*New entry*/
   QH::add_entry(src, NCoA);
   QH::add_entry(HA, NCoA);
   return;
 }
}

QA-RSVP::recv_AU(msg) {

 old_session_id = (PCoA,CN);
 if (old_session_id in QSB list) {
   if (msg.flag.NCR == 0) {
   /*NCR*/
     msg.flag.NCR = 1;
     /*QH:update_QoS*/
```

```
      send_path_msg -> NCoA;
      update_session_id(NCoA);
      send_path_tear_msg -> oldBS;
      return;
    } else {
    /*Common router*/
      update_session_id(NCoA);
      return;
    }
  } else {
    /*New router; do nothing.*/
    return;
  }
}
```

QH receives a trigger from the QA and performs QoS updates immediately. This localization is especially effective for a stateful QoS scheme when signaling message exchanges are normally required to accomplish state updates. Taking RSVP for instance, all the involved routers are put into three groups by the QAA scheme.

- New routers: These are the routers in the new data path segment. The QAs located at these routers do not trigger a Path or Resv message.
- Common routers: These are the routers in the common path segment. They update session id based on the previous CoA and the flow id carried in the AU message;
- The nearest common router (NCR): Once a router identifies itself as the NCR, the QA has three things to do:
  - Trigger the QH to send a Path message towards the NCoA of the MN to establish a reservation at the new routers;

- Update local session id based on the PCoA and flow id in the AU message;
- Send a PathTear message towards the previous BS to explicitly remove the reservation states on the old path segment;

Thus, QoS signaling messages do not need to be propagated end-to-end. When the NCR is closer to the new location of the MN, the reduction of message exchanges and QoS configuration delay is more significant.

For DiffServ, the edge router in the common path segments simply updates the policy entry as shown in the pseudocode. The destination address of the AU message and the previous CoA carried in the AU message will be used to locate the policy entry to be updated. If the edge router is located in the new path segment, it uses the QoS information carried in the AU message to configure the policy entry for the new packets.

During this period, the packets could be forwarded by a node that is closest to the common path segment along the old path. If the MN moves from one access network to another, the QA located in the border router of the old access network will intercept the packets and tunnel them to the new CoA. The knowledge of the new CoA can be obtained from the AU message that is forwarded to the old access network.

During all the above discussions, we assume that MN is a receiver of data. If the MN is a sender, situation is much simpler. First of all, the sender always knows where to send the data traffic during handoff. Thus, there is no unavailable period as in the reverse case. Furthermore, when AU messages are forwarded from the MN towards the CN, QAs associated with the routers along the new path segment can easily obtain the critical information set and trigger a QoS update immediately. Take RSVP as an example, AU messages forwarded towards the CN actually perform the function of Path messages. New routers can establish a session immediately and the NCR can send Resv towards the MN to establish new QoS states promptly.

5.4 Performance Evaluation

5.4.1 Comparisons with Other Proposals

We compare our QAA scheme with the FT scheme [SLSK00] and the QoS Framework for MIPv6 [CK01] in this section, because all the three mechanisms propose to update QoS configuration before binding update is complete.

The purpose of the QAA scheme is to minimize the QoS configuration delay caused by handoff events, so we do not consider the situations when QoS updates are triggered by other causes, such as router change, network congestion, etc. For the ongoing sessions that experience a handoff event, the routers along the data path before handoff must have been satisifying their needs. Therefore, after handoff, the routers along the common path segment should be able to continue the service level, which brings the significance of the NCR, especially when the percentage of common routers is relatively high. Both our scheme and the FT scheme take advantage of the importance of the NCR, while the QoS Framework for MIPv6 does not distinguish the NCR from other routers.

Both our scheme and the QoS Framework for MIPv6 are able to support heterogeneous QoS schemes, while the FT scheme is specifically designed for the RSVP signaling protocol. However, the QoS object option header defined in the QoS Framework for MIPv6 carries QoS data for all possible QoS mechanisms, which significantly increases the length of the binding update message. Our scheme requires the AU message to carry only the QoS data that is necessary for the current QoS model. The QA associated with the MN determines the QoS data when the AU message is created, and the QA in the transition node translates the QoS data between the two QoS models. This reduces the bandwidth consumption due to message delivery.

When the QoS requirements cannot be met along the new path segment, both our scheme and the QoS Framework for MIPv6 have a QoS adaption mechanism to enable the QoS renegotiation, while the FT scheme will have to rely on the RSVP protocol to negotiate the new resource allocation, which increases the QoS handoff

delay.

In summary, the QAA scheme has all the features of promoting the significance of NCR, supporting heterogeneous QoS schemes in a lightweight manner, and adapting to network capacity changes.

5.4.2 Simulation Configurations

In order to evaluate the performance of the QAA scheme quantatively, we have performed a set of simulations with the Network Simulator-2 (NS-2 [LBN06]). We implemented the QAA MIPv6 handoff scheme based on the MOBIWAN [Eri01] module, RSVP/ns [Gre98] module, and the standard Nortel DiffServ module in NS-2.28.

Multiple groups of simulations have been done with randomly generated network topologies. With the same configurations, which are introduced below, different topologies did not show any significant impact on the simulation results. Therefore, we present the results for one group of the simulations. Figure 13 shows the simulation network topology. As we can see, the simulations have been carried out in a small-scaled wireless-and-wired network, which is sufficient for the purpose of demonstrating the performance gain of the QAA scheme. The core network consists of three routers, Node 0, 1, and 2. Node 11 is the mobile node and Node 3 is the corresponding node. All the simulations assume that the MN is the receiver and the CN is the sender of data traffic. The other nodes are grouped into three access networks, one base station in each access network. The wireless link between the base stations and the MN is a simple 802.11 channel.

During any handoff period, two CBR background traffic flows saturate $Link_{3,0}$ and $Link_{0,1}$ to generate the condition of congestion. Four cases of link congestion are described below:

- C1: $Link_{3,0}$ is congested and $Link_{1,0}$ is not;
- C2: $Link_{0,1}$ is congested and $Link_{3,0}$ is not;

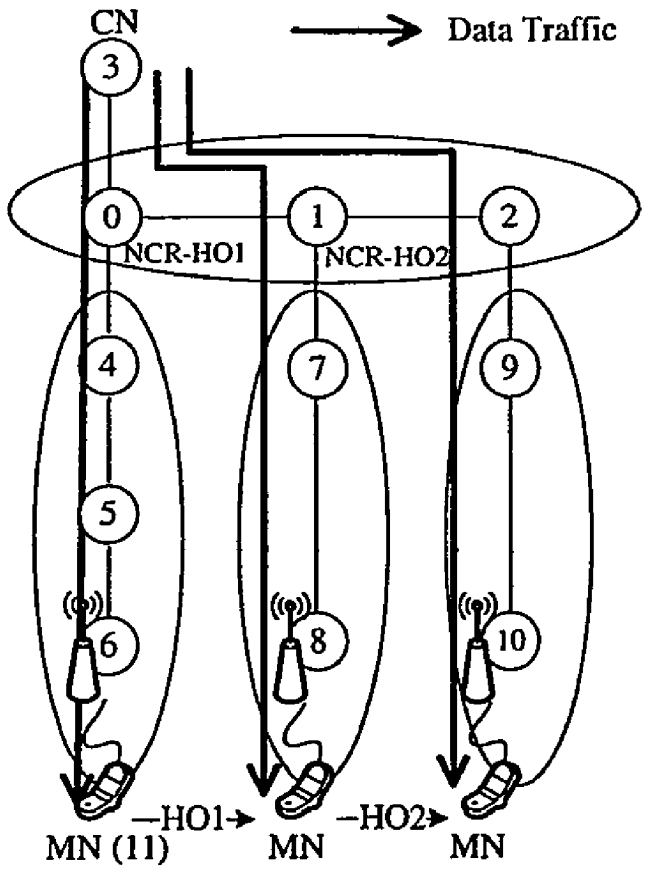


Figure 13: Simulation network topology

- C3: Both $Link_{3,0}$ and $Link_{0,1}$ are congested, and the bandwidth reserved for RSVP messages is sufficient;
- C4: Both $Link_{3,0}$ and $Link_{0,1}$ are congested, but the bandwidth reserved for RSVP messages is insufficient.

In all the four cases, the MN (Node 11) follows the same movement pattern. As the result of movement, the data traffic originating from the CN (Node 3) to the MN experiences two handoffs during the entire simulation time, handoff 1 (HO1) at the 30th second and handoff 2 (HO2) at the 60th second. The NCR of HO1 is closer to the CN side, and the NCR of HO2 is closer to the MN side.

The major parameters are listed in Table 1. All other parameters are set to the default values provided by NS-2.

Table 1: Simulation configuration parameters

Parameter	Value
Simulation time	80 second
Congested link capacity	1 Mb
Background traffic rate	800 kb
Data traffic rate	400 kb
Core network link delay	10 ms
Access network link delay	5 ms
Wireless channel capacity	2 Mb

5.4.3 Simulation Results

Our focus is on minimizing service degradation caused by QoS control re-configuration. Therefore, link layer handoff latency is ignored when evaluating the performance.

We evaluate the performance of the QAA scheme in terms of supporting the following two schemes of QoS control:

- Q1: RSVP as QoS signaling protocol in the entire network;
- Q2: Diffserv in the core network and RSVP in the edge network.

The first criterion of performance evaluation is the difference between the QoS configuration delay (QCD) and the maximum allowed delay (MAD). The QCD is defined as the latency starting from the point when a MN obtains a new CoA until QoS control modules along the new path are properly configured. The MAD is defined as the time duration from the instant when the MN sends AU or BU message to the CN to the time when the first packet destined to the new CoA arrives at the routers along the new path. Both delays are measured at each critical node. If the QCD is greater than the MAD at some critical node, the re-directed packets will be treated as best-effort traffic when they arrive at that node. So, as long as the value of the QCD does not exceed that of the MAD, the packets destined to the new CoA of the MN can be properly treated.

In addition, we have also taken *QoS configuration delay (QCD)*, *end-to-end packet delay*, and *packet loss* as major measurements. We compared the proposed QAA

scheme with the "simple" model in terms of these three measurements. The QAA is designed to support both Q1 and Q2. The simple model for Q1 is the simple interworking of RSVP and MIPv6. The CN receives the BU message from the MN during handoff, and sends refresh Path message immediately to update reservation states. Meanwhile, packets are being sent to the new location of the MN. Once the Resv message is returned to the CN, the packets can receive the desired QoS along the new path. The simple model for Q2 adopts the same interworking scheme as in the simple model for Q1 in the RSVP region. In the DiffServ region, no additional process is applied.

The details of simulation results are presented below.

Delay Differences For Q1, the two delays are measured at each RSVP-aware router. Figure 14 shows the results for all the four scenarios. In all the figures, QCD denotes the QoS Configuration Delay, and MAD denotes the Maximum Allowed Delay. The figures in the left column present the delays measured at each router involved in HO1, and the right column figures present the delays measured at each router involved in HO2. For both events of handoff, the QCD is the lowest at the NCR, because the QoS update at the NCR is performed immediately after the node receives the AU message. On the other hand, the new routers have to wait for the new Path message from the NCR to trigger the update process and the arrival of new Resv messages from the MN to complete the process.

