Towards a Service Level Negotiation Based QoS Provisioning Scheme for Wireless Networks

Towards a Service Level Negotiation Based QoS Provisioning Scheme for Wireless Networks

サービスレベルネゴシエーションに基づくワイヤレスネットワークのサービス提供スキーム

A dissertation presented

By

Juan Carlos Fernandez Perez

Submitted to the Graduate School of Information Sciences

in partial fulfillment of the requirements for the

Doctor of Information Sciences

January 2010

Department of System Information Sciences

Tohoku University

Towards a Service Level Negotiation Based QoS Provisioning Scheme for Wireless Networks

サービスレベルネゴシエーションに基づくワイヤレスネットワークのサービス提供スキーム

A dissertation presented

By

Juan Carlos Fernandez Perez

Approved as to style and content by

Prof. Nei Kato ______________________

Prof. Tetsuo Kinoshita ______________________

Prof. Hideaki Sone ______________________

Abstract

The envisioned next generation wireless networks exhibit important and unique features that qualify them to be an integral part of a global ubiquitous information system, in which an efficient Quality of Service (QoS) provisioning is required. Current QoS architectures are based on centralized and static Service Level Agreement (SLA) mechanisms where SLAs are usually agreed by both the client and the service provider when a client signs up for a service. The contractual duration of such SLAs is in a large time scale, typically in order of months or years. Given heterogeneity in wireless technologies, and diversity of user connection devices, applying constant service level to a wireless user all the time during its contract period, may lead to an unfair service. Effectively it is likely that a user is offered a service level higher than what it can be actually provided by the link layer or is supportable by the user's device. In such an overbooking scenario, the customer will be unfairly charged for a service level, which he/she cannot fully utilize. In case of multiple users from different traffic classes, this unfairness issue becomes more aggravated as the service provider cannot fulfill its QoS commitments to all customers.

As a solution to this issue, a dynamic service level negotiation in a small time scale is highly desirable. This dynamic negotiation should offer users with only what they are seeking for or what is allowable by the current network conditions. This should be beneficial for both users and service providers. From the customer's perspective, a dynamic negotiation of service level is beneficial as users will be charged for only what they have actually requested or indeed used. At the service provider's end, the system scalability can be improved as savings in the network resources become possible and more users can be then served. In this regard, the first part of this thesis proposes a dynamic service level negotiation system that allows users to negotiate their service levels required by the applications that they attempt to

execute. The proposed system allows users to change the negotiated service levels in response to changes in both network conditions and their own resource requirements.

QoS and mobility functionalities go hand in hand in wireless networks, since QoS is offered to a particular user along a specific path that changes as a consequence of handoff. Thus, QoS state is no longer supported in the new path following the handoff scenario. Additionally, the unused QoS state on the former path affects the efficient use of the network resources. Thus, QoS and mobility management should be integrated to ensure the continuity of service levels perceived by users while they perform handoffs between different Base Stations (BSs). One of the hot topics in wireless networks is to track the location of the user that performs handoff and inform the appropriate BS about the Service Level Specification (SLS) of the user. The contemporary schemes to achieve this, such as SLS delivery on demand, traffic pattern prediction, and SLS broadcast approaches, present significant long delay and/or scalability issues. The second part of this thesis proposes a new fast, high scalable, secure, and robust framework of SLS to inform the new BS, following the handoff of a user, about the current service level of the user.

To provide QoS for multimedia applications in wireless networks where resources are the constraints, the proposed service level negotiation system allows users equipped with several wireless interfaces to maintain simultaneous connections with different networks, negotiate the required service level for their traffic, and reach them by aggregating the available bandwidth at these networks. However, such a bandwidth aggregation involves multiple paths in transmission of data of a single application, making packets arrive at the final destination in an out-of-order manner. To cope with this issue, this thesis proposes a multi-path scheduling algorithm to minimize the reordering delay and the associated packet loss rate.

The efficiency of the proposed system is verified by numerical results. Conducted simulations indicate the good performance of the system in handling handoff and distributing data load among multiple paths. The scalability of the system, in terms of signaling overhead and data storage, is also evaluated.

Acknowledgment

I want to express my sincere gratitude to my supervisor, Professor Nei Kato, for his guidance and support over the last five years. I have been really fortunate to have him as my supervisor. Many thanks also go to Dr. Tarik Taleb for guiding me throughout many stage of this work. I would like to thank also the other two professors in my thesis committee, Prof. Tetsuo Kinoshita and Prof. Hideaki Sone, for their interest and valuable counsel on writing this thesis.

Many thanks also to the Japanese Ministry of Education for giving me the opportunity to study in Japan by providing me with the scholarship that allowed me to come to Japan and study in Tohoku University, one of the best universities in Japan.

I would like to thanks to Motoko Shiraishi and Takako Kase for providing me with all the necessary administrative documents. Several other people helped me during the course of this research. I would like to thank all members of Kato laboratory, past and present, for their gracious help in both, academic and non-academic aspects of my stay in Japan. Special thanks go to Zubair who has been a very good friend and have closely reviewed my research work.

Finally, I want to express my admiration and gratitude to my parents, Carlos and Mercedes, my sister Karla, and my love Carmen for all their unconditional support.

January 2010.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Importance of Dynamic Service Level Negotiation

Providing Quality of Service (QoS) in wireless networks has, indeed, become a challenging task because of resource constraints in current wireless networks, mobility of users, and the ever-growing demand for real-time multimedia services (e.g., video streaming, video conferencing, online interactive games, and IPTV) that require high-quality QoS support, such as guaranteed bandwidth, delay, jitter and error rate. In the traditional wireless architectures, the Service Level Agreement (SLA) is initiated by a user via contract with the Internet Service Provider (ISP). The SLA remains static for the contract period and is applied equally to the overall traffic between the end-user and the network, regardless the different service levels required by different applications.

Since each application has different requirements in terms of latency, bandwidth and packet-error rate, the QoS provisioning scheme must cater to each of these needs. Applications requiring low latency (e.g., Voice over Internet Protocol (VoIP) may be given higher priority to use the medium, whereas applications requiring higher bandwidth may be assigned longer transmission times (e.g., video streaming applications). Other traffic may require high reliability (e.g., email and data services) and must be delivered with low packet error rates.

QoS guarantees are actually difficult to obtain in contention-based networks where sharing of the available resources follow a certain degree of randomness [1]. The proposed standards for QoS enhancements for 802.11 and 802.16 standards have put the basis for service differentiation using different MAC parameters with different traffic classes. As a consequence, different medium access priorities have evolved [2]. However, due to the traditional back-off algorithm, separation between traffic classes is rather probabilistic since best effort traffic may still frequently access the medium, especially when the network is fairly loaded. Another problem with current approaches consists in the fact that different flows of the same traffic class have usually the same MAC parameters, which means that they have the same opportunity in accessing the medium [3]-[8]. This obviously penalizes high-bit rate flows because coordinating MAC parameters provide throughput fairness, which limit the flexibility of "Network Operator" in offering QoS guarantees in multi-rate environments (i.e., flows belonging to the same priority class and having different bit-rates).

To ensure the efficient provisioning of QoS-enabled services for mobile users, there are some hot research areas as follows.

Dynamic QoS negotiation: Mobile users should be able to dynamically negotiate their service level requirements, represented by the Service Level Specifications (SLSs), with the access network. This negotiation should be performed per session, based on the resources required by each application. The network operator must guarantee the negotiated SLS during the entire course of the session.

Mobility Managements: The development of efficient ubiquitous systems that can provide a set of bandwidth-intensive and real-time services to multiple users while supporting their full mobility is a challenging task. Thus, the success of any resource allocation and admission control model depends on continuity of QoS guarantees across different Base Stations (BSs), which need to track the location of mobile users and transfer an ongoing data sessions from one BS to another without loss or interruption of service.

Bandwidth Aggregation: In order to cope with the resource constrains in current wireless networks, mobile users equipped with multiple wireless interfaces, makes simultaneous use of these interfaces to connect to the network when the coverage areas of the BSs corresponding to these technologies partially overlap. Such a capacity allows mobile terminals to increase the streaming bandwidth by distributing the load over multiple network paths.

By combining the benefits of these research areas, a user who wants to execute a high bandwidth demanding application, should negotiate the amount of bandwidth required by this application. If the bandwidth of a single interface is not enough to meet the required one, the user may consider using two or more interfaces to ensure the quality of the application. However, the transmission of packets of a single application via multiple paths, with varying characteristics in terms of capacity and propagation delay, makes those packets arrive at the final destination in an out-of-order manner, which results in packet reordering, increases the delay, and causes some packets to exceed their timers and get discarded. As a consequence, a multi-path scheduling algorithm is required.

1.2 Overview of the Proposed Dynamic Service Level Negotiation System for Wireless Networks

1.2.1 Objective

By focusing on improving the user experience while obtaining service from a wireless network, the work presented in this thesis aims at ensuring the continuity of the service level perceived by users while they are on the move between different base stations. To achieve this goal, we propose a service level negotiation system that allows users to dynamically negotiate the service level required by their traffic, and provides them with strict QoS.

1.2.2 Contribution

The major contribution consists in the integration of two essential functionalities to the service level negotiation system; mobility management and bandwidth aggregation to ensure continuity of the service level perceived by mobile users and high data rates in wireless networks , respectively.

For mobility management, we propose a fast, highly scalable and secure mechanism to inform BSs on the SLSs of users coming into their coverage areas, to ensure the seamless mobility, by allocating the negotiated resources at the new BS upon handoff, and releasing the resources allocated in the previous BS.

We define the bandwidth aggregation-aware service level negotiation process to cope with the resource constrains in current wireless networks, where mobile users equipped with multiple wireless interfaces, in combination with ISPs providing services through different wireless technologies, are allowed to make simultaneous use of these interfaces to connect to the network and negotiate the available resources via these interfaces. Thus, users achieve higher service level for their applications. We also propose a new multi-path scheduling algorithm to distribute data load among multiple disparate paths in terms of capacity and delays that minimizes the re-ordering delay and the corresponding packet loss rate.

1.2.3 Scope

- The SL negotiation is valid within the current domain only. Thus, mobile users may roam freely within the domain in which they perform SL negotiation and keep receiving the same service level. However, if a user moves to a different domain, it needs to negotiate a new Service Level (SL) with the negotiation entity of the new domain. The reason behind this lies in the fact that ISPs are unable to guarantee the SL beyond the borders of their own networks.
- Since the last hop of any wireless communications network (from the direction of the BS to the users) is usually considered to be a bottleneck, the dynamic SL negotiation system ensures the negotiated SL at the BSs.
- Our considered QoS parameter is bandwidth. This consideration aims at guaranteeing a strict SL for users instead of pre-defined traffic classes where different flows from the same traffic class may still receive the same SL even if they have different bit rates. The proposed SL negotiation system focuses on individual allocation of bandwidth at the BSs to satisfy the negotiated SLS of each user.

1.3 Summary and Organization of the Thesis

The focus of this thesis is on improving the user experience while obtaining service from a wireless network. In this chapter, we highlight the importance of the dynamic service level negotiation system for ensuring constant service levels to users while they are on the move. Also, we provide a brief overview on the proposed dynamic service level negotiation system.

The rest of this thesis is structured as follows. Chapter 2 provides detailed explanation about the service level negotiation system; architecture, main components, characteristics, negotiation procedures, and the bandwidth allocation strategy adopted to provide users with strict QoS. Chapter 3 presents the Extended Encrypted Service Level specification (EESLS), a novel scheme for mobility management, where the network delegates to users the task of informing their current SLSs to the next point of attachment upon handoff. EESLS ensures the seamless mobility of users, preserves the scalability of the system, and guarantees the security for delivering SLSs of users.

Chapter 4 describes the bandwidth aggregation mechanism, highlights the need for controlling bandwidth aggregation to maintain the scalability of the system, presents in detail the proposed multi-path scheduling algorithm called Time-Slot Earliest Delivery Path First (TS-EDPF), and demonstrates its effectiveness. In addition, to deal with the problem of VBR applications, where huge amount of the reserved bandwidth remain unused, we propose a strategy to serve best-effort traffic during the unused periods of the reserved bandwidth. The thesis concludes in Chapter 5 with a summary recapping the main advantages of the work presented in this thesis.

Chapter 2
The Service Level Negotiation System

2.1 Need for QoS over wireless networks

With the steadily growing synergy between existing heterogeneous networks, wireless networks represent important access network technology in the end-to-end multimedia services distribution chain. Due to its steadily growing capacity, wireless networks are now mature enough to be integrated in many concrete multimedia streams distribution commercial offers [9]. The success of such networks mainly depends on their ability to provide sustainable QoS for real-time multimedia streams. Even though the wireless network technologies can provide high speed (broadband) wireless access to IP networks, they have significant limitations that should be undertaken for allowing scalable and stable QoS.

Besides, audio/video streaming application imposes stringent requirements on communication QoS metrics such as loss rate, delay, and jitter. Delivering video streams to users via a wireless last hop is indeed a challenging task due to the wireless link's varying nature. It is important to guarantee these QoS requirements throughout the multimedia streams' path using appropriate QoS mechanisms. End-to-end QoS guarantees may be achieved at several levels by using appropriate mechanism at application, transport, network, and link layers. While a common approach suggests exploiting the variations of the wireless channel, an alternative is to exploit characteristics of the video streams to improve their resiliency. An integrated combination of these two approaches is a much more optimal possibility in terms of resource savings (e.g., bandwidth) and requirements guarantees (e.g., delay).

Overcoming to the above challenges may be achieved in many manners with, however, different costs in term of network resources. The problems entailed by streaming video over wireless networks are interrelated and require a thorough insight into each component of the system, from the video coding features to the wireless network specificities passing by TCP/IP protocols suite architecture.