Let $T_{msg(n,m)}$ denote the message delivery and processing time of message msg from node n to node m. Without affecting the result of analysis, we assume that link delay of each hop is equal and set to unit value. Then $T_{msg(n,m)}$ equals the hop distance from node n to node m, denoted as d_{n-m}. The relative positions of the critical nodes can be found in Figure 10. The QoS configuration delay at any new

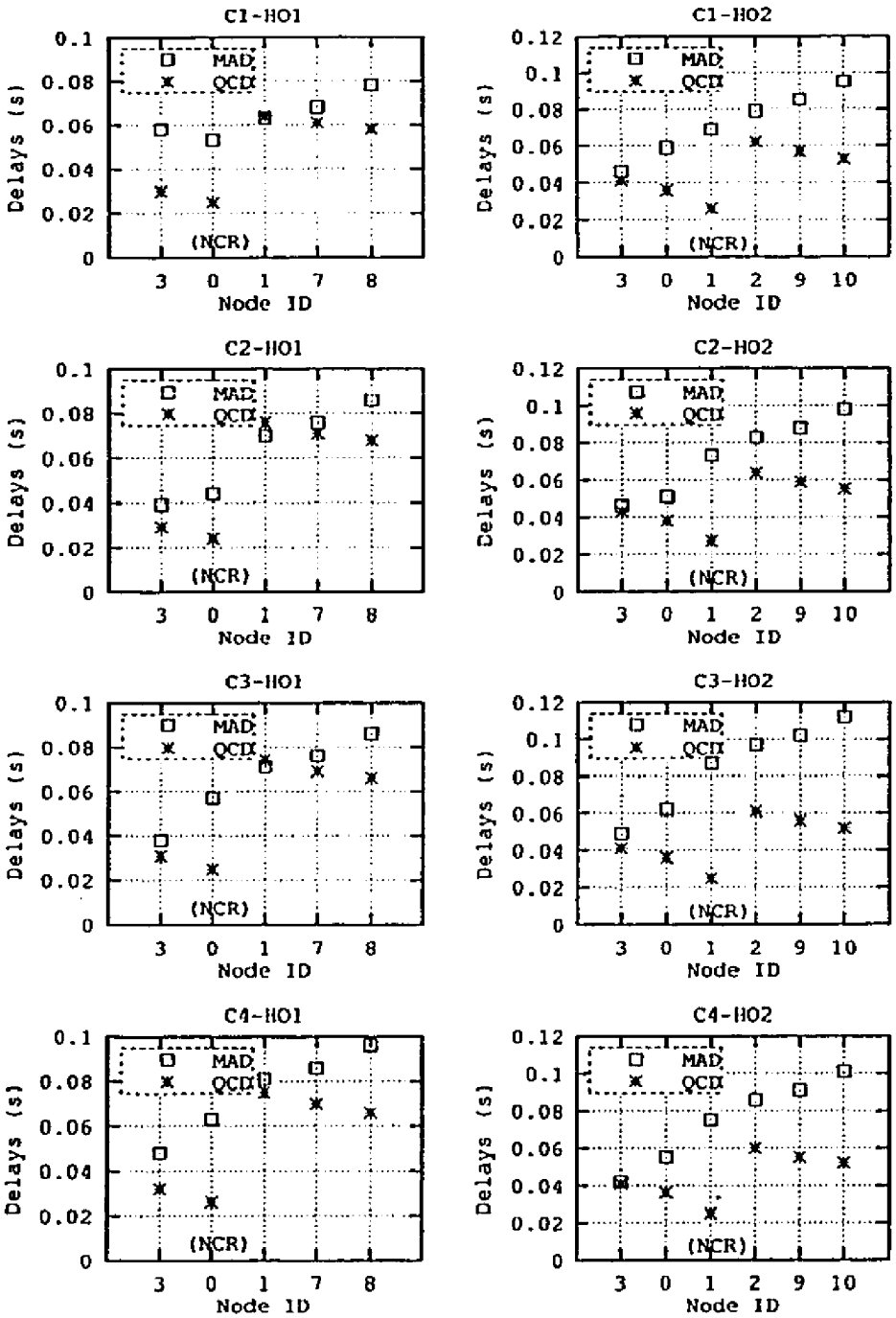


Figure 14: Delay differences

router R, $D_{QCD,R}$, can be computed as:

$$\begin{aligned} D_{QCD,R} &= T_{AU(MN,NCR)} + T_{Path(NCR,MN)} + T_{Resv(MN,R)} \\ &= d_{MN-NCR} + d_{NCR-MN} + d_{MN-R} \\ &= 2 * d_{MN-NCR} + d_{MN-R} \\ &= 2 * d_{MN-NCR} + (d_{MN-NCR} - d_{R-NCR}) \\ &= 3 * d_{MN-NCR} - d_{R-NCR} \end{aligned} \tag{5.1}$$

Given $d_{MN,NCR}$ is fixed for a certain handoff, the $D_{QCD,R}$ increases when $d_{R,NCR}$ decreases. In other words, the closer it is located to the NCR, the longer $D_{QCD,R}$ the router R has to experience. This explains why the highest value of QCD appears at Node 1 in HO1 and Node 2 in HO2, both are the closest nodes to the NCR during the corresponding handoff.

With HO1, the NCR (Node 0) is only one hop away from the CN (Node 3). As shown in Figure 14, the QCD point is above the MAD point at Node 1 with scenarios C1, C2, and C3, which means this node fails to update QoS configuration before the first packet destined to the new CoA arrives. In all the other nodes, the proposed scheme ensured that the QoS update is completed before the new packet arrives. With HO2, the NCR is Node 1, which is closer to the MN. In this case, all the nodes in all the scenarios are able to update the related configuration before the new packets come.

The different results obtained with HO1 and HO2 can be explained by the following analysis. For any router R in the new path segment, we can obtain the QCD, $D_{QCD,R}$ with equation (5.1). With the same assumptions on denotations, the MAD

at R, $D_{MAD,R}$, can be computed by:

$$\begin{aligned} D_{MAD,R} &= T_{AU(MN,CN)} + T_{DATA(CN,R)} \\ &= d_{MN-CN} + d_{CN-R} \\ &= d_{MN-CN} + d_{MN-CN} - d_{R-MN} \\ &= 2 * d_{MN-CN} - d_{R-MN} \\ &= 2 * (d_{CN-R} + d_{R-MN}) - d_{R-MN} \\ &= 2 * d_{CN-R} + d_{R-MN} \end{aligned} \tag{5.2}$$

By comparing the equations (5.1) and (5.2), we can obtain the delay difference at R, $D_{diff,R}$ by:

$$\begin{aligned} D_{diff,R} &= D_{MAD,R} - D_{QCD,R} \\ &= 2 * (d_{CN-R} - d_{NCR-MN}) \\ &= 2 * (d_{CN-NCR} - d_{R-MN}) \end{aligned}$$

Therefore, the QCD is not greater than the MAD at router R as long as d_{R-MN} is not greater than d_{CN-NCR}. In other words, the new routers will have QoS configuration ready for the new packets if NCR is closer to the MN than to the CN, which is usually true in the case of RSVP being used as the signaling protocol throughout the data path.

The worst situation here is that the NCR is so close to the CN that the localized signaling is almost the same as end-to-end signaling. In this case, the QCD will be greater than the MAD at the new routers that are close to the NCR and the first few packets will not receive the desired treatment from these routers. In reality, however, it is very rare that RSVP is used as the global QoS signaling protocol due to the limited scalability of the protocol. Instead, it is often applied in the access network. Moreover, the movement of the MN from one access network to an adjacent access network seldom results in a complete change or a major change of route from the CN to the MN. Therefore, the probability of the NCR being very close to the CN is very low. If the old access network and the new access network share the same edge

router, which is located at the boundary of the DiffServ region adjacent to the access networks, this edge router (e.g., ER2 in case (b) of Figure 10) becomes the NCR in the proposed scheme, because it is associated with a QA-HYBRID. Thus, the NCR is still closer to the MN than to the CN.

With both HO1 and HO2 in all the scenarios, the difference between the MAD at the CN and the MAD at the last hop before the MN is actually the latency of a packet being delivered from the CN to the MN without counting the last hop delay. These values reflect the level of service that the MN receives because the wireless channels are not congested in both scenarios. The detailed analysis on end-to-end packet delay is presented below.

From the results we have shown, it is obvious that our scheme is able to reduce the QoS configuration delay. The other factor we are interested in is the binding update delay. Since the AU message is intercepted at several nodes, the binding update delay should be a little longer than the standard BU message, which is not examined by the intermediate routers. This delay is dependent on the number of active QAs along the message delivery path. For a DiffServ domain, only the edge router touches the message data and causes extra delay. In the case where RSVP signaling is used, each RSVP-aware node needs to check the message data to get the previous CoA. As we discussed above, the number of RSVP-aware routers is not usually large in a global mobility scenario.

QoS Configuration Delay Recall that we define QoS Configuration Delay (QCD) as the time duration from when the MN obtains a new CoA to when QoS states are installed or updated in all the involved routers along the new path. Figure 15 illustrates the QCD that the MN has experienced during HO1 and HO2 with QoS control mode Q1, i.e., RSVP is used as the QoS signaling control protocol. With the simple model applied, the QCD during handoffs is affected by two major factors. First, the congestions at $Link_{0,1}$ and $Link_{3,0}$ force the RSVP-Path messages to compete for resources with the data packets, since they are in the same direction. If the

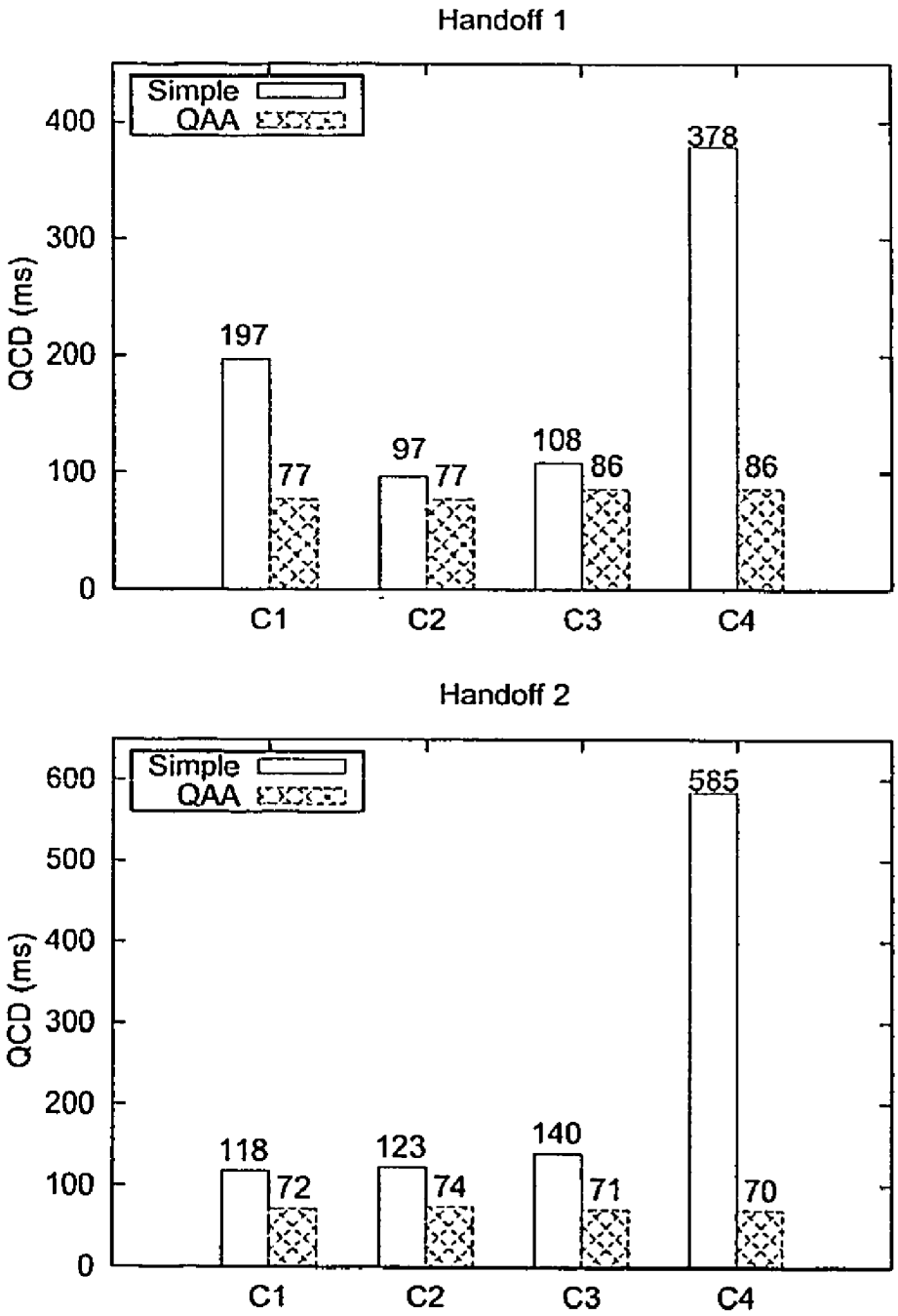


Figure 15: QoS configuration delay with QoS model Q1

bandwidth reserved for RSVP messages is insufficient, the QCD during handoff is greatly increased. As we see in Figure 15, the latencies during HO1 and HO2 in C4 are 378 ms and 585 ms respectively. Second, the QCD increases with the length of data path, because the RSVP messages have to be propagated through all the RSVP-aware routers along the path from the CN to the MN. As shown in Figure 15, the latency during HO2 in C3 (140ms) is larger than that during HO1 (108ms), although the two handoffs both experience link congestion.

The application of the QAA scheme greatly reduces the QCD during handoff in the following aspects. Corresponding to the first factor that affects the performance of simple model, QAA model enables *stable* QCD, no matter whether there is link congestion or whether the congestion affects the delivery of RSVP messages. As clearly illustrated in Figure 15, in the cases of single congestion and double congestions, the latency during HO1 and HO2 does not experience dramatic change. This is because of the localized QoS update, which results in a higher possiblity that the congested link is avoided for the QoS signaling messages. The latencies in C3 and C4 are greater than those in C1 and C2, because the link congestions have actually affected HO1, with its location closer to the CN side. The RSVP messages have to enter two WFQ queues in the cases of two congestions, which incurs a queuing delay of 9 ms. However, the queuing delay is not affected by the fact that bandwidth reserved for RSVP messages is insufficient, because the Path messages are propagated along the new path segment before the data packets come in the same direction. This is the second reason for the stable QCD with QoS model.

Corresponding to the second factor that affects the simple model, the QAA model enables *minimal* QCD. This is simply the result of localized QoS update. The closer to the MN side is the location of handover, the shorter the path segment that the RSVP messages travel, which reduces the QCD, especially in the case of local movements. In Figure 15, HO2 obviously encounters smaller QCD in all the cases of congestion, although the MN is further away from the CN after HO2.

As to the case of Q2, QoS updates in the RSVP regions present similar performance as in case Q1. The DiffServ region faces the problem of updating the policy entry with the simple model. Without any means of negotiation during handoff, the edge router finds that the packets sent to the new CoA cannot be properly marked, since the destination address has already changed. The result is that the packets are treated with lower priority than desired. This situation will last until the data transfer is finished, and the QoS configuration will never be completed unless additional negotiation is applied.

With the QAA model, the policy entry update is triggered when the QA-DS at the edge router intecepts the AU message. Subsequent packets are marked with the same DSCP as before handoff, and receive the desired service level. Therefore, QoS configuration in the DiffServ region does not incur additional latency when the QAA scheme is applied.

End-to-end Packet Delay Figure 16 illustrates the end-to-end delay of CBR traffic in the four cases of congestion, when QoS control scheme Q1 is applied. With the simple model, if the new location of MN causes the data packets and signaling messages to go through a congested link, there is a high peak in the end-to-end delay curve, as the one in (a) and (c), and the two peaks in (e) and (g) of Figure 16. This is because of the long QCD, as we have analysed in the previous section. When the QCD increases in scenarios S3 and S4, as shown in Figure 15, the bursty period of end-to-end delay also lasts longer, as shown in (e) and (g) of Figure 16. With the QA model, the bursty period does not exist. The end-to-end packet delay remains stable during all the handoffs in all the cases, as shown in (b), (d), (f), and (h) of Figure 16.

As to QoS control scheme Q2, we only measure the end-to-end delay in the congestion scenarios C2 and C3, as scenario C1 can not reflect the QoS control effect of the DiffServ region, and scenario C4 does not bring any new element into DiffServ control. With the simple model, at the 30th second when HO1 occurs, Node 0, as the ingress edge router, can not find a policy entry according to the MN's new CoA in its

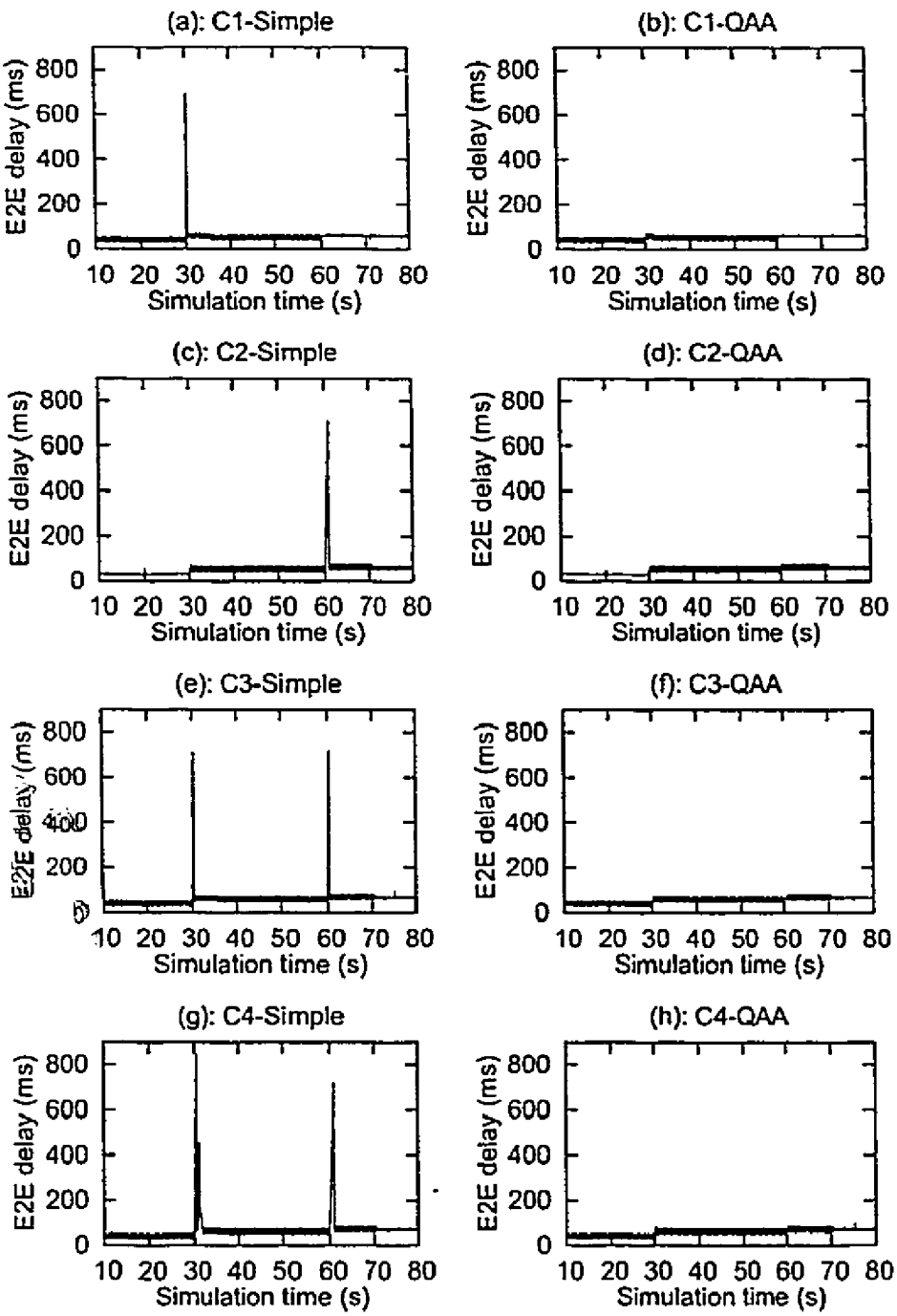


Figure 16: End-to-end packet delay with QoS model Q1

policy table. The result is that the subsequent packets destined to the MN's new CoA have to compete for resources with background traffic when they go through $Link_{0,1}$. Even if RSVP reservation is updated very soon at the RSVP-aware routers in the new access networks, the failure of DiffServ marking at Node 0 causes the end-to-end delay to be very large, as we can see in (a) of Figure 17, which shows congestion scenario C2. When the RSVP reservation can not be updated immediately, as in scenario C3, the situation gets even worse. The sharp peak at the 30th second in (c) of Figure 17 is caused by the long RSVP update latency due to congestion at $Link_{3,0}$.

The performance is largely improved when QAA handoff scheme is applied. QA at Node 0 is configured as type DS, which means it is able to update the local policy entry at Node 0 when the AU message is intercepted. As shown in (b) and (d) of Figure 17, there is no obvious increase in terms of end-to-end packet delay in both congestion scenarios.

Packet Loss There are two parts of packet loss observed in our simulations. The first part is caused by our assmuption of "break-before-make" handoff. When the handoff occurs, the MN first detaches itself from the previous BS. Before the binding update registration is complete, the packets sent to the old CoA will be dropped at the previous BS. This part is not the major focus of performance evaluation, because it is the duration of delivering binding update messages that contributes to the major part of the period during which the packets have to be sent to the old BS. This duration is relatively stable because binding update messages do not experience link congestion. However, we still record the first part of packet loss, as an indication of a valid simulation.

The second part of packet loss is due to inadequate QoS configurations. During handoff, long latency of QoS configuration causes some packets destined to the new CoA to be ill-treated. Furthermore, if the QoS update is not properly accomplished after handoff, the new packets will receive degraded service in a longer period. This part is our main concern.

Table 2 and Table 3 summarize packet loss measurement with Q1 and Q2 being applied as QoS control scheme, respectively. In both tables, Column L1 presents the first part of packet loss, and column L2 presents the second part of packet loss. Column L2 (%) computes the percentage of lost packets due to QCD. The measurement starts from the 5th second and ends at the 80th second, and the total number of packets during the measurement is 3749. The first five seconds of period is skipped to eliminate the influence of the initiation of RSVP reservation.

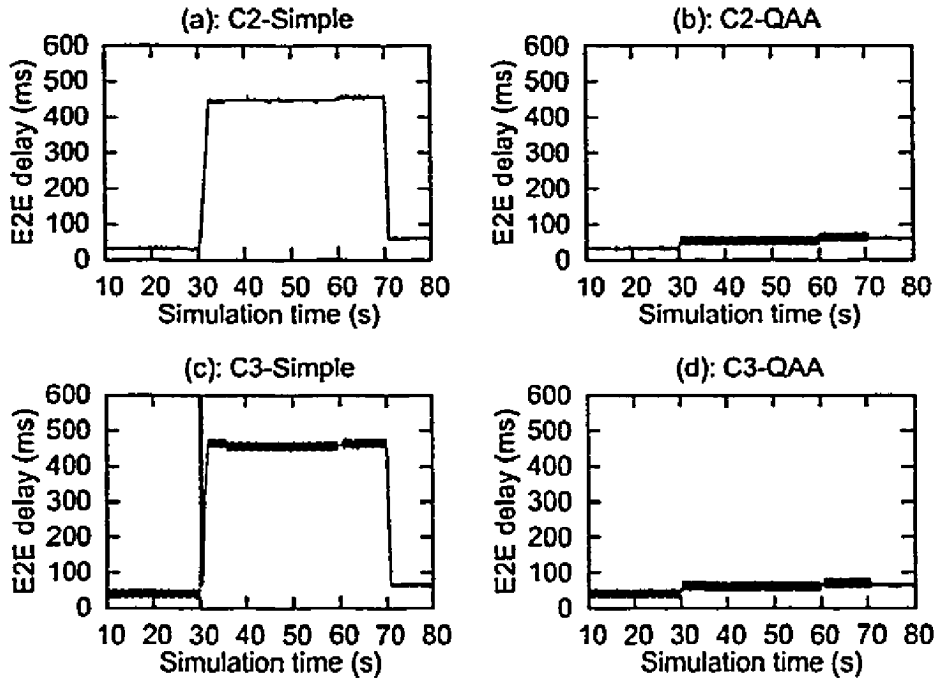


Figure 17: End-to-end packet delay with QoS model Q2

When all the nodes are RSVP-aware, as in the case of Q1, neither the simple model nor the QAA model incurs much packet loss due to QCD, as shown in column L2 and L2 (%) of Table 2. This is due to our simulation configuration. The interval of CBR traffic is 0.02s, and the travel time from the CN to the MN is around 0.04s to 0.06s. After the CN receives the binding update message and before RSVP configuration is updated, the first packets destined to the new CoA may experience long latency, but the number of such packets is not large enough to cause the packets to be dropped at $Link_{3,0}$ or $Link_{0,1}$, as long as RSVP messages can be delivered without extra

Table 2: Packet loss summary with QoS model Q1

	Simple			QAA		
	L1	L2	L2(%)	L1	L2	L2(%)
C1	16	0	0	41	0	0
C2	31	1	0.03	56	0	0
C3	48	1	0.03	57	0	0
C4	38	18	0.48	37	1	0.03

Table 3: Packet loss summary with QoS model Q2

	Simple			QAA		
	L1	L2	L2(%)	L1	L2	L2(%)
C2	41	764	20.38	41	0	0
C3	38	583	15.55	47	0	0

delay. However, as we see in case C4 when RSVP messages are having difficulty going through the congested links, the QAA model is able to eliminate most of the packet loss due to the delayed QoS update. This is simply because Path messages in the QAA model avoid the congestion.

The situation is quite different in the case of Q2. While column L1 in Table 3 presents similar results with both simple model and QAA model, many packets are dropped due to improper QoS configuration update after handoff without assistance of QA, as shown in column L2 and L2 (%).

Overhead Since extra data fields are introduced in the AU messages, it is inevitable that delivering and processing AU messages brings some additional overhead compared to the original MIPv6 BU messages. However, from the simulation results, it is obvious that there is only very minor impact. The major reason is that the cost due to the additional bytes is mostly compensated by the greatly reduced number of signaling messages brought by the localized signaling and the combined procedure of mobility update and QoS update.

5.5 Chapter Summary

In this chapter, we presented a QoS Agent assisted MIPv6 handover scheme as a solution of improving intra-panel communications upon a MIPv6 handoff event. The major contributions of this scheme include:

- The scheme smoothes QoS provisioning during MIPv6 handoff. By introducing a QoS Agent into the interworking of the mobility management and QoS management entities, we reduce the QoS configuration delay during MIPv6 handovers. The simulation results show that the QoS configuration delay is smaller than the maximum allowed delay when the NCR is not extremely close to the CN. In terms of end-to-end packet delay and data loss rate, the performance of the scheme remains stable in different cases of link congestion

- The scheme supports heterogenous QoS models. In our investigation and simulations, the stateful QoS model is represented by the RSVP signaling protocol, and the stateless QoS model is represented by the DiffServ model. By defining different types of QoS Agents as well as their interworking with the QoS control entities, the scheme is able to support multiple QoS models.

- The scheme is lightweight. It defines a new message type for MIPv6. The signaling overhead for MIPv6 is not increased due to the additional new message, Agent Update, as it replaces the original Binding Update message. Furthermore, carrying critical information for QoS Agent in the Agent Update message saves the extra signaling cost incurred in the original process of QoS updates.

- The scheme retains the features of existing QoS schemes, as no modification to any QoS module is required.