2.1.1 Requirements and Challenges for Video Streaming over Wireless Networks

There are a number of problems that affect packet video streaming. In this section, we discuss the problems that influence video streaming quality over IP networks. Video streaming over IP is a challenging task since IP networks are shared environments that provide a rather best effort service. These networks do not offer any quality of service or guarantees of resources in terms of bandwidth, packet losses, delay variation, and transfer delay. Therefore, a key goal of video streaming is to design a system that efficiently handles the video over IP and deals with possible changes in network conditions.

The bandwidth is the most important characteristic that directly affects the quality of provisioned services. The bandwidth between a receiver and a sender is roughly known and generally time-varying. If the sender transmits more than the available bandwidth, video packets may be lost in the channel, not to mention the engendered excessive delays. An obvious technique to deal with bandwidth variation is to use adaptive video streaming. The server estimates the available bandwidth and then adapts its sending rate to match the available bandwidth. This later adaptation is widely dependent on the video coding flexibility and features.

The packet loss process affecting a given source is obviously influenced by the traffic pattern of the source itself. In fact, for the same channel conditions two different sources may experience widely different loss rates. Packet loss occurs in general in network queues and is often accompanied by significant increasing in delays. In wireless network communication, when the network reaches saturation point, the mean medium access interval increases drastically involving excessive packet dropping at wireless station MAC queues. Additionally, wireless networks exhibit high bit error rates that lead to important packet losses even when the network is lightly loaded. To combat the effect of losses, the video streaming system must be designed with error control in mind.

The transfer delay is the time that the packet experiences from its generation at the server until its reception by the client. This delay varies from one packet to another. It is affected by the pattern and volume of traffic entering the network. The variation of this delay is called the jitter. The jitter is a problem, because when a video packet arrives too late at the client it becomes useless. On other hand, when a packet arrives too fast, it must be buffered and then could produce buffer overflow.

Fluctuation in wireless link’s quality is another problem that recently attracted much attention from the research community. One fundamental difference between wired and wireless networks resides in the variation of bandwidth (i.e., overall network capacity), loss rate, and access delays of wireless networks. Based on many factors (e.g., coding/modulation rate, network load, interferences, etc.), wireless networks may exhibit widely different QoS metrics during their operation lifetime.

2.1.1.1 Bandwidth Management

As highlighted earlier, transmitting packet video stream over IP encounters three major problems: bandwidth fluctuation, excessive and variable delays, and packet losses. It is much suited to develop end-to-end techniques to adapt coded video streams to varying channel conditions, i.e., adaptive video streaming. Adaptive video streaming aims to adapt itself in any possible manners, e.g., video rate adaptation, multi-resolution streams adjustment, etc.

2.1.1.2 Video Compression Standards

In this section, we discuss possible application-level mechanisms that may be deployed at end systems to manage bandwidth fluctuation and heterogeneity of receivers. Audio and video coding algorithms are highly correlated to the growth of computation speed of systems. Nowadays, the most important compression standards for streaming video (such as H.261, H.263, MPEG-1, MPEG-2, MPEG-4 and recently H.264) are proposed by ITU (International Telecommunication Union) and ISO (International Organization for Standardization) standardization bodies. The majority of available video codec do not perform well when used to streaming media in wireless IP environments, because they require a high scalability, lower computational complexity, high resiliency to network losses and lower encoding/decoding latency. Building a video codec that respond to all these requirements is not sufficient. The codec must be aware of network conditions in order to switch to its best operation point given a particular network condition, producing the highest possible user perceived quality. Current research is investigating a new scalable and flexible coding algorithm, and ways to adapt existing codec to heterogeneous environment such as Internet and WLAN. We present in this subsection a non-exhaustive list of video codec that are widely used in packet video applications.

Basically, most video codec are able to generate either Constant Bit Rate (CBR) [10] or Variable Bit Rate (VBR) traffic [11]. However, coded video sequence typically has time-varying complexity which cannot be generally achieved by a CBR coding. For instance, during scene changing, the predictive (Inter) coding involves high data rate. Therefore, VBR coding produces a constant visual quality while CBR coding produces a rather time-varying quality. In variable network channel condition, it is important for streaming system to match the available bandwidth.

2.2 Current Dynamic SL Negotiation Approaches

Several protocols for service level negotiation have been proposed, such as Common Open Policy Service for Service Level Specification (COPS-SLS) [12], which is an extension of the Common Open Policy Service (COPS) [13] protocol. While such an extension to COPS does enable a network entity to negotiate, it increases the complexity of the already sophisticated protocol. Hence, it remains to be seen if COPS-SLS will be viable for devices with limited resources such as PDAs. The Resource Negotiation And Pricing (RNAP) protocol developed by Wang et al [14] enables a user/service provider to dynamically negotiate a contracted service, allowing price and transmission parameters to be adjusted according to changes in network conditions and user requirements. RNAP requires the routers along the signaling path to maintain a *soft state*, resulting in an increased storage overhead. Furthermore, the protocol necessitates a host to periodically send messages to refresh the soft state, wasting bandwidth and energy, both of which are at a premium in wireless networks. In addition, this protocol does not take mobility into consideration. As a result, whenever a mobile moves to a new location, additional signaling is required between the user and the network for establishing an already negotiated service. Service negotiation protocol (SrNP) [15] is a protocol dedicated for service negotiation in wired networks. This protocol is not specific to any SLS format and is general enough to be applied for negotiating any document which in the form of attribute-value pairs. On account of the generality, SrNP messages could potentially have computationally expensive, verbose textual encodings which affects the protocol's applicability for devices with limited capabilities.

Simple Inter-domain Bandwidth Broker Signaling (SIBBS) [16] is a protocol proposed by the QBone group to enable communication between two bandwidth broker peers in adjacent domains. This protocol requires the signaling end-points to maintain long lived TCP connections, and therefore may not be suitable for a wireless environment with mobile hosts. IETF's Next Step in Signaling (NSIS) charter has recently proposed a QoS signaling protocol referred to as .QoS NSIS Signaling Layer Protocol (QoS-NSLP) [17]. QoS-NSLP too uses soft-state, peer-to-peer refresh messages as the primary state management mechanism. As said earlier, such periodic refresh messages consume both wireless bandwidth as well as the battery power of a wireless device. Hence QoS-NSLP may not be well suited for wireless environments.

Furthermore, two protocols have been proposed to support QoS negotiation in wireless networks by considering users' mobility, namely, QoS Generic Signaling Layer Protocol (QoS GSLP) [18] and Dynamic Service Negotiation Protocol (DSNP) [19]. QoS GSLP uses mobility and traffic pattern prediction to prefigure the next point of attachment of a mobile user and delivers the SLS to that access point, reducing thereby the handoff negotiation delay. This method highly increases the complexity of the system and makes it shortly scalable. DSNP informs all neighboring Base Stations (BSs) of the current BS of the SLS of a user. Each time the user negotiates for a new SLS, the QoS Global Server (QGS) delivers the new SLS to the current BS and its neighbors. In this fashion, all potential points of attachment after the users' handoff already have information on the users' SLS. This mechanism presents scalability problems in terms of signaling overhead and data storage.

There are other works in the open literature that discuss QoS negotiation [20][21][22]. They primarily focus on devising strategies for accepting or rejecting a request so that the system utilization is maximized. They do not describe the mechanics of QoS negotiation. Our work is not concerned with devising optimal strategies for admission control. We aim at developing a standard methodology for service negotiation. A comparison of some service negotiation protocols can be found in [23].

In addition to the above, several link layer service negotiation protocols have also been proposed. 3GPP has proposed a protocol for dynamically negotiating and re-negotiating the RAB service in 3GPP networks [24]. However, it is a link-layer protocol and can be used only for negotiating services over the Iu interface in 3GPP networks. At the IP layer, the *PDP (Packet Data Protocol) Context Modification* in 3GPP networks can modify the QoS profile of an active session [25][26]. However, the PDP context is confined between mobile station and Gateway GPRS Support Node (GGSN) only. It cannot be used for end-to-end service negotiation in heterogeneous networks. Both RAB negotiation and PDP context modification can be used only in 3GPP networks, and cannot be used in different wireless environments.

2.3 Desirable Characteristics of a Service Negotiation System

In this section, we delineate the essential characteristics of any system for dynamic service negotiation.

- **Transparency to link layer technologies;** Although diverse wireless communications systems exist, they essentially consist of several *Radio Access Networks (RANs)* and a *Core Network*. A RAN provides radio resources (e.g., radio channels) for mobile users to access the core network. The core network is typically a wireline network used to interconnect RANs and to connect the RANs to other networks such as the Internet. It is expected that different kinds of RANs, each optimized for distinctive environments and service requirements, will coexist in the future. To inter-operate and support universal roaming, the core network potentially will be based on IP.

- **Generic QoS Architecture; t**raffic between the end hosts may have to pass through networks owned by several service providers. Hence for compatibility, the protocol used for end-to-end service negotiation should adopt a QoS architecture that is commonly used.
- **Light-weight;** the service negotiation protocol will be used across devices with varying capabilities in terms of battery, computing power, and memory. Therefore, it should be lightweight. Extending protocols that are originally developed for other purposes to do service negotiation, might become too complex for a mobile device, when compared to a protocol dedicated just for this purpose.
- **Reduced Signaling Overhead;** the protocol should be scalable in terms of the signaling required between the mobile and the service provider. For example, consider a mobile enjoying a certain service that has moved to an adjacent network/cell with enough resources. The mobile should not be required to negotiate for the same service again, just because it has moved to a new network/cell. In addition, the protocol should not demand periodic signaling between network entities to refresh a service that has already been agreed upon. Signaling consumes precious resource like bandwidth and battery power, and hence should be minimized.

2.4 IP QoS Network Management

Currently QoS is not widely deployed. But with the push for applications such as multicast, streaming multimedia, and Voice over IP (VoIP) the need for certain quality levels is more inherent. Especially because these types of applications are susceptible to jitter and delay and poor performance is immediately noticed by the end-user. Internet Service Providers can proactively manage new sensitive applications by applying QoS techniques to the network. However, QoS is not the solution to every congestion problem. It may very well be that upgrading the bandwidth of a congested link is the proper solution to the problem.

Among the applications for QoS we can find: To give priority to certain mission critical applications in the network, to maximize the use of the current network resources, better performance for delay sensitive applications such as Voice and Video, and to respond to changes in network traffic flows

QoS can be broken down into three different levels, also referred to as service models. These service models describe a set of end-to-end QoS capabilities. End to end QoS is the ability of the network to provide a specific level of service to network traffic from one end of the network to the other. The three service levels are best-effort service, integrated service, and differentiated service. The following sections discuss each service model in detail.

2.4.1 Best Effort Services

Best-effort service, as its name implies, is when the network will make every possible attempt to deliver a packet to its destination. With best-effort service there are no guarantees that the packet will ever reach its intended destination [27]. An application can send data in any amount, whenever it needs to, without requesting permission or notifying the network. Certain applications can thrive under this model. FTP and HTTP, for example, can support best-effort service without much hardship. This is, however, not an optimal service model for applications which are sensitive to network delays, bandwidth fluctuations, and other changing network conditions. Network telephony applications, for example, may require a more consistent amount of bandwidth in order to function properly. The results of best-effort service for these applications could result in failed telephone calls or interrupted speech during the call.

2.4.2 Integrated Services

The integrated service (IntServ) model provides applications with a guaranteed level of service by negotiating network parameters end-to-end. Applications request the level of service necessary for them to operate properly and rely on the QoS mechanism to reserve the necessary network resources prior to the application beginning its transmission. It is important to note that the application will not send the traffic until it receives a signal from the network stating that the network can handle the load and provide the requested QoS end-to-end [28]. To accomplish this, the network uses a process called admission control. Admission control is the mechanism that prevents the network from being overloaded. The network will not send a signal to the application to start transmitting the data if the requested QoS cannot be delivered. Once the application begins the transmission of data, the network resources reserved for the application are maintained end-to-end until the application is done or until the bandwidth reservation exceeds what is allowable for this application. The network will perform its tasks of maintaining the per-flow state, classification, policing, and intelligent queuing per packet to meet the required QoS.

There are two features to provide integrated service in the form of controlled load services. They are Resource Reservation Protocol (RSVP) and intelligent queuing. RSVP is currently in the process of being standardized by the Internet Engineering Task Force (IETF) in one of their working groups.

RSVP is a protocol used to signal the network of the QoS requirements of an application. RSVP is not a routing protocol. RSVP works in conjunction with the routing protocols to determine the best path through the network that will provide the QoS required. RSVP enabled routers actually create dynamic access lists to provide the QoS requested and ensure that packets are delivered at the prescribed minimum quality parameters.

The drawback of the IntServ approach is scalability. The IntServ needs to detect each single flow and both packet scheduling and buffer management act on per-flow basis. Cost and complexity of the control system increase with the number of flows because there is no single label to identify a flow or a group of flows with similar performance requirements and traffic features. Moreover, RSVP signaling does not have any release message, so a periodic refresh message is needed to confirm the resource (bandwidth) request; a connection is dropped only if the refresh message has not been received for some time. Thus, even if the IP flow is terminated, the resources are not immediately released. Additionally, refreshing signaling is bandwidth consuming. Even if the IntServ approach has clear drawbacks, its features allow assuring a specific QoS to each flow.

2.4.3 Differentiated Services

The Differentiated Services (DiffServ) have been proposed to cope with the scalability problem faced by Integrated Services. The solution uses the DSCP (DiffServ Code Point) field of the IP packet header. The 6-bit DSCP field specifies the forwarding behavior that the packet has to receive within the DiffServ domain of each operator. The behavior is called “Per Hop Behavior” (PHB) and it is defined locally. It is not an end-to-end specification but it is strictly related to a specific domain. The same DSCP may have two different meanings in two different domains. Negotiations between all adjacent domains are needed to assure a correct end-to-end forwarding behavior.