Chapter 6

QoS Signaling during Vertical Handoff

6.1 Overview

As described in the literature, cellular networks and WLANs, along with other wireless access networks, will both be included in the future generation networks. Multiple cellular/WLAN interworking architectures have been proposed in order to take advantage of the wide coverage and almost universal roaming support of cellular networks and the high data rates of WLANs. Based on the interdependence between the two access networks, there are three categories of UMTS-WLAN integration ([SJ+05], [VR+03], [XL+05]):

- *Tight coupling* : WLAN works as an SGSN emulation. Mobility from UMTS to WLAN or from WLAN to UMTS results in inter-SGSN handoffs and NO IP address change. Mobility largely follows UMTS mobility management.

- *Loose coupling*: A functional entity located in WLAN provides protocol interworking between WLAN and cellular networks for signaling and interworking between WLAN and IP core network for data traffic. WLAN and cellular networks are in different IP address domains. Mobility is only considered when the

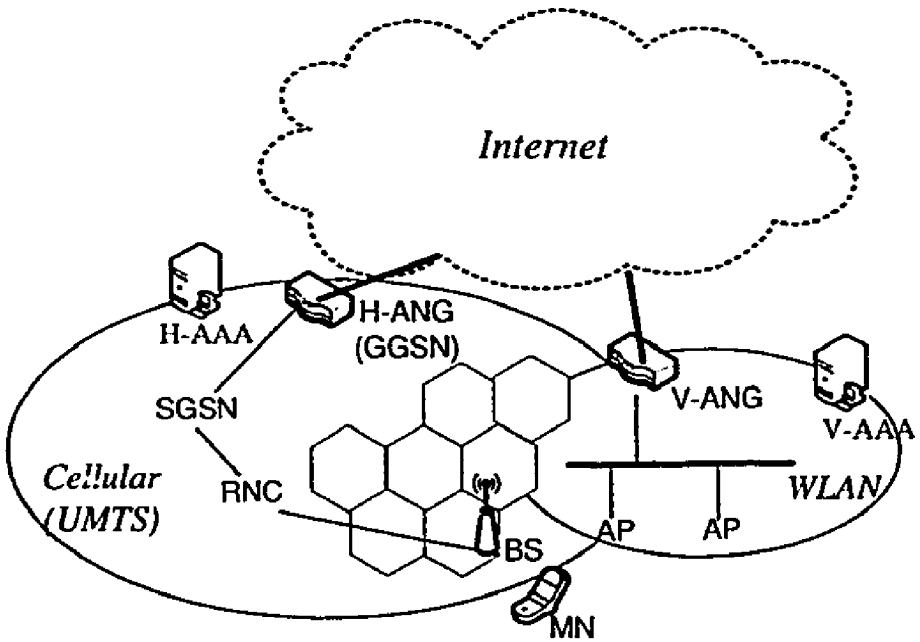


Figure 18: Peer integration of cellular/WLAN

user is not in an active session. Otherwise changing the IP address may result in loss of connectivity for the upper layers.

- *Peer coupling*: There is no direct connection between WLAN and UMTS, as shown in Figure 18. They are both connected to the Internet. Mobility management is provided by mobile IP protocols.

We choose peer coupling as the integration mode of our heterogeneous network model because it is the least complex from the perspective of mobility management. It requires the least changes to cellular and WLAN networks. The two types of networks remain independent, with the AAA linkage being the only new element in the architecture. Moreover, MIPv6 is used as global mobility management approach, which provides transparency of location update to higher layer protocols. This actually enables continuous provisioning of service to mobile users, which is critical for delay-sensitive applications such as voice over IP.

In this chapter, we propose a QoS control communication architecture in order to meet the new challenges of cellular/WLAN handoffs. Closely related to mobility management, QoS control during vertical handoff is based on the QA-assisted signaling scheme introduced in the previous chapter. Triggered by new challenges from the vertical handoff event, the intra-panel communications in the QoS Panel are enhanced by major components such as Service Mapper, QoS Agent-Enhanced Signaling (QAES), and Service Adaptor. We will present the architecture with details of the major components, and evaluate the performance of the architecture both analytically and experimentally.

6.2 Problem Description

In peer-coupling integrated networks, mobility between different types of access networks brings the following new challenges regarding to QoS provisioning during handoff.

- Potential QoS domain switch

  Different QoS approaches may be adopted in different access networks. RSVP may be used as IP QoS signaling protocol in the WLAN, while DiffServ is standardized as the IP layer QoS scheme in UMTS networks. It is also appropriate to assume that DiffServ is the major QoS architecture in the Internet core network. Therefore, a DiffServ-to-IntServ switch occurs when the MN moves from UMTS to WLAN. When the MN moves back to UMTS, an IntServ-to-DiffServ switch occurs. We refer to this type of switch as *QoS domain switch*.

  We argue that QoS scheme switch should meet the following requirements:

  - The switch procedure should be executed immediately after the MN's attachment to the new access point.
  - Related knowledge should be available for the prompt domain switch.

- Control signaling involved during the switch should not overburden the limited channel capacity of wireless networks
- The resource allocation for handoff sessions in the new domain should be available as soon as possible.
- The reserved resources in the old domain, if any, should be released within a reasonably short time duration.

- Potential service adjustment

  One of the major motivations of cellular/WLAN integration is that WLANs provide higher data rates than cellular networks do. It is reasonable for mobile users to expect better service quality for their data traffic when the MN moves from a UMTS network to a WLAN, and vice versa. In general, service differentiation due to capacity change between different networks is very common in heterogeneous networks. Therefore, fast service adjustment without adding too much system overhead during vertical handoff is desired in such environments.

  We argue that the following requirements should be met regarding service adjustment during handoff.

  - End users should not be involved during service adjustment. Service upgrade or downgrade should not interrupt the ongoing sessions.
  - Service adjustment should be completed as soon as possible after the change of attachment.
  - Flexible service adjustment should be allowed, i.e., a tolerable service level should be maintained all the time even if the expected service upgrade cannot be achieved.
  - Service upgrade to some handoff sessions should neither decrease the overall system performance nor result in poor accessibility to new sessions in the domain.

6.3 Related work

Various work has been done to address problems of QoS support in cellular/WLAN interworking environment. In [SJ+05], the authors discussed QoS-related issues such as resource allocation and call admission in cellular/WLAN interworking, taking network heterogeneity and user mobility into consideration. More issues are addressed in [XL+05], such as mobility management and handoff, mapping UMTS QoS to WLAN QoS and resource management in individual wireless access networks. Wang et al. proposed an adaptive QoS framework for cellular/WLAN interworking ([WM+05]). It uses a reservation-based QoS model, where RSVP is the end-to-end signaling protocol. With the help of user's degradation profile and an adaptation mechanism, the framework aims at providing end-to-end QoS with presence of cellular/WLAN handoff. In [MP07], the authors proposed a novel integrated architecture, called Integrated InterSystem Architecture (IISA), to enable the integration and interworking of various wireless networks with the introduction of a new entity, Interworking Decision Engine (IDE). This architecture supports AAA, QoS, and mobility management functionalities and aims at seamless roaming and service continuity.

However, not much attention has been paid to potential QoS domain switch during handoff, when end-to-end RSVP signaling is not always available. Service adjustment during QoS domain switch is hardly found in the previous works either. Based on the QAA scheme presented in previous chapter, we propose an enhanced QAA signaling scheme to accommodate new challenges during vertical handoff, which takes both QoS domain switch and service adjustment into consideration.

6.4 QoS Control Communication Architecture for Vertical Handoff

When an MN is roaming into a new access network, Candidate Access Router Discovery (CARD, [LS+05]) and interactions between Security Agents and related entities

ensure that the ongoing sessions with the MN can be accepted by the new access router. MIPv6 ensures that a new CoA is assigned to the MN. However, whether the desired QoS can be maintained in the new access network remains a question. Another problem is how to minimize QoS disruption during handoff.

In this section, we propose a generic architecture for a QoS-Agent enhanced vertical handoff scheme to address the above problems. The architecture is based on the signaling framework in Chapter 4 and the handoff signaling scheme in Chapter 5. Figure 19 shows the major components of the architecture. New components such as Service Adaptor and Service Mapper are introduced in the mobile terminal (MT) and the access network gateway (ANG). Peer QoS Agents interact with each other via QoS Agent-Enhanced Signaling (QAES). Configuration and activation of individual QA conform to the rules introduced in Chapter 5.

Based on the QAA signaling scheme during MIPv6 handoff, the MN sends AU messages to its correspondents and the home agent once the NCoA is obtained. Upon receiving an AU message, the QA at NAR extracts "critical information" and passes it to its Service Adaptor module. The critical information normally includes new and old address of the MN and the QoS profiles of ongoing sessions. The Service Adaptor then performs automatic service adjustment. The "adjusted service requirement" is then passed to the Service Mapper, which maps the IP QoS parameters to wireless QoS parameters. Afterwards, wireless QoS is enforced in the lower layers to actually provide desired service to mobile user traffic. Meanwhile, the QoS Entity at NAR is triggered to perform QoS updates via IP QoS signaling. QoS-aware routers along the new path segment perform resource allocation according to received information. Routers in the common path segment update resource allocation in order to complete the end-to-end QoS update. Our goal is that QoS updates can be completed before packets are delivered to the new location of the MN at most QoS-aware nodes.

The details of the major components and the interworking between them are explained in the subsequent text.

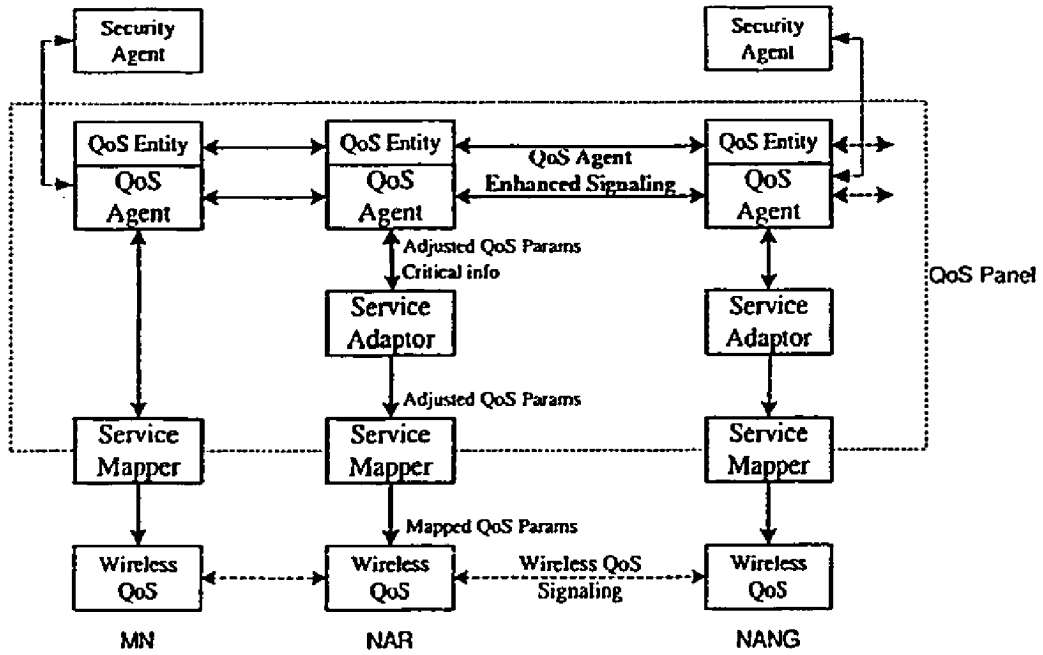


Figure 19: Architecture of QoS-Agent enhanced vertical handoff scheme

6.4.1 Security Agent

We assume that service upgrade will not cause unexpected session interruption due to any charging issue. The details of billing procedures are outside the scope of this thesis. The Security Agent located at the NANG interacts with the local AAA server to complete authentication and authorization procedures before the mobile node can receive or send any data. The details are also out of scope.

6.4.2 Service Mapper

IP QoS signaling in wireless access networks delivers QoS requests and traffic characteristics data between QoS-aware routers and hosts in the same manner as in legacy networks. In the lower layer, wireless signaling takes its own parameters. Thus, QoS mapping becomes one of the most essential components. Mapping QoS parameters properly can shield details of QoS enforcement in the lower layer from users. As a result, an end-to-end QoS signaling protocol, such as RSVP, can be applied uniformly among various access networks.

Table 4: Guidelines for IP-wireless service class mapping

Application type	IP QoS class	Wireless QoS class
Voice	IntServ GS/DiffServ EF	UC1/WC1
Video/streaming	IntServ CLS/DiffServ AF	UC2/WC2
Web browsing	IntServ BE/DiffServ AF	UC3/WC3
Email/file transfer	IntServ BE	UC4/WC4

The enforcement of QoS is a matter of implementation of wireless QoS. In UMTS networks, a Service Mapper should be located at the UE and at the GGSN. Once established, the PDP connection will be reserved for certain data traffic. In WLAN networks, the Service Mapper should be located at the corresponding gateway router. IEEE 802.11e specifies two modes of QoS schemes. Contention-based EDCA provides service differentiation with classification and prioritization, which can not guarantee fine-grained QoS. If RSVP is used to carry a guaranteed service request, the highest priority can be allocated to the data traffic. With polling-based mode, 802.11e signaling messages are used in the lower layer. Therefore, translation from RSVP FLOWSPEC to 802.11e TSPEC is required.

We propose a two-level QoS mapping scheme. The first level is *Service class mapping*. When QoS is implemented via classification and prioritization, it is sufficient to identify the service class of incoming packets to provide service differentiation. The accurate QoS parameters such as data rate, delay or jitter do not affect the actual service level. Table 4 presents the proposed guidelines for the mapping relationship among application type, IP QoS class and wireless QoS class. We use $UC_i(i = 1..4)$ to denote the four UMTS QoS classes, *conversational*, *streaming*, *interactive*, and *background*. We use $WC_i(i = 1..4)$ to denote the four 802.11e traffic categories, *voice*, *video*, *best effort* and *background*.

When fine-grained QoS is desired, it is necessary to deliver QoS parameters among routers and make the proper mapping between different schemes. Therefore, our second level mapping is *QoS parameter mapping*. Among all the parameters defined in different QoS schemes, we summarize the most common and decisive factors in Table 5, which can serve as a mapping guide. These three parameters are commonly

Table 5: The most common QoS parameters

UMTS	WLAN	RSVP	NSIS
Guaranteed Bit-rate	Mean data rate	rate	rate
Maximum Bit-rate	Peak data rate	peak rate	peak rate
Maximum SDU size	Maximum MSDU size	bucket length	bucket size

defined in all the QoS models. They define the most important characteristics of the data traffic. Furthermore, traffic engineering and resource scheduling algorithms rely on these parameters to work effectively. Therefore, the mapping relationship among different schemes enables continuous provisioning of desired service level to data traffic when the MN is roaming.

6.4.3 QoS Agent-Enhanced Signaling (QAES)

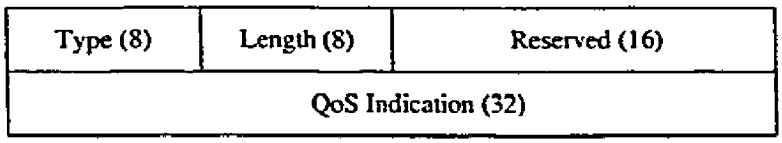
QAES refers to a group of protocols that are involved during vertical handoff. The major members of QAES include:

- IP QoS signaling

  Due to the heterogeneity of networks, IP QoS signaling may not be available all the time. However, there has to be an IP-level QoS signaling protocol to negotiate fine-grained QoS provisioning. RSVP is selected as the example of end-to-end QoS signaling in the following discussions. However, no matter which signaling protocol is being used, there is no modification to message exchanges and message definitions. The only change is that localized QoS update is triggered by the local QA of a router.

- MIPv6

  The mobility management protocol MIPv6 has been enhanced by defining a new message type, Agent Update, and several new flags to carry QoS-related information. Besides, we propose a new option in the Router Advertisement (RADS) message, *QoS Indication Option*, to announce the IP QoS capability in the new domain. Figure 20 illustrates the fields in the QoS Indication Option.



Type:	8 (to be considered by IANA)
Length:	1
Reserved:	This field is unused. It MUST be initialized to zero by the sender and MUST be ignored by the receiver.
QoS Indication:	32-bit unsigned integer. The indication of QoS scheme available in the current access network. Currently three values are defined:
1:	Parameterized QoS only;
2:	Prioritization-based QoS only;
3:	Both.

Figure 20: QoS Indication Option in RADS message

With this option, the new access point (NAP) of the MN includes the type of QoS scheme in the RADS message when the network prefix information is being sent to the mobile host. Upon receiving the RADS message, the QA-MN is able to prepare QoS-related information, which is placed in the QoS Requirement Data field of the Agent Update (AU) message.

- Enhanced QAA scheme

  The signaling scheme proposed in Chapter 5 is enhanced to meet the new challenges. *Dynamic information collection* is proposed for the QA to collect QoS-related information for stable traffic flows to and from the MN while the MN dwells in the access network before handoff.

  QA-MN collects QoS-related information dynamically before the MN switches the access network. The recorded information is used by the Service Adaptor to execute the service adjustment algorithm, as seen in Section 6.4.4. When handoff is taking place, the QA-MN prepares the information according to the QoS indicator in MIPv6 RADS and passes it to the NAP and other involved routers so that QoS updates can be triggered as soon as possible. Table 6 presents a guideline for dynamic information collection. If DiffServ is indicated as the new QoS scheme, recommended DiffServ Codepoints (DSCP) can be included

Table 6: Guidelines for dynamic information collection

New QoS scheme	QoS-related info	Source
DiffServ (EF,AF)	DSCP	IETF standard
DiffServ (Others)	Service type, rate	Application layer protocol
RSVP (new)	Service type, rate	Application layer protocol
RSVP (existing)	RSVP TSPEC parameters	Previous RSVP messages

if the service differentiation is offered using one of the IETF-standardized PHB groups, such as Assured Forwarding (AF) and Expedited Forwarding (EF). Otherwise, the service type and bandwidth information are obtained from application layer protocol message headers, such as SIP, of the previously received packets. If RSVP is newly enabled in the new access network, the same set of information can be obtained from the application layer protocol. If the MN was in an RSVP-enabled domain before handoff, RSVP parameters have been recorded by the QA-MN, and are thus available for indicating reservation requirements in the new domain.

6.4.4 Service Adaptor

In our architecture, there are two types of service adjustments during a vertical handoff. The first one is from coarse-grained QoS to fine-grained QoS provisioning or vice versa. Coarse-grained service differentiation is usually provided by prioritization. DiffServ and 802.11e EDCA are two examples of coarse-grained QoS provisioning schemes. Fine-grained QoS requires specification of QoS parameters and exchanges of QoS signaling messages. RSVP and the 802.11e HCCA QoS scheme are able to provide such fine-grained QoS. In the proposed architecture, this type of service upgrade is enabled by QAES (Section 6.5) no matter what QoS scheme is being applied in the old and new access networks.

The second type of service adjustment is triggered by the change of network capacity after handoff. A UMTS-WLAN handoff often results in a lower-capacity-to-higher-capacity upgrade. This type of upgrade is especially useful for delay-sensitive

or bandwidth-sensitive traffic flows, because they have stricter requirements on networks than others. Parameterized QoS control is mainly used for this type of application. Therefore, we assume that QoS signaling is present at least after service upgrade. We propose an automatic service adjustment algorithm for real-time traffic flows to receive continuous high quality of service during handoff.

Based on "Service Type" and other information provided by dynamic information collection, traffic flows are mapped to certain serving classes in wireless networks according to Table 4. We assume that a minimum serving bandwidth $B_{min}(i)$ and a maximum serving bandwidth $B_{max}(i)$ are defined for each class i.

The automatic adjustment algorithm is inspired by the algorithm proposed in [CA05]. Assume flow f is destined to the MN from the CN. Let A be the actual serving rate received by flow f before handoff. When the MN obtains a new CoA and generates an AU message, A is recorded in the QoS Requirement Data field. The NAP receives AU and intercepts A. Let $C(t)$ be the bandwidth available for flow f at time t. First, the NAP computes $C(t)$ with equation 6.4.1:

$$C(t) = C - \sum_{i=1}^{n} B_{avg}(i,t)N(i,t) \tag{6.4.1}$$

where n is the total number of traffic classes provided by the access network, C is the amount of reserved bandwidth for best-effort flows, $B_{avg}(i,t)$ is the bandwidth consumed by flows in class i at time t, and $N(i,t)$ is the number of traffic flows in class i at time t.

Assume flow f is mapped to wireless QoS class c after handoff. The NAP then determines an updated value $\hat{A}$ according to the relationship between $C(t)$, A, and $B_{max}(c)$:

$$\hat{A} = \begin{cases} B_{max}(c) & \text{if } C(t) \geq B_{max}(c) > A \\ A & \text{if } A \leq C(t) < B_{max}(c) \\ B_{min}(c) & \text{if } C(t) < A \end{cases} \tag{6.4.2}$$

$\hat{A}$ represents the maximum bandwidth available to flow f in the new wireless hop. $\hat{A}$ replaces A in the AU message that is passed to the next hop towards the CN.

NCR-RSVP obtains $\hat{A}$ from AU and compares it with token bucket parameter r. If $\hat{A}$ does not equal to r, r is updated with value $\hat{A}$ and other RSVP parameters are adjusted accordingly. Both the AU message sent to the CN and the Path message towards the MN carry the updated parameters. Upon receiving the messages, the CN can adjust the data rate according to the parameters, and the MN can request adequate resources from the network. Thus, service upgrade from lower bandwidth channel to higher bandwidth channel is completed.

When the MN moves from a higher-bandwidth network to a lower-bandwidth network, service downgrade is possible. Similarly, the MN indicates the actual serving bandwidth A in the AU message sent to the NAP. According to information in the QoS Requirement Data field, the NAP determines the serving class c and computes $\hat{A}$ as below:

$$\hat{A} = \begin{cases} A & \text{if } C(t) > A \\ B_{min}(c) & \text{if } B_{min}(c) \leq C(t) < A \\ 0 & \text{if } C(t) < B_{min}(c) \end{cases} \tag{6.4.3}$$

where $C(t)$ is computed with Equation (6.4.1). If $\hat{A}$ is zero, the flow will be treated in a best-effort manner.

6.5 QAES Signaling

Generally speaking, users expect service upgrade in a cellular-to-WLAN handoff, and accept service degrade in a WLAN-to-cellular handoff. In either case, a vertical handoff may cause one of the four cases, listed in Table 7, determined by the presence of end-to-end QoS signaling in the old access network and in the new access network. QAES signaling is designed to meet user expectations upon cellular/WLAN handoff and enable smooth QoS domain switch. In all the illustrations of event sequences, we chose RSVP as the example of end-to-end QoS signaling protocol because of its popularity and maturity in research. We chose DiffServ as the example of prioritization-based QoS model because it is the standard IP QoS model in cellular networks.

Table 7: Presence of signaling protocol during vertical handoff

Case	before handoff	after handoff	Major schemes
1	Yes	Yes	QAA handoff, service adjustment
2	Yes	None	QAA handoff, mapping and adjustment
3	Transparent	Yes	Upgrade to finer QoS
	None	Yes	Two-stage QoS update
4	None	None	Case 2 without signaling update

The details of QAES signaling in the four QoS domain switch are introduced in the following text.

Case 1 If an end-to-end signaling protocol is enabled in both the old access network and the new access network, a cellular/WLAN handoff follows the rules of a basic QA-assisted handoff, as proposed in Chapter 5. The only difference is that service adaptors located at the NAP and the NANG perform service adjustment during handoff, which enables service upgrade in a cellular-to-WLAN handoff or service degrade in a WLAN-to-cellular handoff.

Case 2 If end-to-end QoS signaling is enabled in the old access network but disabled in the new access network, a potential service upgrade can still be enabled if it is a cellular-to-WLAN handoff. Since the MN was participating an end-to-end QoS signaling before handoff, it is natural for the MN to generate and send AU messages. The NANG receives the AU and extracts QoS information. QoS mapping is performed at the NANG so that new packets can receive prioritized forwarding. A QoS signaling update is going on following the basic rule of QAA so that packets still receive desired QoS in other parts of the data route. QoS signaling is transparent in the new access network except for the MN, which is still an endpoint of the signaling. Figure 21 illustrates signaling events for this case of QoS domain switch.

Case 3 If end-to-end signaling is enabled only in the new access network, there are two possibilities. The first one (referred to as Case 3.1) is that end-to-end signaling is

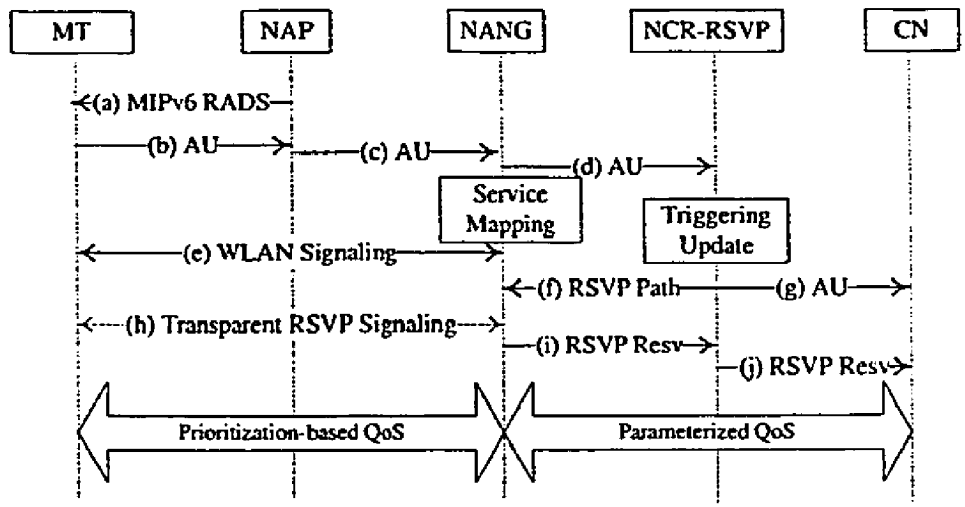


Figure 21: QoS domain switch Case 2 in a UMTS-to-WLAN handoff

transparent in the old access network but the MN is able to process the signaling messages. In this case, the handoff results in a service upgrade from coarse-grained QoS to fine-grained QoS. Figure 22 illustrates the event sequence during such a UMTS-WLAN handoff. Before handoff, RSVP signaling (a) is transparent in the UTRAN. When the MN changes its point of attachment, it receives router advertisement (b) from the NAP. An AU message (c) is composed and sent towards the CN from the MN. NCR-RSVP, usually located in the access network of the CN side, identifies itself and takes actions according to the algorithm in Chapter 5. RSVP Path messages (g-i) are propagated towards the MN and the MN replies with a Resv message (j). When Resv message (k) is processed at the NANG and the corresponding QoS provisioning is established, fine-grained QoS is available to data flows destined to the MN. Meanwhile, the CN starts sending data packets to the NCoA of the MN upon receiving the AU message (f). In most cases, data packets arrive at the NAN after QoS configuration is complete, which ensures a smooth QoS domain switch and minimal QoS disruption.