The DiffServ [29] approach does not distinguish each single user flow throughout the network. The traffic is classified and aggregated in different traffic classes, each of them individuated by a label provided by setting bits in the DSCP field. The identification is performed at the network edges. Within the network core, packets are managed according to the behavior associated with the specific identification label. DSCP is assigned on the base of the TCP/IP header content by a router called “Edge Router”, which is the entrance gate to the DiffServ network. In theory, DSCP might also be assigned directly by the source but, in regular DiffServ networks, the operation is performed by Edge Routers.

The class selector PHB offers three forwarding priorities:

Expedited Forwarding (EF) characterized by a minimum configurable service rate, independent of the other aggregates within the router, and oriented to low delay and low loss services.

Assured Forwarding (AF) consisting of 4 independent classes (AF1, AF2, AF3, and AF4) although a DiffServ domain can provide a different number of AF classes. Within each AF class, traffic is differentiated into 3 “drop precedence” categories. The packets marked with the highest drop precedence are dropped with lower probability than those characterized by the lowest drop precedence.

Best Effort (BE), which does not provide any performance guarantee and does not define any QoS level.

2.5 The proposed Dynamic Service Level Negotiation System

This section provides a description of the service level negotiation system, which aims to guarantee the continuity of the service level perceived by user over wireless networks. To achieve this goal, the users should negotiate the resources required by their applications, and the network should guarantee the negotiated resources for the entire course of the users’ sessions.

The proposed dynamic QoS negotiation system allows users to define and request their desired service levels, which can be accepted or rejected by the service negotiation entity. In case of rejection, the service negotiation entity proposes a different service level to MS. MS accepts or rejects such an offer. Moreover, at any time an user can upgrade or downgrade a previously negotiated service level. On the other hand, the service negotiation entity may require degrading the service level when resources become scarce. Disregarding the situation, a new Service Level Specification (SLS) is established when both user and the service negotiation entity receive positive responses from each other. Some important aspects of the QoS negotiation system are:

- As to the bottleneck for most wireless communications is the last hop; from the BS to the MS, the SL negotiation system focus on the service provided by the base station to the mobile station. Thus, after a successful service level negotiation, the BS allocates resources guarantee the negotiated SL to the MS.

- The SLS is valid within the current domain only, the MS may move freely within the domain in which it performed SL negotiation and keep receiving the same service level, but if the MS move to a different domain, it need to negotiate a new SL with the negotiation server of the new domain. The reason for that limitation is that any internet service provider may guarantee the service level beyond the borders of it own network.
- The QoS parameter considered is bandwidth. To guarantee a strict service level for MSs instead of predefined traffic classes where different flows from the same traffic class may still receive the same Service level even if they have different bit rates. The proposed SL negotiation system focuses on the allocation of bandwidth to satisfy the negotiated Sl of each MS.

2.5.1 SLS Model

A main feature of the service level negotiation system is the capability to set up a QoS Service Level Specification (SLS) between networks and users. The SLS model shall be used to define a set of possible parameters as well as valid values and options for them. As the SL negotiation system assumes a heterogeneous environment, the SLS Model must be uncoupled from the underlying QoS network technologies. The QoS requirements included in a SLS must be expressed in a generic template, which will also facilitate the automatic translation of the QoS requirements for use in the specific control planes of each network. The specification of the SLS Model assumes that a SLS is restricted to an bidirectional agreement between a network and a user for a specific QoS level.

An SLS provides an assertion that user’s traffic will be carried by the network with one particular QoS guarantees. However, this will occur only if the user’s traffic meets certain stated traffic descriptor.

The traffic descriptor includes parameters that will be used for the configuration of the QoS legacy mechanisms, such as policing, admission control, and resource management. The traffic descriptor also includes the statistical source characterization and a way to identify the traffic belonging to the service.

The QoS guarantees category captures the nature of the service that is going to be provided, that is, the QoS level that the provider offers to the customer. Parameters such as bandwidth, end-to-end delay and packet loss ratio, etc., are included in this category.

Therefore, when designing the SLS template, the approach of network guarantees and user requirements has been applied:

• The user will describe the service requirements by means of a traffic descriptor and will inform the network about the acceptable values for the different parameters describing the QoS guarantees

• The network will accept the traffic descriptor and the desired QoS guarantees or will propose alternative QoS guarantees to ensure the traffic descriptor.

In addition to the *traffic descriptor* and the *QoS guarantees*, an SLS includes other categories, such as the *geographical and temporal scopes* of the service being specified. Additionally, the SLS template is open to include additional categories such as compensation and charging parameters.

SLSs can be established for different levels of abstraction from supporting an individual flow to describing the long-term traffic engineering relationships between user and network. This is all achieved using a common negotiation protocol. It can happen at the start of service or anytime after the service level negotiation by updating a previously agreed SLS.

Network administrators will configure through policies which kind of services the network can provide, the set of parameters and value ranges that must be included in the SLSs used to negotiate the provision of those services. With this approach, the SLS negotiation will be based on performance parameters of the traffic flow: throughput, bandwidth, packet losses, delay, etc.

2.5.2 Negotiating, realization and releasing QoS

Users are able to automatically and dynamically establish QoS agreements with the network based on SLSs. This dynamic and automatic establishment of QoS agreements involves the following sequential steps:

QoS negotiation, performed based on proposals and counterproposals of SLSs between the network and the user.

QoS realization, involve the allocation of resources to guarantee the negotiated QoS.

QoS release, release of resources previously reserved for established SLSs that are no longer needed.

Next subsections address the negotiation, realization and release of QoS resources focusing on the envisioned architecture for QoS negotiation over wireless networks.

2.5.3 Envisioned Architecture

The QoS negotiation takes place at the IP layer and is based on Differentiated Services (DiffServ). The components of the envisioned architecture for dynamic QoS negotiation are schematically shown in Fig. 2.1. The figure depicts one of the multiple domains administered by different ISPs and offering services through different wireless technologies.

The domain consists of a QoS global server (QGS), an authentication, authorization, and accounting (AAA) server, a network proxy, a number of base stations (BSs), and a population of mobile users, termed henceforth as mobile stations (MSs).

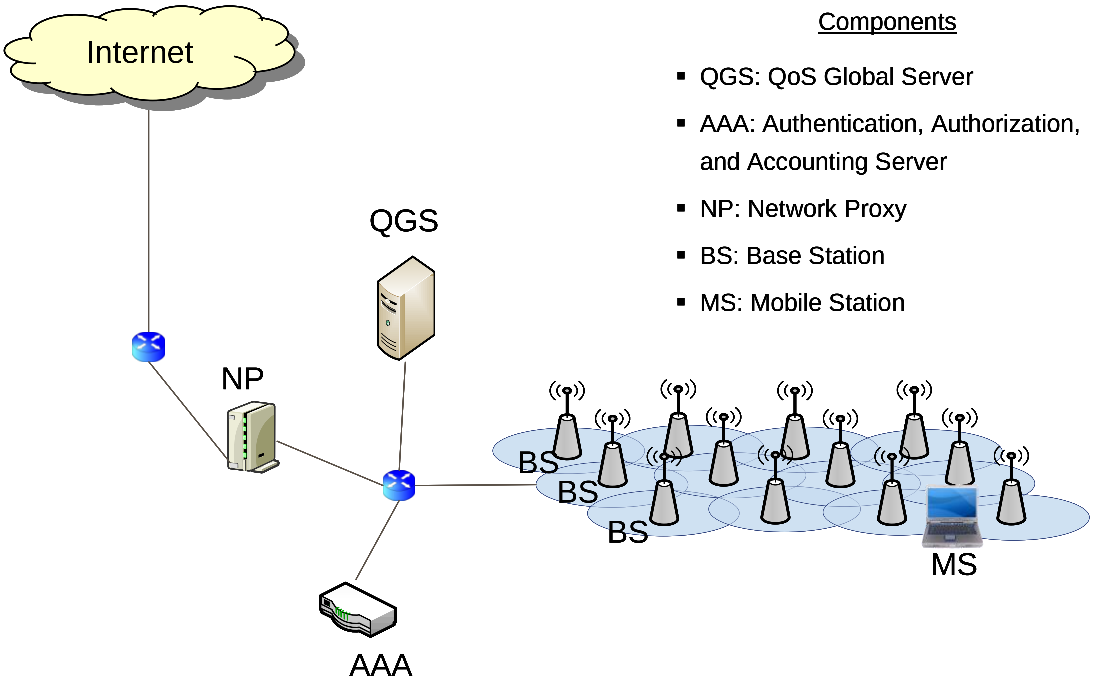

Figure 2.1 Envisioned architecture for dynamic service level negotiation

QGS is the entity for service level negotiation; it decides the admissibility of service requirements based on the service level that the user is allowed to receive as well as the current available resources in the network [30].

The AAA server is used to confirm that a user, who is requesting a specific service level, is permitted to obtain it.

BS is the entity where SLSs are applied. It enforces different service levels to users. BSs inform the QGS about their local resource availability and receive SLSs of users for traffic conditioning.

MS is the device that allows users to communicate, and also provides means of interaction between users and the networks. Traffic is generated/received by MS and may be queued in the MS while waiting for transmission/reception. The MS interacts with the QGS when it requests certain degree of QoS in the domain.

2.5.4 QoS Negotiation

There are two procedures to perform service level negotiation. Initial Service level negotiation and service level renegotiation. When a MS is powered up, it needs to perform initial service level negotiation with the network. Service level renegotiation is required when MS is receiving service from the network and one of the following three scenarios emerge: First, the service level requirements of the MS changes. Second, the network resources become scarce and The QGS requires the MSs to degrade their service levels. And thirds, when a MS performs handoff and the available resources in the new BS are not enough to satisfy the service level of the MS.

For the first two scenarios, MS keeps receiving the same SL until the new SL is successfully agreed. On the other hand, for the thirds scenario, the service is stopped until the new SLS is successfully agreed.

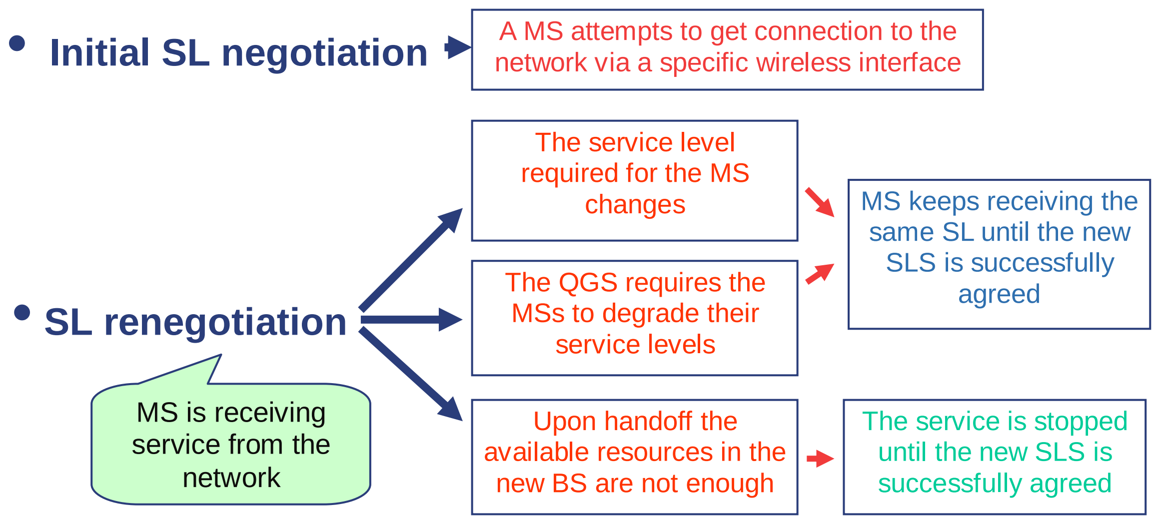

Figure 2.2 Service level negotiation procedures

2.5.4.1 Initial SL Negotiation

Upon connecting to the network, a MS negotiates with the QGS the service level required for the application it attends to execute. Firstly the MS searches for a BS it can communicate through, and then requests predefined services available in the network. When the MS obtains the requested information, it sends a service negotiation request; the request is received by the BS and forwarded to the QGS.

The QGS consults with the AAA server to verify whether the MS is authorized to receive the requested service. In case of acceptance, the QGS sends the new SLS to the corresponding BS in order to perform traffic conditioning. The QGS also notifies the successful service level negotiation to the MS via the BS. The BS allocates resources to the MS for data transmission and notifies the NP about the new path for data transmission belonging to the MS. Right after that, the MS starts using the service. This procedure is conceptually depicted in Fig. 2.3. If the MS is not authorized to acquire the requested service or there are not enough resources to satisfy it, the request is rejected and a negative negotiation response is sent to the MS, which includes the reasons for turning down the request and the available resources that the MS can currently negotiate for.

Even when an MS is able to negotiate SLSs through all its interfaces, each SLS is associated to one specific interface. Thus, MS should perform an initial negotiation through each interface it attempts to use to connect to the network.

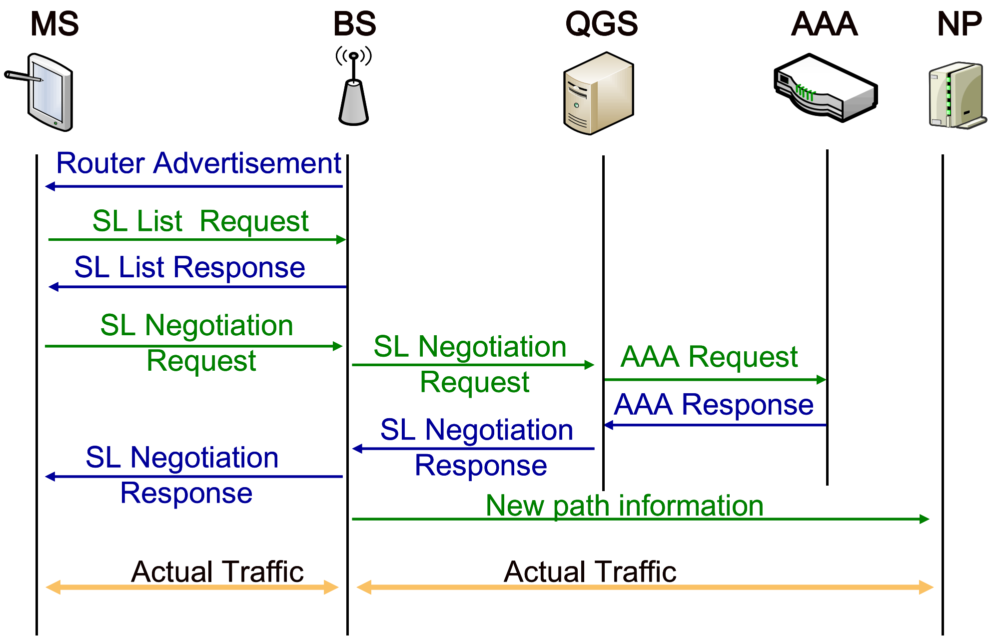

Figure 2.3 Initial Service Level Negotiation

2.5.4.2 SL Renegotiation

Service level renegotiation is required when a MS is currently receiving services from the network and one of the following three cases occurs: *i)* the service requirement of the MS changes, *ii)* the resources in the network become scarce and the QGS requires the MSs to degrade their existing SLSs, and *iii)* the MS performs handoff and the available resources in the new subnet are not enough to guarantee the current SLS.

For the two first cases, the renegotiation is similar to the initial QoS negotiation procedure apart from the fact that the MS keeps receiving services during the renegotiation period. If the QGS rejects the new service level requested by the MS, its current service level is retained. On the other hand, in the third case, the service is stopped until a new SLS is successfully negotiated.

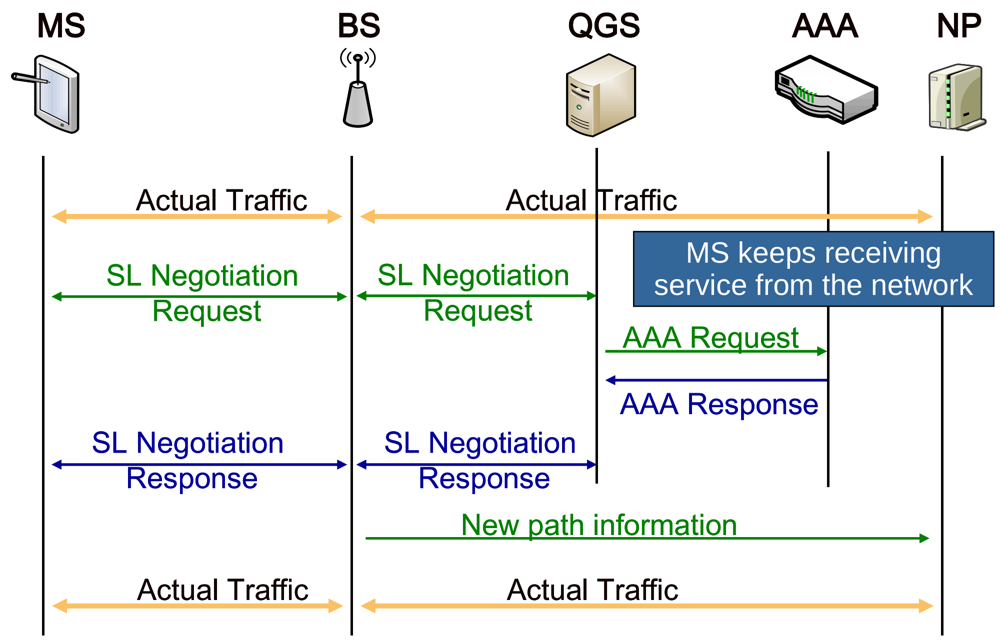

Figure 2.4 Service Level Renegotiation (case 1 and case2)

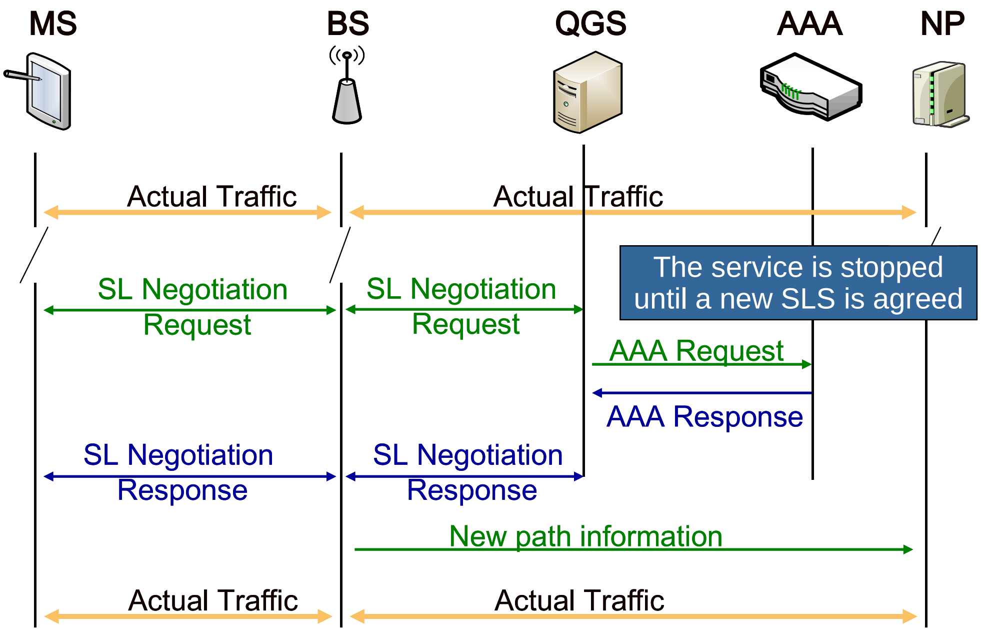

Figure 2.5 Service Level Renegotiation (case 3)

2.5.4.3 SL negotiation messages

- *SL List Request:* This message is sent by a MS to the QGS to request for a list of SLs offered by the network. The MS sends this message when it does not have knowledge about the service levels that the network can support.

- *SL List Response:* This message is sent in response to the *SL List Request* message, by the QGS. This message lists all the SLs that are provided by the network. The time of availability for each service may also be included in the list.

- *SL Negotiation Request:* This message is usually sent by a MS to the QGS to request a particular SL. The requested SL could be one of those listed in the SL List Response message. Alternatively, it could also be customized to include a set of predefined SLs. This message is used in both requesting for a new SLS as well as for updating an existing one. This message can be sent by both the MS as well as the QGS. For example, if resources in the network become scarce, the QGS sends this message to the MSs requesting them to degrade their existing SLS to suit the current network conditions. Additionally, if the network wants to forcefully terminate a SLS of a MS due to some reason, it sends a SL Negotiation Request message to the user with appropriate fields set to *ZERO*.

- *SL Negotiation Response:* This message is sent in response to the SL Negotiation Request. This message indicates whether the requested SL is accepted or not. If the requested SL is not accepted, then the reason for not accepting is also provided. For example, if the

QGS does not accept the SL of a MS due to lack of resources, it sends back a response indicating a reject along with the maximum SL that could be supported. Including such information might induce the MS to negotiate again for a different service.

2.5.5 QoS Realization

After a successful negotiation of a SLS between a MS and the network, the BS enforces the negotiated service level to the user.

2.5.5.1 Multiplexing Strategies

In this section we give a rough overview about several multiplexing techniques. These describe how several independent data channels have access to a single physical signal carrier medium like the air for wireless transmission systems. In general the multiplexing schemes are based upon time, frequency and code.

- **CDMA Code Division Multiple Access:** CDMA refers to any of several protocols used in so-called second-generation (2G) and third-generation (3G) wireless communications. CDMA is a form of multiplexing, allowing numerous signals to use a single transmission channel, optimizing the use of available bandwidth. The technology is used in ultra-high-frequency (UHF) cellular telephone systems in the 800-MHz and 1.9-GHz bands.

CDMA employs analog-to-digital conversion (ADC) in combination with spread spectrum technology. Audio input is first digitized into binary elements. The frequency of the transmitted signal is then made to vary according to a defined pattern (code), so it can be intercepted only by a receiver whose frequency response is programmed with the same code. The CDMA channel is nominally 1.23 MHz wide. CDMA networks use a scheme called soft hand-off, which minimizes signal breakup as a handset passes from one cell to another. The combination of digital and spread-spectrum modes supports several times as many signals per unit bandwidth as analog modes. CDMA is compatible with other cellular technologies which allows for nationwide Roaming.

The original CDMA standard [31], also known as CDMA One and still common in cellular telephones in the US, offers a transmission speed of only up to 14.4 Kbps in its single channel form and up to 115 Kbps in an eight-channel form. CDMA2000 and Wideband CDMA (WCDMA) deliver data many times faster.

- **FDMA Frequency Division Multiple Access:** FDMA is the division of the frequency band allocated for wireless cellular telephone communication into 30 channels, each of which can carry a voice conversation or, carry data of a digital service. FDMA [32] is a basic technology in the analog Advanced Mobile Phone Service (AMPS), the most widely-installed cellular phone system installed in North America. With FDMA, each channel can be assigned to only one user at a time. The Digital-Advanced Mobile Phone Service (D-AMPS) also uses FDMA but adds time division multiple access (TDMA) to get three channels for each FDMA channel, tripling the number of calls that can be handled on a channel.

- **TDMA Time Division Multiple Access:** TDMA [33] is a technology used in digital cellular telephone communication that divides each cellular channel into time slots in order to increase the amount of data that can be carried. TDMA is used by Digital-American Mobile Phone Service (D-AMPS), Global System for Mobile communications (GSM), and Personal Digital Cellular (PDC). However, each of these systems implements TDMA in a somewhat different and incompatible way. An alternative multiplexing scheme to FDMA with TDMA is CDMA, which takes the entire allocated frequency range for a given service and multiplexes information for all users across the spectrum range at the same time.

TDMA was first specified as a standard in EIA/TIA Interim Standard 54 (IS-54). IS-136, an evolved version of IS-54, is the United States standard for TDMA for both the cellular (850 MHz) and personal communications services (1.9 GHz) spectrums. TDMA is also used for Digital Enhanced Cordless Telecommunications (DECT).

- **OFDM Orthogonal Frequency Division Multiplexing:** OFDM [34] is fundamentally different from other modulation schemes because it may be transmitted via AM, FM, QAM (Quadrature Amplitude Modulation), and so on. OFDM is defined as a mathematically technique for the generation and demodulation of radio waves.

Frequency division multiplexing (FDM) is a technology that transmits multiple signals simultaneously over a single transmission path, such as a cable or wireless system. Each signal travels within its own unique frequency range (carrier), which is modulated by the data (text, voice, video, etc.). Orthogonal FDM's (OFDM) spread spectrum technique distributes the data over a large number of carriers that are spaced apart at precise frequencies. This spacing provides the "orthogonality" in this technique which prevents the demodulators from seeing frequencies other than their own. The benefits of OFDM are high spectral efficiency, resiliency to RF interference, and lower multi-path distortion. This is useful because in a typical terrestrial broadcasting scenario there are multipath-channels (i.e. the transmitted signal arrives at the receiver using various paths of different length).

The multicarrier transmission technique OFDM is seen as a key technology for high-rate communications and is already part of the IEEE 801.11 and ETSI BRAN standard for wireless local area networks. OFDM became a serious alternative by applying modern, digital signal processing methods based on the Fast Fourier Transform (FFT). The main advantages are the high spectral efficiency, the simple channel equalization and the suppression of inter-symbol interference. However OFDM suffers from a high peak-to-average power ratio of the transmitting signal. The unfavorable PAR prohibits power efficient operation of the amplifier which is especially intolerable in portable systems. Thus OFDM is not confined to mobile communications but is used in Digital Audio Broadcast (DAB), Digital Video Broadcast (DVB) and xDSL systems.

- **COFDM (Coded OFDM):** COFDM is an expansion of the already available OFDM modulation technique [35]. The special performance of COFDM with respect to multipath and interference, burst errors and fading are the reasons why COFDM is well-suited to the needs of terrestrial broadcasting channels. COFDM is resistant to multipath effects because it uses multiple carriers to transmit the same signal. Thus COFDM has been chosen for the two standards DAB (Digital Audio Broadcast) and DVB-T (Digital Video Broadcast-Terrestrial). COFDM is ideal for single frequency networks.

- **WCDMA Wide Band Code Division Multiple Access:** WCDMA is an ITU standard and was derived from the CDMA standard. WCDMA is a third-generation (3G) mobile wireless technology offering much higher data speeds to mobile and portable wireless devices than older technologies. WCDMA can support mobile/portable voice, images, data, and video communications at up to 2 Mbps (local area access, no movement) or 384 Kbps (wide area access, movement in trains, cars, etc.) [36]. The input signals are digitized and transmitted in coded, spread-spectrum mode over a broad range of frequencies. A 5 MHz-wide carrier is used, compared with 200 kHz-wide carrier for narrowband CDMA.