The second possibility (referred to as Case 3.2) is that there is no end-to-end signaling at all before handoff. However, the QoS Indication Option indicates to QA-MN that fine-grained QoS is available in the new access network. In this case, we

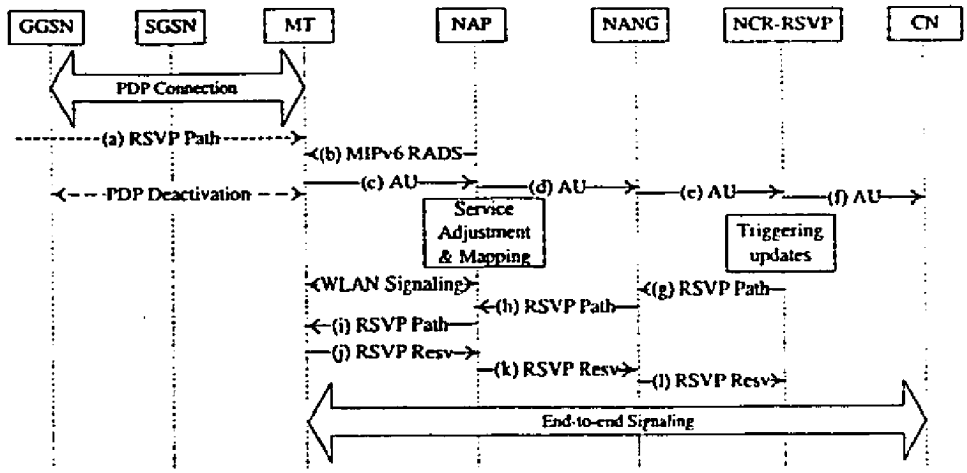


Figure 22: QoS domain switch Case 3.1 in a UMTS-to-WLAN handoff

propose a two-stage QoS setup method to enable fast establishment of QoS states after handoff. As in the other cases, the QA-MN collects QoS parameters and the AU message carries these parameters to the new access point and other routers in the new access network. In fact, the delivery of AU messages acts as QoS signaling during the establishment of the first-stage QoS. Afterwards, the setup of second-stage QoS states is executed by the QoS signaling protocol in the new access network. Although the desired QoS may not be completely provided to the MN right after handoff, this method prevents the new packets from being treated in a best-effort manner in the new access network before the desired service level is available. It is especially useful for traffic flows destined to the MN. Figure 23 illustrates the events in sequence that will occur during such a UMTS-WLAN handoff. When MIPv6 RADS with QoS Indicator (a) indicates that the new QoS domain is RSVP enabled, QA-MN determines whether RSVP signaling was previously available. If not, general characteristics of traffic flows are extracted from the database provided by the dynamic information collection procedure. The QA-MN maps the traffic class to the proper RSVP-compatible QoS class and fills the QoS Requirement Data field of the AU message accordingly. The NAP receives this AU message (b) and translates the IP QoS requirement to 802.11 EDCF compatible parameters. The first stage QoS, WLAN QoS, is now available to the flows. When the AU (c) arrives at the CN, the QA-CN notifies the CN to

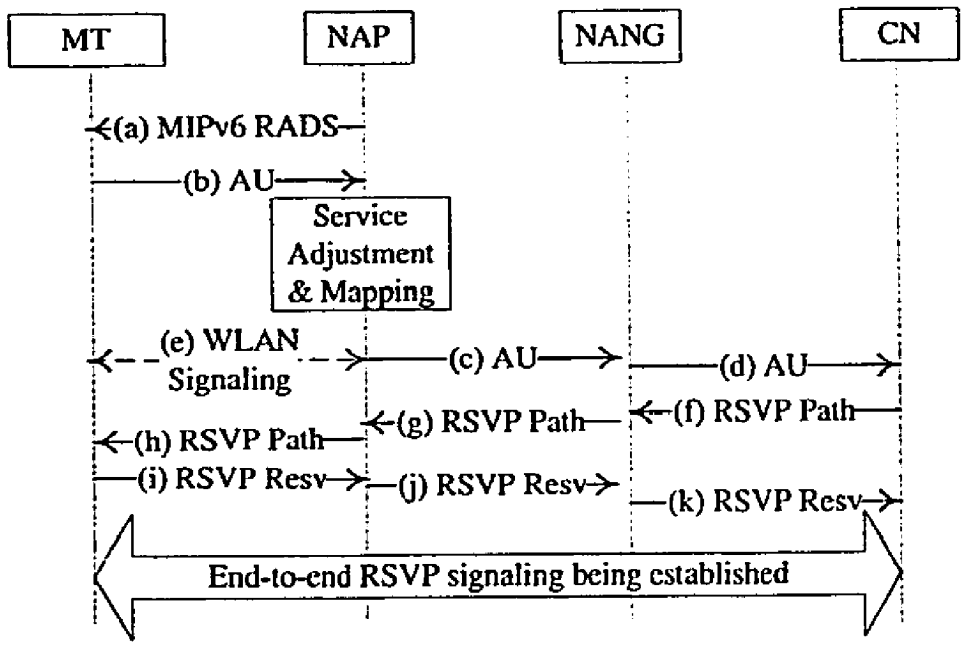


Figure 23: QoS domain switch Case 3.2 in a cellular-to-WLAN handoff

generate RSVP Path message (f). Path messages (g, h) are propagated towards the new location of the MN, which replies with a Resv message (i). Thus, the second stage QoS, end-to-end IP QoS, can be established with the propagation of Resv messages.

Case 4 If end-to-end QoS signaling is not available either before or after handoff, the situation is very similar to case 2, except that AU messages do not trigger a signaling update along the route. Instead, the new edge routers and the NANG configure prioritized forwarding according to the AU messages.

6.6 Performance Evaluation

We argue that the proposed UMTS-WLAN handoff signaling scheme is able to improve handoff performance in various aspects. It enables smooth QoS domain switch, reduces QoS configuration latency during handoff, provides automatic service adjustment, and reduces overhead on message delivering and processing. In this section we evaluate the performance of the signaling scheme through analytical study.

6.6.1 Performance Analysis

We analyze three measurements in this section: delay difference, QoS configuration latency, and signaling overhead. The definitions of the first two measurements are the same as in Chapter 5. Due to different cases of QoS domain switch, it is only meaningful to analyze delay difference and QoS configuration latency when end-to-end signaling is present explicitly or implicitly both before and after handoff. In addition, signaling overhead is briefly analyzed to justify that the proposed scheme does not introduce extra signaling to the precious wireless channel resources.

Delay difference This measurement actually indicates how much the proposed scheme can reduce the influence of handoff on packets sent to the new location of the MN after handoff. If delay difference is non-negative, i.e., QoS configuration delay equals or is smaller than maximum allowed delay, the packets can be treated with desired QoS. Otherwise, the first packets arriving at routers in the new segment will not receive any special treatment.

Our major concern is the QoS configuration delay at the NANG. QoS update is performed upon receiving the AU message at all the QoS-critical routers located on the path segment from the NANG to the CN. This ensures that new packets can receive desired QoS when they arrive at these routers, because the CN will not send packets to the new location of the MN until it receives the AU message and performs the corresponding update.

We use the same denotations as in Chapter 5 to analyze delay difference at the NANG. Let $T_{msg(n,m)}$ denote the message delivery and processing time of message msg from node n to node m. Then $T_{msg(n,m)}$ is equal to the hop distance from node n to node m. Let d_w be the hop distance from the MN to the NANG, d_n be the hop distance from the NANG and the NCR-RSVP, and d_c be the distance from the NCR-RSVP to the CN. At the NANG, QoS configuration delay, $D_{QCD,NANG}$, and maximum allowed delay, $D_{MAD,NANG}$, can be computed by the following equations

respectively:

$$\begin{aligned} D_{MAD,NANG} &= T_{AU(MN,NCR-RSVP)} + T_{AU(NCR-RSVP,CN)} + T_{DATA(CN,NANG)} \\ &= (d_w + d_n) + (d_c) + (d_c + D) \\ &= d_w + 2 * d_c + 2 * d_n \end{aligned}$$

$$\begin{aligned} D_{QCD,NANG} &= T_{AU(MN,NCR-RSVP)} + T_{Path(NCR-RSVP,MN)} + T_{Resv(MN,NANG)} \\ &= (d_w + d_n) + (d_n + d_w) + (d_w) \\ &= 3 * d_w + 2 * d_n \end{aligned}$$

Thus, delay difference at the NANG, $D_{diff,NANG}$ is equal to:

$$\begin{aligned} D_{diff,NANG} &= D_{MAD} - D_{QCD} \\ &= 2 * (d_c - d_w) \end{aligned}$$

It is obvious that $D_{diff,NANG}$ has non-negative value if $d_c \geq d_w$, which means that the wireless network does not have more hops than the access network where the CN is located does. Given the fact that wireless access networks usually do not have many hops, this condition is true in most cases.

QoS configuration latency If RSVP signaling is available before and after handoff, the following events happen during handoff without QAES:

- The MN switches to a new AP in a new AN. (At least some routers in the NAN are RSVP-aware so that RSVP signaling exchanges are expected in the NAN.)
- The MN updates the new CoA with CN via BU messages.
- The CN initiates a new round of RSVP negotiation with the MN, by issuing new Path messages destined to the NCoA.
- The arrival of the Resv message at the CN completes the round of message exchanges.

- Afterwards, packets delivered to the new location of the MN can receive desired quality of service. During the handoff, however, packets are generally treated in best-effort manner.

  Hence, QoS configuration latency without the QAA scheme can be computed with the following:

$$\begin{aligned} T_{Orig.} &= T_{BU(MN,CN)} + T_{Path(CN,MN)} + T_{Resv(MN,CN)} \\ &= 3 * (d_w + d_n + d_c) \\ &= 3 * d_w + 3 * d_n + 3 * d_c \end{aligned}$$

Based on the major events explained in Section 6.5, QoS configuration latency with QAES is

$$\begin{aligned} T_{QAA} &= T_{AU(MN,NCR-RSVP)} + T_{Path(NCR-RSVP,MN)} + T_{Resv(MN,NANG)} \\ &= 2 * (d_w + d_n) + d_w \\ &= 3 * d_w + 2 * d_n \end{aligned}$$

It is clear that QoS configuration latency is reduced by $d_n + 3 * d_c$ with the QAA scheme.

Signaling overhead In any case, the QAES signaling does not introduce new signaling overhead to the handoff process. AU replaces BU to perform binding update. AU and end-to-end signaling together perform QoS update. Furthermore, end-to-end signaling update is localized to the new path segment. Therefore, we argue that signaling overhead is reduced in the proposed signaling framework.

6.6.2 Simulation Configurations

In order to demonstrate the functionality of the Service Adaptor, we have done simulations with NS-2. Since there is no implementation of the UMTS QoS architecture in

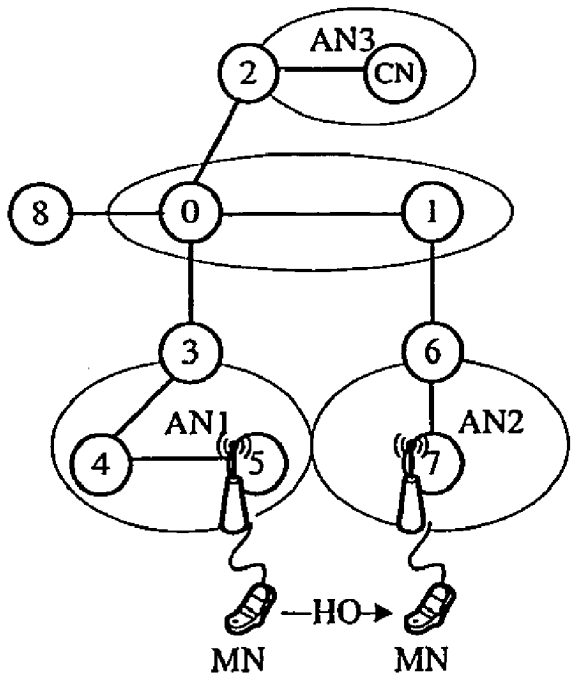


Figure 24: Simulation network topology

NS-2, we use an 802.11a network to emulate the UMTS access network. The emulation does not change the effects of the automatic service adjustment algorithm because the algorithm is independent of the actual enforcement of wireless QoS schemes.