- **SDMA Spatial Division Multiple Access:** SDMA has proven to be an interesting option for capacity increase of wireless communications systems. The idea is to allow several users to use the same frequency band (and time slot) simultaneously and to identify them from their positions. SDMA makes use of antenna-array processing and advanced digital signal processing techniques which rely heavily on concepts from linear algebra [37]. The matrix decompositions involved are complex, mostly O^{n^3}, with n the problem size. Moreover, data have to be processed at a high rate (e.g. GSM 270 kbits/s), so that the computational requirement is in the Mflops/s or even Gflops/s range. Hence there is a strong need for efficient algorithms which can be obtained by using adaptive matrix decomposition techniques.

2.5.5.2 Multiplexing Strategies in Wireless Broadband Communication Networks

Here we discuss the current and future broadband communication networks that can support real-time applications. The channel used here is a general concept representing the smallest unit of radio resources that a user can be assigned for transmitting data, such as frequency band and time slot. Taking into consideration the different channel conditions and users' transmission requirements, dynamic channel assignment can improve the utilization of system resources by exploring the diversities over multiuser, time, and frequency. Scheduling and random access are two special types of channel-assignment schemes that enable multiple users to take turns occupying the limited radio resources over time. Scheduling is a centralized control usually applied in cellular network to determine which user can transmit at a specific time. In contrast, random access can reduce transmission delay in lightly loaded networks such as WLANs and provide an autonomous way to avoid conflict of resource usage.

- ***3G Cellular Networks: CDMA/Scheduling:*** 3G wireless communication systems employ code-division multiple access (CDMA). CDMA uses unique spreading codes to spread the baseband data before transmission. The signal occupies a much broader bandwidth than narrow-band transmission and is transmitted with power density below noise level. The receiver uses a correlator to despread the signal of interest, which is passed through a narrow band-pass filter to remove unwanted interferences. This brings many benefits, such as immunity to narrowband interference, jamming, and multiuser access.

Fig. 2.6 on shows the CDMA system with scheduler to allocate resources to users. Widely adopted in 3G networks, the scheduler allocates a different number of CDMA codes or CDMA codes with various spreading factors to users at different times according to the channel conditions, QoS types, bandwidth requirements, and buffer occupancies.

- ***WLAN: OFDM, CDMA/Random Access:*** WLAN can provide a higher transmission rate within local areas. There are two major current standards for WLAN, namely, IEEE 802.11b and IEEE 802.11g. IEEE 802.11b uses CDMA technology and supports up to 11 Mb/s; and IEEE 802.11g uses orthogonal frequency-division multiplexing (OFDM) technology and supports up to 54 Mb/s. OFDM splits a high rate data stream into a number of lower-rate streams and transmits them over a number of frequency subcarriers simultaneously. In addition, guard time with cyclical extension is inserted in each OFDM symbol. Thus, inter-symbol and inter-carrier interference are almost eliminated in OFDM systems. Fig. 2.7 illustrates multiple accesses in the current IEEE 802.11 standard. The IEEE 802.11 MAC protocol supports two access methods: the distributed coordination function (DCF) and the point coordination function (PCF). At each time slot, both functions only allow one user to occupy all radio resources. The DCF is the basic random-access mechanism using carrier-sense multiple access with collision avoidance (CSMA/CA) or ready-to-send (RTS)/clear-to-send (CTS). In contrast, the PCF is based on polling controlled by a point coordinator.

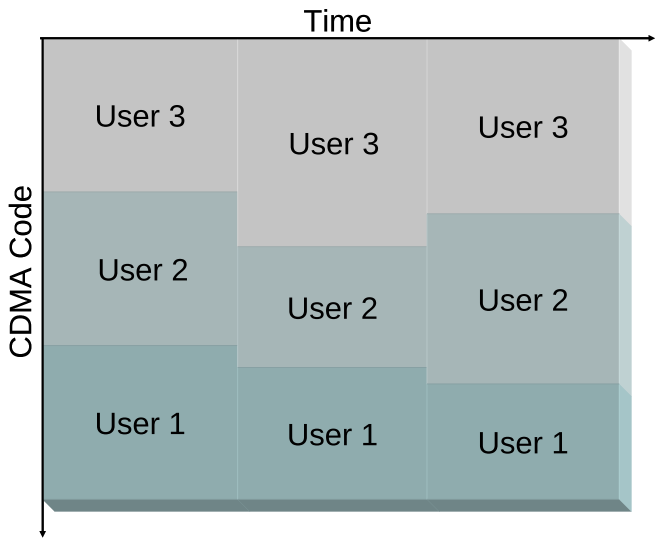

Figure 2.6. Multiplexing strategies for 3G

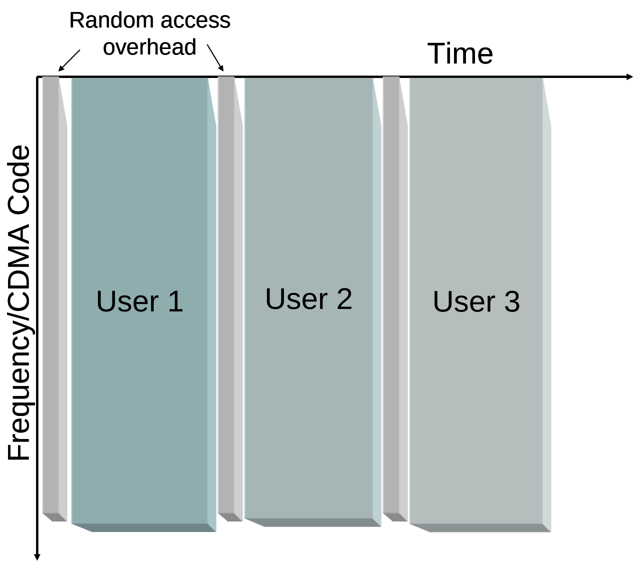

Figure 2.7. Multiplexing strategies for WLAN

- **4G Cellular Networks, WLAN and WiMax/ OFDMA:** In current OFDM systems, all subcarriers are assigned to a single user at each moment, and multiple users are supported through time division. However, for a given subcarrier, different users experience different channel conditions and the probability for all users to have deep fades in the same subcarrier is very low. Orthogonal frequency-division multiplexing access (OFDMA) allows multiple users to transmit simultaneously on the different subcarriers, while each subcarrier is assigned to the user who is experiencing a good channel condition. Fig. 2.8 shows the multiuser resource-allocation strategy for OFDMA system. Users' transmission can be allocated to different time-frequency slots. By doing this, the multiuser, time, and frequency diversity can be fully explored to improve the system performance.

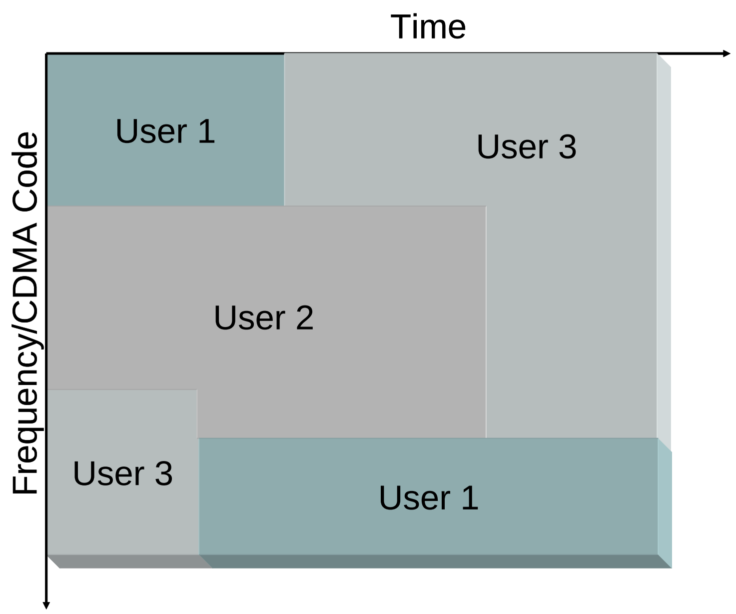

Figure 2.8. Multiplexing strategies *for WiMax.*

2.5.5.3 Multiplexing Strategies in the SL Negotiation System

Since, the multiplexing strategies assign codes and/or frequencies to users over time, and multiple users are supported through time division, time-code division and time-frequency division. The BS in the service level negotiation system implement a time-slot division approach, which allows each MS to use the channel during the time-slots assigned to the MS. Thus, BSs avoid collisions in the wireless channel and provide users with strict QoS, rather than relative QoS where data packets belonging to users with a similar service level compete among them to get access to the link. Each MS is allocated a specific period of time to use the wireless link. However, the process to assign the specific time to MS to use the channel that ensure the negotiated amount of bandwidth, vary from one wireless technology to another, as they use different multiplexing strategies.

- **OFDM** represent the simplest case; because it allows only one user occupy the entire bandwidth at each time. Thus, the time-slot assigned to a MS depends on the amount of bandwidth specified in its SLS and the capacity of the channel. The time-slot size varies from one MS to another. It is calculated as shown in Equation 2.1.

$$\delta_i = \frac{BW_{SLS_i}}{BW_T} * \Delta \tag{2.1}$$

where BW_{SLSi} , BW_T , and Δ denote the bandwidth specified in the SLS of the MS_i, the total bandwidth of the wireless link from the BS, and the time-slot interval, respectively. The time-slot interval is the continuously repeated time period in which all MSs will be served.

- **CDMA** is able to provide bandwidth-on-demand platform by allocating multiple codes over time according to users' negotiated bandwidth. Thus, the time-slot assigned to a MS depends on the amount of bandwidth specified in its SLS, the capacity of the channel, and the number of codes assigned to the MS. The time-slot size varies from one MS to another and it is calculated as shows Equation 2.2.

$$\delta_i = \frac{BW_{SLS_i}}{BW_T * N} * \Delta \tag{2.2}$$

Where N represents the number of codes assigned to the MS_i.

- **OFDMA** provides bandwidth-on-demand by allocating multiple subcarriers over time according to users' negotiated bandwidth. Thus, the time-slot assigned to a MS depends on the amount of bandwidth specified in its SLS and the capacity of each subcarrier assigned to the MS. The time-slot size varies from one MS to another and it is calculated as shown in Equation 2.3.

$$\delta_i = \frac{BW_{SLS_i}}{\sum_{j=1}^{m} BW_j} * \Delta \tag{2.3}$$

Where m represents the number of subcarriers assigned to the MS_i.

Let's define slot synchronization delay as the time a packet has to wait at the BS queue since its arrival time until the beginning of the time-slot of its corresponding MS. Burst delay is in turn defined as the time a packet has to wait at the BS queue due to the bursty nature of real-time traffic. These packets have to wait for later time-slots to be transmitted. In addition, if the next packet to be transmitted is too large to be processed during the remaining of the time-slot, the packet will wait for the time-slot of its corresponding MS in the next round. This also means that the current time-slot will have an unused remainder at the end.

Based on the above discussion, an appropriate size of the time-slot interval should be carefully determined to decrease the synchronization delay and the burst delay, while keeping the network utilization high. In [38], it was demonstrated that the unused bandwidth at the end of time-slots does not depend on the time-slot size. It only depends on the distribution of packet sizes. The same work also demonstrated that:

$$U = \frac{\Delta - E(R)}{\Delta}; \qquad E(R) \le \Delta \tag{2.4}$$

where U and E(R) denote the network utilization and the expected value of the unused portion of the time-slot of a MS, respectively. Eq. 2.4 indicates that increasing the time-slot interval length yields better network utilization.

On the other hand, decreasing the time-slot interval size lowers the packet delay. Thus, the size of the time-slot interval should be appropriately selected to keep a balance between the network utilization and the communication delay. In [39] we empirically demonstrated that (Δ = 0.1s) achieves a low delay as well as high utilization of the bandwidth.

2.5.5.4 Ensuring Wireless Channel Bandwidth

The service level negotiation system provides guaranteed bandwidth at the congestion points of the network (the BSs), ensuring the negotiated bandwidth to the MSs. To perform this task, we review the queuing strategies for QoS and then present the queuing strategy adopted for the SL negotiation system.

- **PQ Priority Queue:** PQ ensures that important traffic gets the fastest handling at the BSs. It was designed to give strict priority to important traffic. Priority queuing can flexibly prioritize according to network protocol (for example IP, IPX, or AppleTalk), incoming interface, packet size, source/destination address, and so on. In PQ, each packet is placed in one of four queues; high, medium, normal, or low, based on an assigned priority. Packets that are not classified by this priority list mechanism fall into the normal queue (see Fig. 2.9). During transmission, the algorithm gives higher-priority queues absolute preferential treatment over low-priority queues.

PQ is useful for making sure that mission-critical traffic gets priority treatment. PQ currently uses static configuration and thus does not automatically adapt to changing network requirements.

- **CQ Custom Queue:** CQ was designed to allow various applications or organizations to share the network among applications with specific minimum bandwidth or latency requirements. In these environments, bandwidth must be shared proportionally between applications.

CQ handles traffic by assigning a specified amount of queue space to each class of packets and then servicing the queues in a round-robin fashion (see Fig. 2.10).

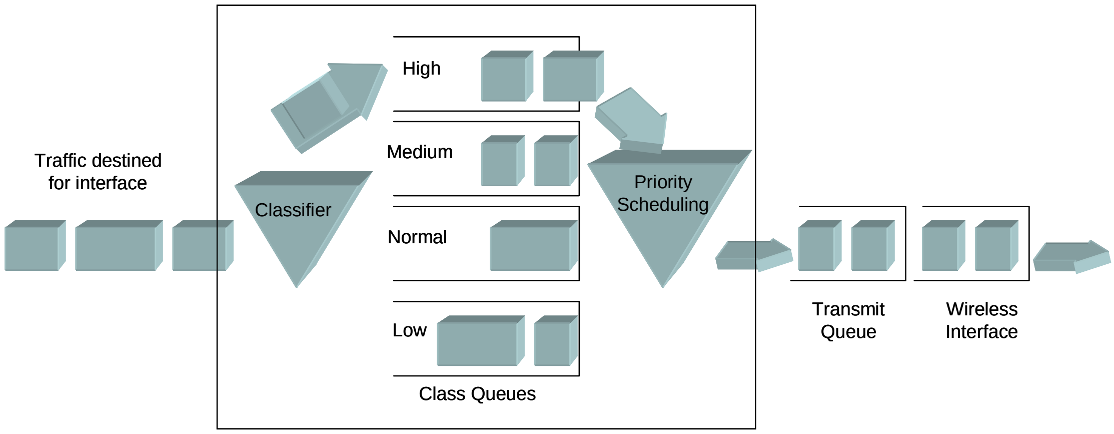

Figure 2.9 Priority Queuing

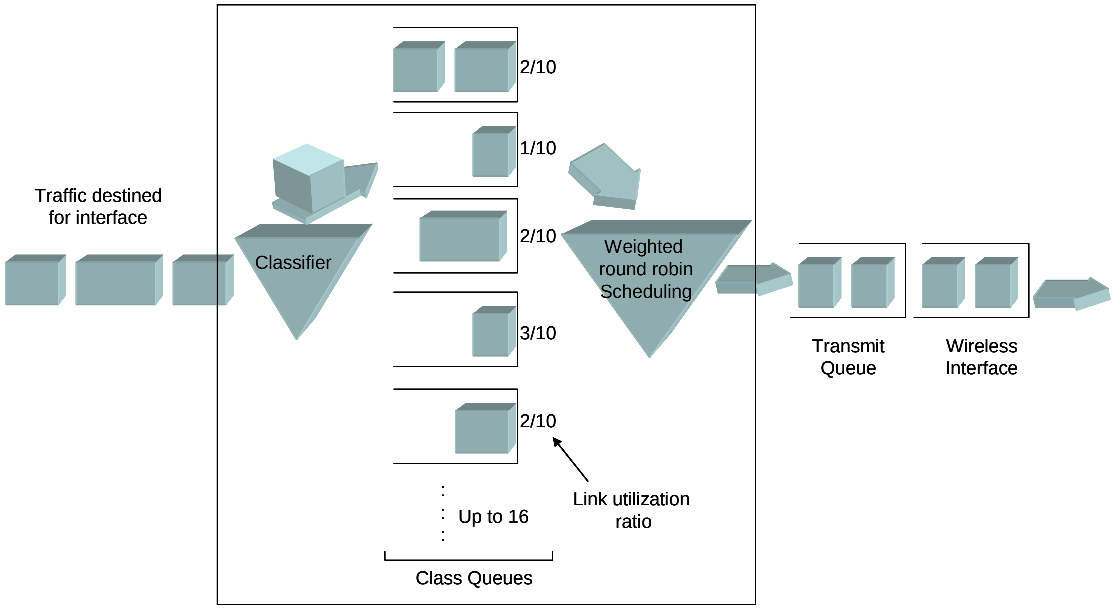

Figure 2.10 Custom Queuing

The queuing algorithm places the messages in one of 17 queues (queue 0 holds system messages such as keep alive, signaling, and so on) and is emptied with weighted priority. The router services queues 1 through 16 in round-robin order, dequeuing a configured byte count from each queue in each cycle. This feature ensures that no application (or specified group of applications) achieves more than a predetermined proportion of overall capacity when the line is under stress. Like PQ, CQ is statically configured and does not automatically adapt to changing network conditions.

2.5.5.5 Queuing Strategies in the SL Negotiation System

The dynamic service level negotiation system implements the Individual Queues (IC) at the BSs. IQ handles traffic by assigning a different queue to each MS's traffic and then serving the queues during the time-slot assigned to the corresponding MS to use the wireless channel. Each queue is served in round-robin fashion. In IQ there is a individual queue for each MS that negotiates for QoS, and a common queue for best-effort traffic; the traffic for the MSs that did not negotiate for QoS. Unlike PQ and CQ, IQ is dynamically configured and adapt to changes in the service level negotiated by the MSs. (see Fig. 2.11).

In this fashion, the system avoids collision in the wireless channel and guarantees the negotiated SL to the MSs.

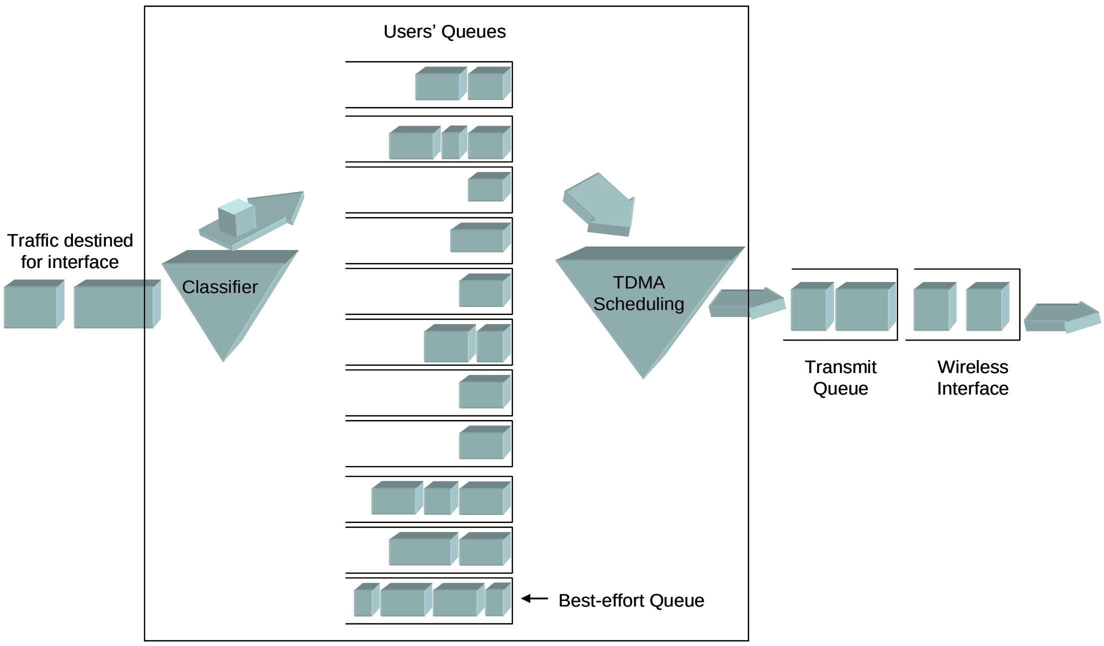

Figure 2.11 Individual Queuing

2.5.6 QoS Releases

There are three scenarios for releasing resources previously allocated at the BS. First, when the MS's application finish, the negotiated resources are not longer used, therefore the MS sends a SL Negotiation Request message to the QGS with appropriate fields set to *ZERO*. After that, the QGS order to the BS to release the resources allocated to that MS and to erase the MS' SLS. Additionally, The QGS updates the available resources at the BS on it's table and also erases the MS' SLS of it table. Second, when the MS moves to the coverage area of another BS, in this case, after the MS starts receiving service from the new BS, the MS send a message to the previous BS via the new BS informing about the handoff. Then the BS releases the resources allocated to that MS, erases the SLS of the MS, and informs the QGS about it new state of its available resources. And thirds, when the BS does not receive any signaling or traffic from the MS for a certain period of time. The BS releases the allocated resources to this MS, erases the SLS of the MS, and informs the QGS about it new state of its available resources.

2.6 System Scalability

The dynamic service level negotiation system have been designed to be highly scalable, by requiring low signaling for SL negotiation, SL renegotiation, and not signaling at all for maintaining the negotiated SLS.

Given the centralized nature of QGS, it is the more susceptible component to scalability. However, the proposed architecture separates signaling traffic from data traffic and the QGS is responsible for handling only the signaling traffic. When compared to data traffic, the signaling traffic is much smaller and hence QGS should be able to handle a large number of MSs. Such a centralized controller has been successfully employed in other IETF protocols such as Megaco [40], COPS [12], and Middlebox [41]. Therefore, implementing a centralized QGS is certainly scalable. Alternatively, to guarantee higher scalability of the system, we consider to distribute the population of BSs in several geographic group controlled by different QGSs installed around the domain.

2.7 Summary

In this chapter, we presented a detailed description of the service level negotiation system exclusively designed for wireless networks. The proposed system allows users to dynamically negotiate the SL required for their traffic whilst the network guarantees the negotiated SL during the entire course of the user's session. The system was designed to be generic to any wireless technology, light weight to be used by any capability connection device, and highly scalable.

We described the SL negotiation and renegotiation process, and explored the resource allocation strategies currently used by the most relevant wireless networks in order to select the resource allocation strategy to be adopted in our system, which should be supported by all the considered wireless technologies. The SL negotiation system provides strict QoS to users, by individually allocating bandwidth to each user.

Chapter 3

Mobility Management

In the previous chapter, we described the dynamic service level negotiation system and explain the way how the BSs allocate resources for MSs to guarantee the negotiated SL. In this chapter, we describe the mechanism to maintain the continuity of the service level perceived by MSs while they are on the move performing handoff between different BSs.

QoS and mobility functionalities are not independent. Coupling between the two functionalities occurs because QoS is tied to a specific path and paths change as a result of handoff. Without this coupling, when handoff occurs, the state becomes not associated with the new path. Thus there is no QoS on the new path and the unused QoS state on the old path affects the use of the network resources. Thus, QoS and mobility management should be coupled to ensure the continuity of service level perceived by MSs while they perform handoffs between different BSs. The handoff process can be in homogeneous networks (horizontal handoff) [42] [43], or between heterogeneous networks (vertical handoff) [44] [45]. The work in [46] presents an interesting classification of mobile applications based on their mobility management requirements and also investigates the handoff performance of the existing mobility management protocols for these applications.

In wired networks, it is easy to identify the edge router that has to do the traffic conditioning for a specific user. However, in wireless networks one of the most relevant issues is to track the location of an MS and inform the appropriate BS of the SLS of the MS. Based on the centralized nature of the QGS in the proposed dynamic SL negotiation system, where the QGS is the repository that retains the SLS of all the MSs in the domain, there are several ways to inform the new BS of MS's SLS. We can classify them into two categories as shown in Fig. 3.1; proactive and reactive strategies. Proactive strategies usually provide seamless handoff as they involve very short delay. However, they are shortly scalable as they require high signaling, data storage and in some cases they involve high complexity. On the other hand, reactive strategies are more scalable, requiring low signaling, data storage and computational capacity at the cost of longer delay. One exception is the users informing their SLS strategy, which involve very short delay.

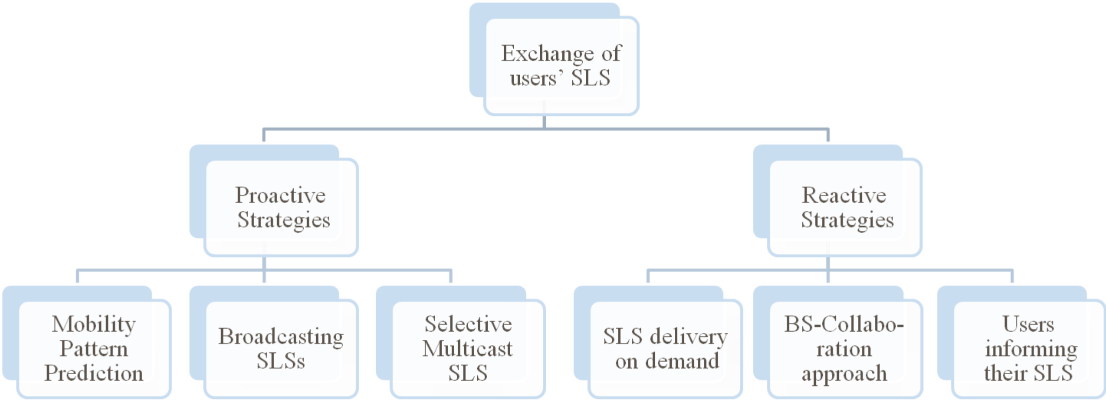

Figure 3.1 Strategies to inform the new BS of MS's SLS

- *Mobility pattern prediction:* The QGS prefigures the next BS to which the MS will perform handoff and delivers the SLS to that BS, thus reducing the handoff negotiation delay [47]. However, this method increments highly the complexity of the system and accordingly impact its scalability.
- *Broadcasting SLSs:* The QGS delivers the SLS of each MS to every BS in the domain. Each time a MS negotiates for a new service, the new SLS is broadcast to all BS in the domain. This method is simple and minimizes the handoff negotiation delay, as all the BSs know in advance the SLS of every MS. Thus upon handoff, the new BS is able to immediately perform traffic conditioning. However, this mechanism presents scalability problems in terms of signaling overheads and data stored at the BSs, as the database in the BSs will be huge if there are many MSs in the domain. In addition, if a MS performs SL renegotiation, the same transaction for updating the database must be performed at all the BSs.
- *Selective Multicasting SLS:* It is an enhancement to broadcasting SLSs. The QGS delivers the SLS of the MS to the current BS and the most likely next BSs. Thus, possible new BSs already know the MS's SLS. This method decreases the signaling overheads as well as the amount of SLSs stored at the BSs. However, both signaling and data stored still remain high, as several BSs receive and store in their tables the SLS of the MS every time it performs handoff.