Figure 24 shows the topology of the simulations. AN1 emulates a UMTS access network, which has lower capacity. In contrast, AN2 simulates a WLAN with higher capacity. The core network is based on DiffServ and the access network AN2 is RSVP-aware. This configuration corresponds to QoS domain switch Case 1 and Case 3. CN at AN3 generates CBR data traffic and sends it to the MN. Background traffic flows are always saturating bottleneck links, i.e., links between each BS and its attaching router. In other words, these links become congested when both background traffic flows and the CBR data traffic flow arrive. Since we do not have wireless QoS implemented, these background traffic flows assist to demonstrate the effect of QoS control in wireless access networks.

Three scenarios have been simulated:

- S1: CBR data traffic is generated at a rate of 1.4Mbps all the time. The basic

Table 8: Simulation configuration parameters

Parameter	Value
Simulation time	60 second
Congested link capacity	2 Mb
Background traffic rate	800 kb
Core network link delay	10 ms
Access network link delay	5 ms
Wireless data rate at AN1	2 Mb
Wireless data rate at AN2	11 Mb

QAA scheme proposed in Chapter 5 is applied.

- S2: CBR data traffic is generated at a rate of 1.0Mbps all the time. The basic QAA scheme proposed in Chapter 5 is applied.

- S3: CBR data traffic is generated at a rate of 1.0Mbps initially. QAES with service adjustment is applied.

The major parameters are listed in Table 8. The difference in Congested Link Capacity compared with that used in Table 1 is due to the different wireless tehnologies assumed in this chapter. All other parameters are set to the default values provided by NS-2.

6.6.3 Simulation Results

Since the major purpose of the simulations is to demonstrate the service adjustment algorithm, we present the measurements of throughput and end-to-end delay for the CBR data traffic from the CN to the MN in Figure 25.

In scenario S1, CBR traffic flow obtained a throughput of around 1 Mbps in spite of the data rate of 1.4 Mbps and the resource request of 1.4 Mbps before handoff. This is because of the limited capacity of AN1. The end-to-end delay remained stably high all the time as a result of relatively low data forwarding rate. When handoff happened at the 30th second, the basic QAA scheme ensured that QoS configuration latency was small, which is demonstrated by the short gap in the plot of throughput.

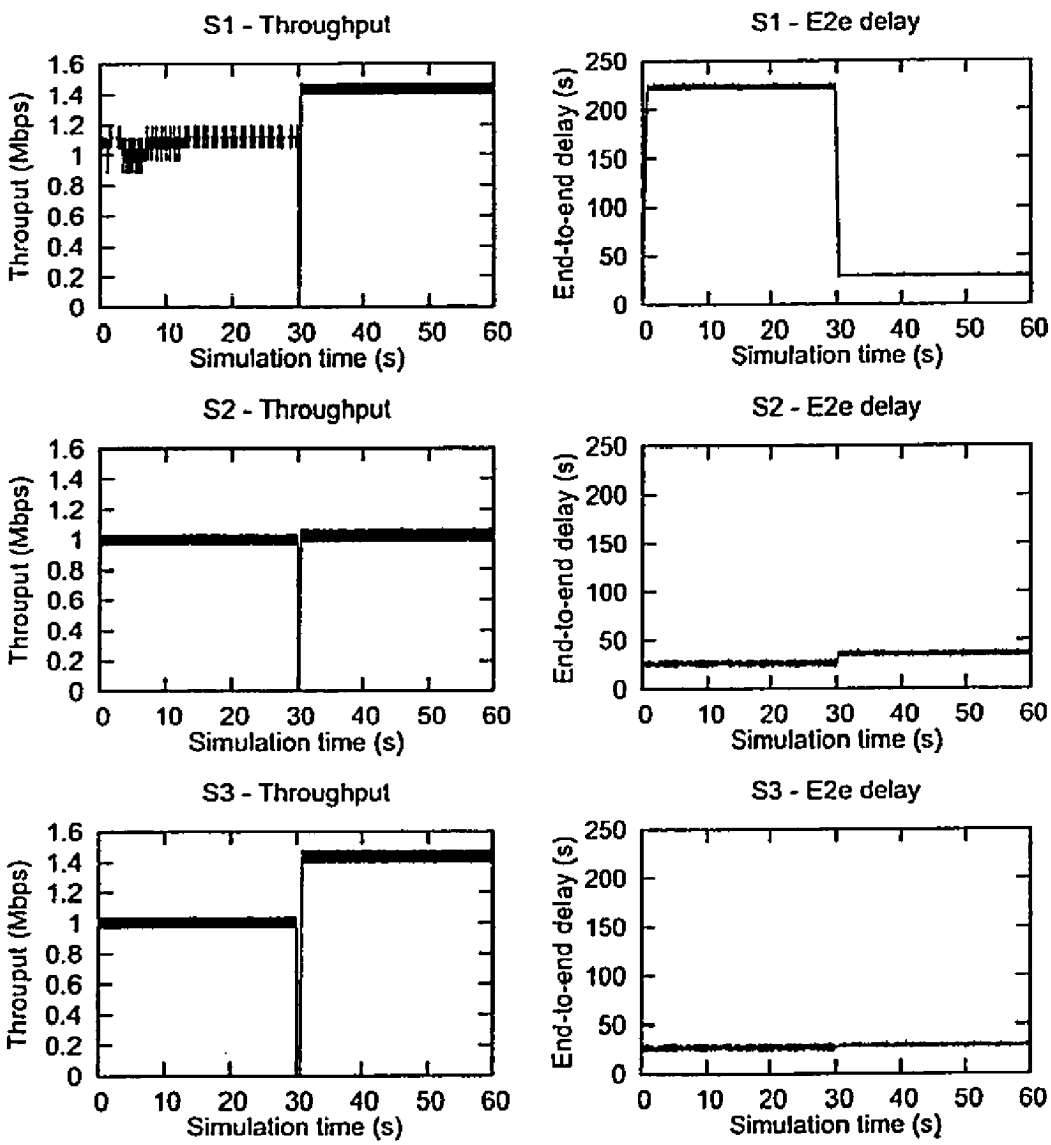


Figure 25: Throughput and end-to-end delay of CBR traffic from the CN to the MN

Given the largely increased network capacity, the data traffic obtained the desired throughput of 1.4 Mbps and the end-to-end delay decreased to around 28 ms.

In scenario S2, AN1 was able to meet the resource request of data traffic. So the throughput at the MN was stable at 1.0 Mbps and the end-to-end delay was low. After handoff, the increased network capacity did not bring much effect to the data traffic. Throughput remained at 1.0 Mbps and the end-to-end delay was about 35 ms, slightly larger than in S1 due to the lower data rate.

In scenario S3, CBR traffic was first generated at 1.0 Mbps, resulting in stable throughput and low end-to-end delay before handoff. Upon handoff, however, the service adjustment algorithm determined that the traffic flow was entitled to a higher data rate. In the simulation, we assume the maximum bandwidth granted to video flow in AN1 is 1.4 Mbps. According to the algorithm, the flow obtained data rate of 1.4 Mbps after handoff. End-to-end delay became 30 ms accordingly.

In summary, the service adjustment algorithm is able to bring service upgrade during handoff. With the adjustment, traffic flow can receive optimal service with the varying network capacity.

6.7 Chapter Summary

In this chapter, we address the challenges imposed on maintaining the service level for vertical handoff sessions. We propose a QoS control communication architecture for handoff, including service adaptor, service mapper, and signaling group, QAES, to meet the challenges. The enhancements are expected to enable fast QoS re-configuration in the new access network after a vertical handoff. Furthermore, automatic service adjustment enables service upgrade and degrade during a cellular/WLAN handoff. We use a combination of analysis and simulation to evaluate system performance and present our results.

Chapter 7

Conclusions

QoS control communications in future generation networks are complicated by the heterogeneity of the network technologies and the diversity of QoS provisioning mechanisms. QoS control communications are further challenged by the user expectation of "roaming anywhere doing anything", i.e., continuous access to desired quality of data services for a wide spectrum of applications. In this thesis, we have conducted extensive research to meet these challenges. The major goal is to design a unified QoS control communication framework for heterogeneous network environments. The results of our research are summarized in the following section.

7.1 Summary of Results

The achievements of this thesis are presented in the three aspects below.

1. **The multi-panel control communication framework is effective for heterogeneous network environments with different QoS models.** Generalization of control functionalities has enabled a unified signaling framework for multiple QoS and mobility management mechanisms in different wireless platforms. Moreover, control communications in the framework are triggered by user events and network events, supporting effective performance analysis and optimization, which has been proved by the design and improvement of the

QoS Panel. Finally, the framework is open to future extensions, because only functionalities in each panel and their interactions inside and between panels are specified.

2. **The QoS Agent-assisted MIPV6 handoff scheme has greatly improved handoff performance in terms of QoS updates** We have utilized and extended MIPv6 binding update messages to carry critical information to the nodes that are critical to QoS updates during handoff. By defining different types of QoS Agent and their functions in assisting MIPv6 handoffs, multiple QoS models can perform quick, smooth QoS re-configurations. Simulation results have shown that the scheme reduces QoS configuration latency during handoff, which further minimizes service disruption in terms of end-to-end latency and packet loss.

3. **The QoS Agent-enhanced vertical handoff scheme supports adjustable service provisioning to wireless users roaming among networks with different capacities.** The QoS Panel of the framework has been further completed by the proposal of a control communication architecture for vertical handoff. The two major problems of QoS update during vertical handoff, QoS domain switch and network capacity change, have been solved by QoS Agent-enhanced signaling and automatic service adjustment respectively. Moreover, guidelines for service mapping are provided for heterogeneous wireless technologies and QoS models. We have demonstrated that handoff performance can be greatly improved in terms of QoS configuration latency and optimal service can be provided to mobile users when they switch between different networks.

7.2 Future Research Directions

The control communication framework is far from complete. Potential future work is suggested in the following three directions:

1. Develop more realistic simulation environments to better evaluate the framework. The framework involves multiple network platforms, each with its own QoS mechanisms. More modules are expected to be implemented in the simulation software we have chosen, in order to evaluate and improve the framework. Specifically, the implementations of different wireless QoS schemes are desired. It could be a further step to implement a prototype of the framework in heterogeneous wireless platforms.

2. Apply the concept of QoS Agent to more QoS models and QoS signaling protocols. We expect the QoS Panel to accommodate emerging protocols if the corresponding types of QoS Agents are properly defined. However, the details need much investigation and evaluation.

3. Complete other panels and the inter-panel communications as well. The Access Panel and the Service Panel are as important as the QoS Panel in the framework. Their interactions determine the overall performance of the control communications at different trigger events. We leave the details to future researchers.

Bibliography

[3GP00] 3GPP2. *3GPP2 C.S0017-0-2 Version 2.0, Data Service Options for Spread Spectrum Systems Addendum 2*, August 2000.

[3GP03] 3GPP. *TS 25.401 v5.6.0,UTRAN Overall Description (Release 5)*, June 2003.

[3GP04] 3GPP2. *P.S0001-B Version 2.0, cdma2000 Wireless IP Network Standard*, September 2004.

[3GP05] 3GPP. *TS 23.107 v.6.3.0, Quality of Service (QoS) Concept and Architecture (Release 6)*, June 2005.

[BB+98] S. Blake, D. Blake, et al. An Architecture for Differentiated Services. RFC 2475, December 1998.

[BCS94] R. Braden, D. Clark, and S. Shenker. Integrated Services in the Internet Architecture: an Overview. RFC 1633, June 1994.

[Ber00] Y. Bernet. A Framework for Integrated Services Operation over Diffserv Networks. RFC 2998, November 2000.

[Bru04] M. Brunner. Requirements for Signaling Protocols. RFC3726, April 2004.

[BZ+97] R. Braden, L. Zhang, et al. Resource ReserVation Protocol -Version 1 Functional Specification. RFC 2205, September 1997.

[CA05] Yan Cheng and J. W. Atwood. QoS Signaling Model for Heterogeneous Mobile Environments. In *Proceedings of The International Conference on Communication and Information (ICCI 2005)*, pages 395–401, Beijing, China, June 2005.

[CA06a] Yan Cheng and J. W. Atwood. A hierarchical agent assisted MIPv6 handover scheme with QoS support. In *Proceedings of the 3rd international conference on Quality of service in heterogeneous wired/wireless networks*, Waterloo, Ontario, Canada, August 2006.

[CA06b] Yan Cheng and J. W. Atwood. Towards Minimizing Service Degradation during MIPv6 Handoffs. In *Proceedings of The 31st IEEE Conference on Local Computer Networks (LCN 2006)*, Tampa, FL, October 2006.

[CA07] Yan Cheng and J. W. Atwood. Performance Evaluation of QoS Agent-Assisted MIVv6 Handoff Scheme. In *Proceedings of The 3rd ACM International Workshop on QoS and Security for Wireless and Mobile Networks (Q2SWinet 2007)*, pages 55–62, Chania, Greece, October 2007.