- *SLS delivery on demand:* The QGS delivers the MS's SLS to the new BS in response to an SLS solicitation message sent by the MS and forwarded by the new BS. This method reduces the signaling overhead and the data stored at the BSs, as the SLS of the MS is delivered to the appropriate BS only on demand. However, such a reduction on the signaling overhead and data stored comes at the price of handoff negotiation delay that is increased by the round trip delay between the BS and the QGS.
- *BS-collaboration approach:* An enhancement to SLS delivery on demand method. When an MS performs handoff, the new BS consults adjacent BSs for the SLS of the MS. By this way, this method reduces the handoff negotiation delay as the round trip delay among two adjacent BSs should be shorter than that among BSs and QGS. This approach is highly scalable as the MS's SLS is delivered only to the new BS.
- *Users informing their own SLS*: in this method when a MS negotiates the service level for the first time, it receives its SLS in an encrypted form. Upon handoff, it sends its own encrypted SLS to the new BS. The BS decrypts the encrypted SLS and performs traffic conditioning.

3.1 Former SLS Delivery approaches

In this section we present the three most relevant approaches for SLS delivery; DSNP, BS-collaboration approach, and ESLS.

- Dynamic Service Negotiation Protocol (DSNP) [48] is a protocol developed by the ITSUMO (Internet Technologies Supporting Universal Mobile Operation) team. DSNP fall into the Selective Multicast SLS strategies, it was developed exclusively for dynamic service negotiation and provides support for user's mobility. When a MS negotiate for the service level required by its traffic, the QGS delivers the resulting SLS not only to the BS in which the MS is currently located, but also to all the BSs that serve the wireless networks adjacent to the MS's current location. This intuitively requires storing the SLS of users at BSs. Consequently, when the MS performs handoff, any possible new BS already has the SLS of that MS. Thus the MS continues to enjoy the negotiated service from the new BS. After that, the new BS informs the QGS that it is serving that MS and the QGS delivers the SLS of that MS to the BSs that serve the wireless networks adjacent to the MS's current location. However, both signaling and data stored still remain high, as several BSs receive and store in their tables the SLS of the MS every time it performs handoff or perform SL renegotiation. Fig. 3.2 shows the handoff procedure for DSNP.

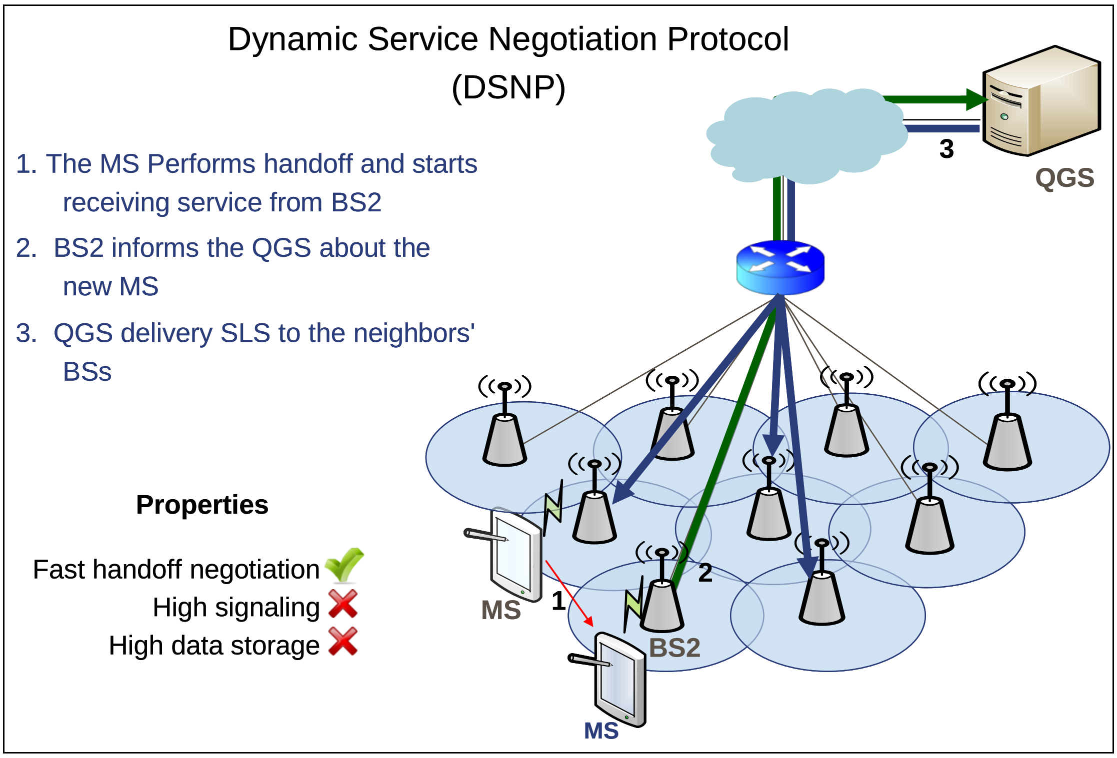

Figure 3.2 DSNP SLS' delivery Strategy

- BS-collaboration approach was introduced in our previous work [49]. When a MS performs handoff to a new BS in the same domain, it sends the IP address of the previous BS to the new BS, via a service negotiation request. In response, the new BS confirms the SLS from the previous BS. If the new BS can guarantee this SLS, it sends a positive service negotiation response to the MS. The MS starts enjoying the service from the new BS immediately. Then, the new BS informs the QGS that it is currently providing service to the MS to update available resources of the new BS and the previous BS. Additionally, the previous BS erases the SLS of the MS from its database. This operation ensures that BSs store information on SLSs of only users they are currently serving. This method reduces the handoff negotiation delay as the round trip delay among two adjacent BSs should be shorter than that among BSs and QGS. This approach is highly scalable as the MS's SLS is delivered only to the new BS. Fig. 3.3 portrays the mayor procedures for this approach.

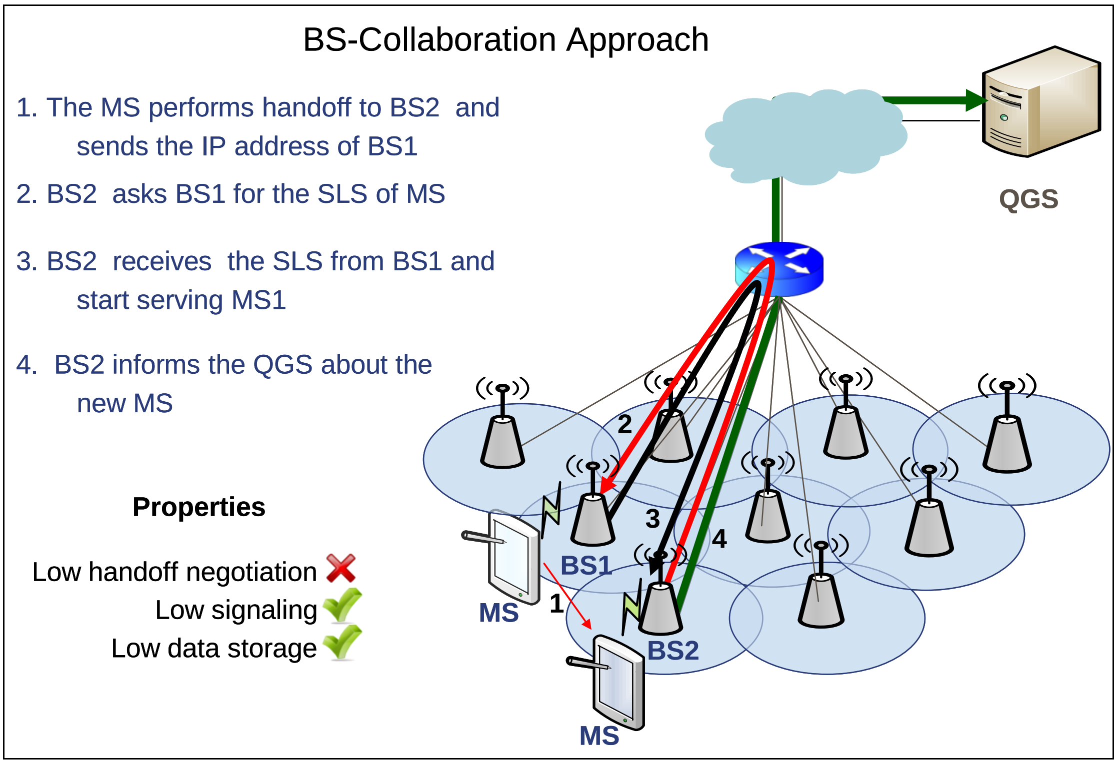

Figure 3.3 BS-Collaboration SLS' delivery Strategy

- Encrypted SLS, as in [50] after a user negotiates its service level with the network, the QGS delivers an SLS token to the user. The token contains all details associated with the service level and traffic specifications. The token is encrypted and cannot be deciphered by MSs. However, any BS into the same domain can decrypt it by using a network specific secret key. When a user performs handoff within the domain, it simply sends the SLS token to the new BS, which decrypts it and performs traffic conditioning. Accordingly, the handoff negotiation delays as well as the signaling overheads are minimized. However, this method presents some security concerns. Indeed, malicious users can obtain the SLS token from genuine users and steal their service levels. Fig. 3.4 illustrates the signaling required in this strategy.

ESLS shows to be the best approach in terms of scalability and short delay, however, its poses some security issues as shows Fig. 3.5 and Fig 3.6, ESLS presents two security gaps, first, a malicious user can intercept the token sent to a legitimate user by the QGS and move to the coverage area of another BS to use this token to steal the service. And second, since the legitimate users keep their token during long time (the entire course of their session), they can decrypt the token by brute force, modify it, and encrypted again to obtain a better service level from the network.