[CAV97] G. Chiruvolu, A. Agrawal, and M. Vandenhoute. Mobility and QoS support for IPv6-based real-time wireless Internet traffic. In *Proc. Int. Conf. Communication*, pages 334–338, Vancouver, BC, Canada, June 1997.

[CC+03] R. Chakravorty, J. Crowcroft, et al. A framework for dynamic SLA-based QoS control for UMTS. *IEEE Wireless Communications*, 10(5):30–37, October 2003.

[CGC01] A. T. Campbell and J. Gomez-Castellanos. IP Micro-Mobility Protocols. *ACM SIGMOBILE Mobile Comp. Comm. Rev.*, 2001.

[CK01] H. Chaskar and R. Koodli. A Framework for QoS Support in Mobile IPv6. *draft-chaskar-mobileip-qos-01.txt*, March 2001.

[CKK02] F. M. Chiussi, D. A. Khotimsky, and S. Krishnan. Mobility Management in Third-Generation All-IP Networks. *IEEE Communication Magazine*, September 2002.

[DB95] L. Delgrossi and L. Berger. Internet Stream Protocol Version 2 (ST2) Protocol Specification- Version ST2+. RFC 1819, 1995.

[DWZ03] J. Diederich, L. Wolf, and M. Zitterbart. A mobile differentiated services QoS model. In *Proceedings of the 3rd IEEE Workshop on Applications and Services in Wireless Networks (ASWN)*, Berne, Switzerland, July 2003.

[Ern01] Thierry Ernst. MobiWan: NS-2 extensions to study mobility in Wide-Area IPv6 Networks. http://www.inrialpes.fr/planete/pub/mobiwan/, June 2001.

[FKK02] X. Fu, H. Karl, and C. Kappler. QoS-Conditionalized Handoff for Mobile IPv6. In *Proceedings of NETWORKING 2002*, pages 721–730, Pisa, Italy, May 2002.

[FST03] X. Fu, H. Schulzrinne, and H. Tschofenig. Mobility functions in the NTLP. draft-jeong-nsis-mobility-ntlp-01.txt, October 2003.

[GAD02] Y. Guo, Z. Antouniou, and A. Dixit. IP Transport in 3G Radio Access Networks: an MPLS-based Approach. In *Proceedings of WCNC2002*, 2002.

[Gre98] Marc Greis. RSVP/ns: An implementation of RSVP for the Network Simulator ns-2. http://titan.cs.uni-bonn.de/~greis/rsvpns/index.html, February 1998.

[GWM04] X. Gao, G. Wu, and T. Miki. End-to-end QoS provisioning in mobile heterogeneous networks. *IEEE Wireless Communications*, 11(3):24–34, June 2004.

[H+99] J. Heinanen et al. Assured Forwarding PHB Group. RFC 2597, June 1999.

[H+04] J. Hillebrand et al. Quality-of-Service Signaling for Next Generation IP-based Mobile Networks. *IEEE Communications Magazine*, pages 72–79, June 2004.

[HK+05] R. Hancock, G. Karagianni, et al. Next Steps in Signaling: Framework. RFC4080, June 2005.

[Hus00] G. Huston. Next Steps for the IP QoS Architecture. RFC 2990, November 2000.

[IEE04] IEEE. *IEEE Standard for Local and Metropolitan Area Networks–Part 16: : Air Interface for Fixed Broadband Wireless Access Systems*, October 2004.

[IEE06] IEEE. *Amendment to IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems - Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands*, February 2006.

[J+99] V. Jacobson et al. An Expedited Forwarding PHB. RFC 2598, June 1999.

[Jai] Raj Jain. Quality of Service Over IP. http://www.cse.ohio-state.edu/~jain/talks/ftp/ipqos/.

[Jas03] M. Jaseemuddin. An Architecture for Integrating UMTS and 802.11 WLAN Networks. In *Proceedings of IEEE Symposium on Computers and Communications (ISCC 2003)*, pages 716–723, Antalya, Turkey, July 2003.

[JPA04] D. Johnson, C. Perkins, and J. Arkko. Mobility support in IPv6. RFC 3775, June 2004.

[JZM02] M. Jaseemuddin, J. A. Zubairi, and O. Mahmoud. A study of profiled handoff for diffserv-based mobile nodes. *Wireless Communications and Mobile Computing*, 2(4):339–356, 2002.

[Koo05] R. Koodli. Fast Handovers for Mobile IPv6. RFC 4068, July 2005.

[LBN06] LBNL. The Network Simulator - ns-2. http://www.isi.edu/nsnam/ns/, 2006.

[LF+02] S. Lepaja, R. Fleck, et al. QoS Provisioning to Mobile Internet Users. In *Proceedings of EU2002: Next Generation Wireless Networks: Technologies, Protocols, Services and Applications*, Florence, Italy, February 2002.

[LS+05] M. Liebsch, A. Singh, et al. Candidate Access Router Discovery (CARD). RFC4066, July 2005.

[MA01] B. Moon and H. Aghvami. RSVP Extensions for Real-Time Services in Wireless Mobile Networks. *IEEE Commuications Magazine*, pages 52–59, December 2001.

[MA+03] V. Marques, R.L. Aguiar, et al. An IP-based QoS architecture for 4G operator scenarios. *IEEE Wireless Communications*, 10(3):54–62, June 2003.

[MF05] J. Manner and X. Fu. Analysis of Existing Quality of Service Signaling Protocols. RFC4094, May 2005.

[MK04] J. Männer and M. Kojo. Mobility Related Terminology. RFC 3753, June 2004.

[MKM08] J. Manner, G. Karagiannis, and A. McDonald. NSLP for Quality-of-Service Signaling. draft-ietf-nsis-qos-nslp-16.txt, February 2008.

[MP07] C. Makaya and S. Pierre. Reliable Integrated Architecture for Heterogeneous Wireless Networks. *Journal of Networks*, 2(6):24–32, December 2007.

[MP08a] C. Makaya and S. Pierre. Adaptive Handoff Scheme for Heterogeneous IP Wireless Networks. *Computer Communications*, 31(10):2094–2108, July 2008.

[MP08b] C. Makaya and S. Pierre. An Analytical Framework for Performance Evaluation of IPv6-based Mobility Management Protocols. *IEEE Transactions on Wireless Communications*, 7(3):972–983, March 2008.

[MP08c] C. Makaya and S. Pierre. An Architecture for Seamless Mobility Support in IP-based Next Generation Wireless Networks. *IEEE Transactions on Vehicular Technology*, 57(2):1209–1225, March 2008.

[Net] NORTEL / Bay Networks. IP QoS - A Bold New Network. White Paper.

[NRT04] Q. Ni, L. Romdhami, and T. Turletti. A survery of QoS enhancements for IEEE802.11 wireless LAN. *Wireless Communications and Mobile Computing*, 4(5):547–566, 2004.

[NVFA06] Q. Nguyen-Vuong, L. Fiat, and N. Agoulmine. An Architecture for UMTS-WIMAX Interworking. In *Proceedings of 1st IEEE International Workshop on Broadband Convergence Networks (BCN 2006)*, Vancouver, Canada, April 2006.

[P+05] R. Prior et al. Heterogeneous Signaling Framework for End-to-end QoS Support in Next Generation Networks. In *Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS'05)*, Big Island, HI, USA, January 2005.

[PC02] S. Pack and Y. Choi. Seamless QoS Handling Mechanism for Macro and Micro Mobility. *Electron. Lett.*, 2344:62–73, February 2002.

[PC+03] H. Parikh, H. Chaskar, et al. Seamless Handoff of Mobile Terminal from WLAN to CDMA2000 Network. In *Proceedings of 2003 World Wireless Congress (3GWireless'2003)*, San Francisco, CA, USA, May 2003.

[Per96] C. Perkins. IP Mobility Support. RFC 2002, 1996.

[PS98] P. Pan and H. Schulzrinne. YESSIR: A Simple Reservation Mechanism for the Internet. In *Proceedings of NOSSDAV*, Cambridge, UK, July 1998.

[Rob05] L. Roberts. QoS Signaling for IP QoS Support. TIA 1039, July 2005.

[RP+02] R. Ramjee, T. La Porta, et al. HAWAII: A Domain-based Approach for Supporting Mobility in Wide-area Wireless Networks. *IEEE/ACM Transactions on Networking*, 6(2), June 2002.

[SCEMB05] H. Soliman, C. Castelluccia, K. El-Malki, and L. Bellier. Hierarchical Mobile IPv6 Mobility Management (HMIPv6). RFC 4140, August 2005.

[SH08] H. Schulzrinne and R. Hancock. GIST: General Internet Signaling Transport. draft-ietf-nsis-ntlp-17.txt, October 2008.

[Sin96] S. Singh. Quality of Service Guarantees in Mobile Computing. *Computer Communication*, 19(1):359–371, January 1996.

[SJ+05] W. Song, H. Jiang, et al. Resource Management for QoS Support in Cellular/WLAN Interworking. *IEEE Network*, pages 12–18, September/October 2005.

[SK98] M. Stemm and R. H. Katz. Vertical Handoffs in Wireless Overlay Networks. *Mobile Networks and Applications*, 3(4):335–350, 1998.

[SLSK00] Q. Shen, A. Lo, W. Seah, and C.C. Ko. On Providing Flow Transparent Mobility Support for IPv6-based Wireless Real-time Services. In

Proceedings of Seventh International Workshop on Mobile Multimedia Communications (MoMuC2000), October 2000.

[SP04] D. Skyrianoglou and N. Passas. A Framework for Unified IP QoS Support Over UMTS and WirelessLANs. In *Proceedings of the Fifth European Wireless Conference: Mobile and Wireless Systems beyond 3G (European Wireless 2004)*, Barcelona, Spain, February 2004.

[T+01] A. K. Talukdar et al. MRSVP: A Resource Reservation Protocol for an Integrated Services Network with Mobile Hosts. *Wireless Networks*, 7(1):5–19, January/February 2001.

[THM05] A.-E.M. Taha, H. S. Hassanein, and H. T. Mouftah. Extensions for internet qos paradigms to mobile ip: a survey. *IEEE Communiations Magazine*, 43(5):132–139, May 2005.

[TSZ99] A. Terzis, M. Srivastava, and L. Zhang. A simple QoS signaling protocol for mobile hosts in the integrated services internet. In *Proceedings of IEEE INFOCOM'99*, pages 1011–1018, New York, 1999.

[TW+99] A. Terzis, L. Wang, et al. A Two-Tier Resource Management Model for the Internet. In *Proceedings of IEEE GLOBECOM'99*, December 1999.

[VCG98] A. G. Valko, A. T. Campbell, and J. Gomez. Cellular IP. Internet Draft, draft-valko-cellularip-00.txt, November 1998.

[VJFD99] B. Vandalore, R. Jain, S. Fahmy, and S S. Dixit. AQuaFWiN: Adaptive QoS Framework for Multimedia in Wireless Networks and its Comparison with other QoS Frameworks. In *Proceedings of the 26th IEEE Conference on Local Computer Networks*, pages 88–97, Lowell, Massachusetts, USA, October 1999.

[VR+03] V. K. Varma, S. Ramesh, et al. Mobility Management in Integrated UMTS/WLAN Networks. In *Proceedings of 2003 IEEE International*

Conference on Communications (ICC'03), pages 1048–1053, Anchorage, Alaska, May 2003.

[WM+05] X. Wang, G. Min, et al. An adaptive QoS framework for integrted cellular and WLAN networks. *Computer Networks*, 47:167–183, 2005.

[WMAB03] X. Wang, J. Mellor, and K. Al-Begain. Towards Providing QoS for Integrated Cellular and WLAN Networks. In *Proceedings of the 4th Annual Postgraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting (PGNet 2003)*, pages 207–211, Liverpool, UK, June 2003.

[WMM04] X. Wang, G. Min, and J. Mellor. Adaptive QoS Control in Cellular and WLAN Interworking Networks. In *Proceedings of the 2nd International Working Conference: Performance Modelling and Evluation of Heterogeneous Networks (HET-NETs '04)*, West Yorkshire, UK, July 2004.

[XL+05] Y. Xiao, K. K. Leung, et al. Architecture, mobility management, and quality of service for integrated 3G and WLAN networks. *Wireless Communications and Mobile Computing*, 5(7):805–823, November 2005.

[YL+00] S. U. Yoon, J. H. Lee, et al. QoS Support in Mobile/Wireless IP networks using Differentiated Services and Fast Handoff Method. In *Proceedings of WCNC2000*, Chicago, IL, USA, September 2000.

[ZMH04] Y. Zhang, D. Makrakis, and D. Hatzinakos. Supporting of QoS and Micro-Mobility in MPLS-based IPv6 Wireless Networks. In *Proceedings of the Fifth European Wireless Conference: Mobile and Wireless Systems beyond 3G (European Wireless 2004)*, Barcelona, Spain, February 2004.

[ZPM06] L. Zhang, S. Pierre, and L. Marchand. A New Seamless Method to Support CDMA2000/WLAN Vertical Handover. In *Proceedings of CCECE/CCGEI 2006*, pages 1122–1126, Ottawa, Canada, May 2006.