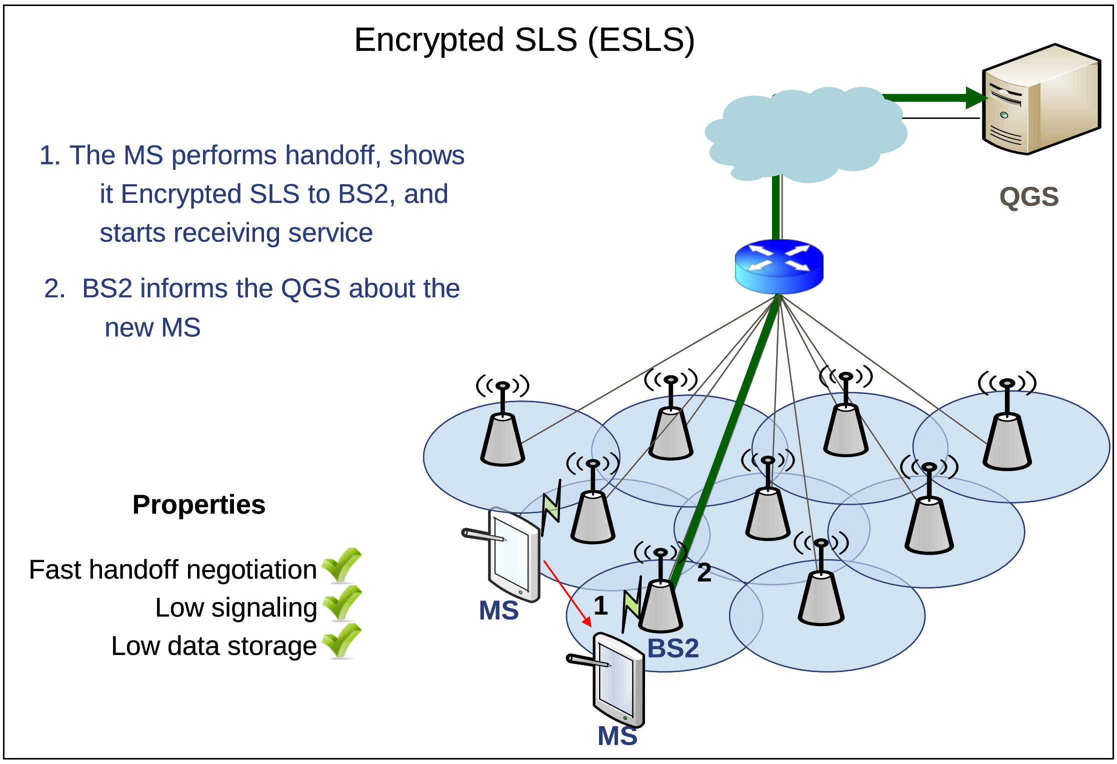

Figure 3.4 Encrypted SLS delivery Strategy

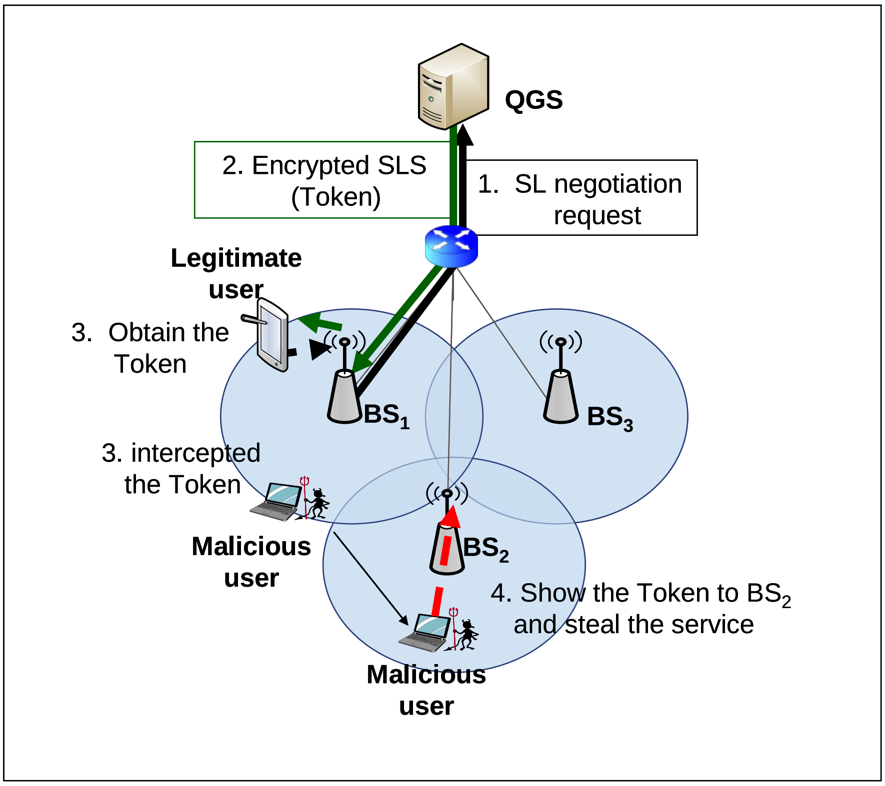

Figure 3.5 Encrypted SLS' vulnerability 1

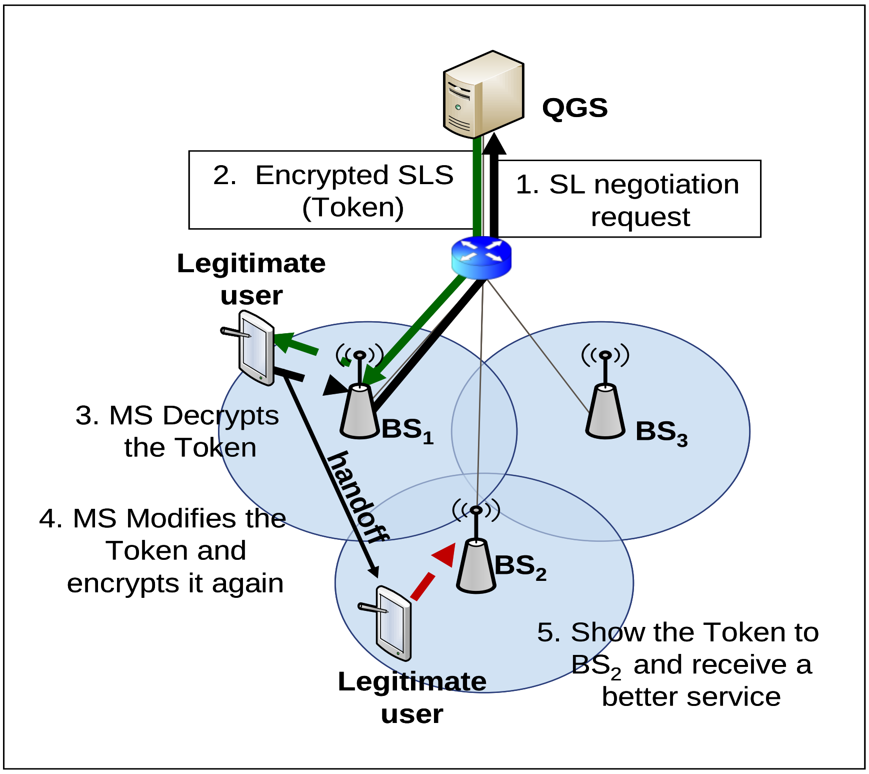

Figure 3.6 Encrypted SLS' vulnerability 2

In the next section we developed an enhanced version of the Encrypted SLS (ESLS) approach to be adopted in the dynamic SL negotiation system to inform the new BS of the SLS of the MS.

3.2 Propose Extended Encrypted SLS (EESLS)

In this section, we develop an enhanced version of ESLS by tackling its security issues. We implement public-key cryptography at the BSs, also known as asymmetric cryptography, in which the key used to encrypt a message differs from the key used to decrypt it. Thus, each BS has a pair of cryptographic keys: a public key and a private key. The private key is kept secret while the public key is announced to the MSs by using the router advertisement message. To keep the complexity at the BSs low, the MSs encrypt the messages related to authentication only at the beginning of the service negotiation process and upon handoff. Additionally, the generation of the public and private keys pair for each BS is delegated to the QGS. Moreover, the QGS also uses asymmetric cryptography with all the BSs within the domain.

The initial QoS negotiation is shown in Fig. 3.7. The BSs include their public keys into the router advertisement messages. The MS generates and includes a password into the SLS negotiation request message and encrypts this message with the public key of the BS. The BS decrypts the message and forwards it to the QGS. Upon acceptance, the QGS encrypts the new SLS of the MS along with the password of the MS, termed as token, and sends it to the BS into the SLS negotiation response message. The QGS distribute its public key among all the BSs into the domain; the BSs use that public key to decrypt the tokens.

The BS forwards the message to the MS and decrypts the token to obtain the SLS and perform traffic conditioning. Thus, the MS gets its token. That token works only when it is sent by the QGS to a BS. Therefore, even if a malicious user steals the token, it cannot do anything with the same, because the MS should include some security information when it attempts to get services from other BSs.

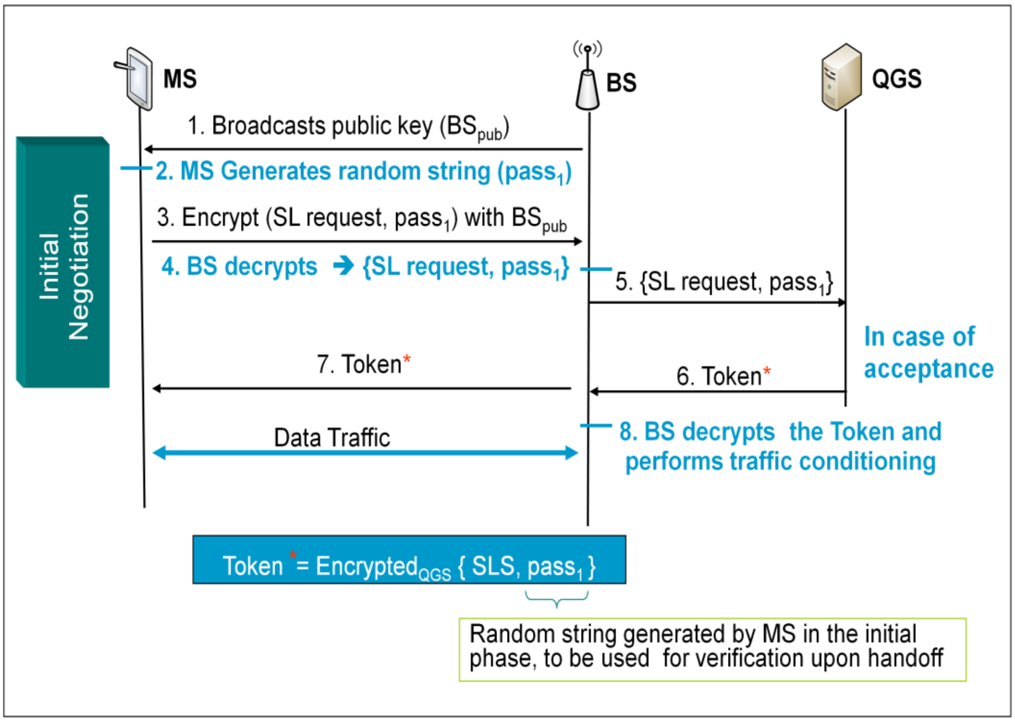

Figure 3.7 Initial SL negotiation for EESLS

As an MS moves and attempts to change its point of attachment to the network within the same domain (intra-domain handoff), it receives the router advertisement message from the new BS, and uses the public key of that new BS to encrypt a handoff negotiation message, which contains the token, the MS' password, and the current time, as shown in Eq. 3.1.

$$HNM = Encrypted_{BS}(Token, \quad Pass_2, \quad time) \tag{3.1}$$

where *HNM* denotes the handoff negotiation message that is encrypted with the public key of the BS. And Token is shows Eq. 3.2.

$$Token = Encrypted_{QGS}(SLS, \quad Pass_1) \tag{3.2}$$

where the *Token* is encrypted with the private key of the QGS, and can be decrypted by any BS in the domain.

Then the MS sends the HNM message to the new BS, which decrypts the message with its private key, decrypts the token with the QGS's public key, and compares the MS's password into the token with the password in the message. If they match, the new BS verifies if the message is a fresh one by verifying the time into the message. If the time is recent enough, the new BS performs traffic conditioning by applying the SLS of the MS. This procedure is depicted in Fig. 3.8.

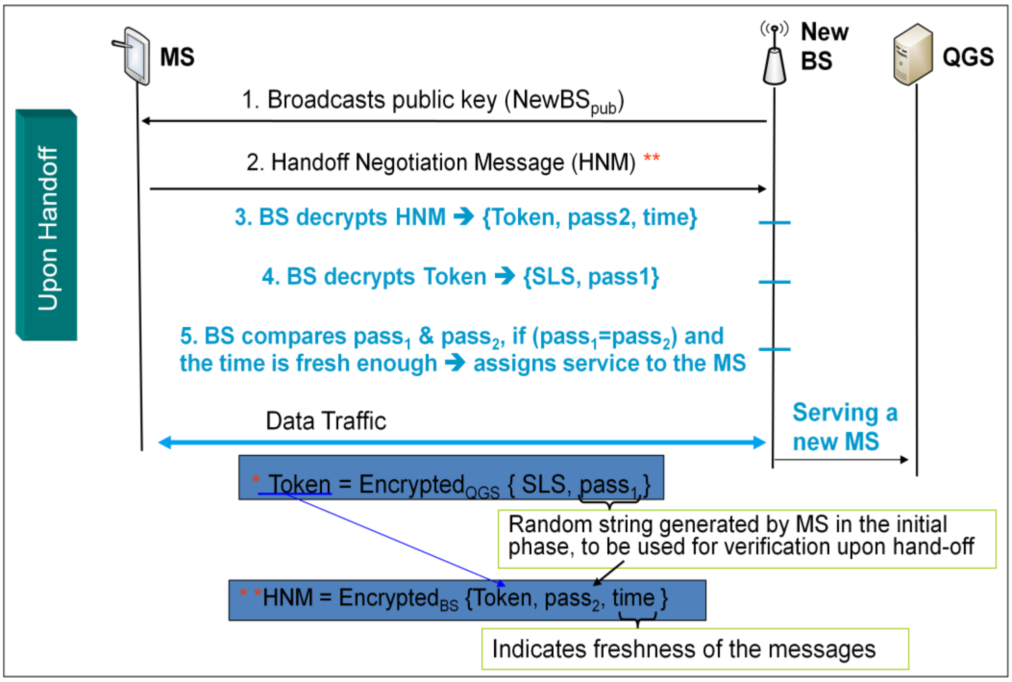

Figure 3.8 Handoff SL negotiation for EESLS

The MS receives the handoff negotiation response message and starts receiving the service. In this fashion, we ensure that a malicious user cannot steal the service level of an MS just by intercepting its token. On the other hand, if a malicious user gets the handoff negotiation message of a MS, it cannot be used at the same BS because only one MS can receive services with a single SLS. The time in the handoff negotiation message is used to ensure that a malicious user cannot use an intercepted handoff negotiation message when the MS performs handoff to another BS. In such a case, the time in the message is not recent. The malicious user cannot use the handoff negotiation message of the legitimate MS in a different BS because the pairs of keys are different for the BSs; therefore, the new BS is unable to decrypt the message.

After that, the new BS informs the QGS that it is currently providing services to the MS in order to update the available resources of the new BS in the QGS database. Moreover, when the previous BS detects that the MS is no longer active in the coverage area, it erases the SLS of the MS from its database, releases the associate resources, and informs the QGS of its new state of available resources. These operations ensure that BSs store information on SLSs of only users that they are currently serving. In case the new BS is unable to guarantee the SLS, it forwards the handoff negotiation messages to the QGS. Then the QGS sends a negative handoff negotiation response to the MS, informing the MS of available service levels that the new BS can offer.

When an MS moves out to a new domain (inter-domain handoff), the MS negotiates a new SLS with the QGS of the new domain, because getting the SLS of the MS from the QGS in the previous domain may be more costly than negotiating a new one.

The main goal of contemporary researches in mobile networks is to provide seamless handoff, which is not always possible; in some cases the available resources in the new BS may not be enough to guarantee the SLS of the MS. In such a case, the QGS asks the MS to downgrade its SLS. Such downgrade of the service level affects the quality perceived by the user. Therefore, the user notifies the corresponding source of the new service level and the sender accordingly adjusts its streaming rate.

3.3 Performance Evaluation

This section presents and discusses the performance of the proposed mobility management strategy for the dynamic SL negotiation System.

We set up a simulation environment using the network simulator (NS-2) [51] to evaluate the applicability of the proposed mobility management mechanism. As mobility management deals with MSs performing handoffs between BSs of the same wireless technology, for this evaluation, we consider that MSs are employing only one wireless interface (the same wireless technology for all the users). The mobility of MSs follows the Reference Point

Group Mobility (RPGM) model. To provide a wide area for the users to move around, we consider the coverage area of five BSs where is located as shown in Fig. 3.9. The number of MSs roaming over the coverage area varies from 10 to 100. The simulation starts when all MSs have already initiated their service levels. The major issue in providing QoS in wireless networks consists in the mobility of users (where seamless and lossless handoffs need to be guaranteed). Therefore, the focus of this evaluation is on the service level negotiation upon intra-domain handoff, as this is the most frequent handoff performed by MSs. The background traffic consists of Constant Bit Rate (CBR) applications running between each pair of BSs.

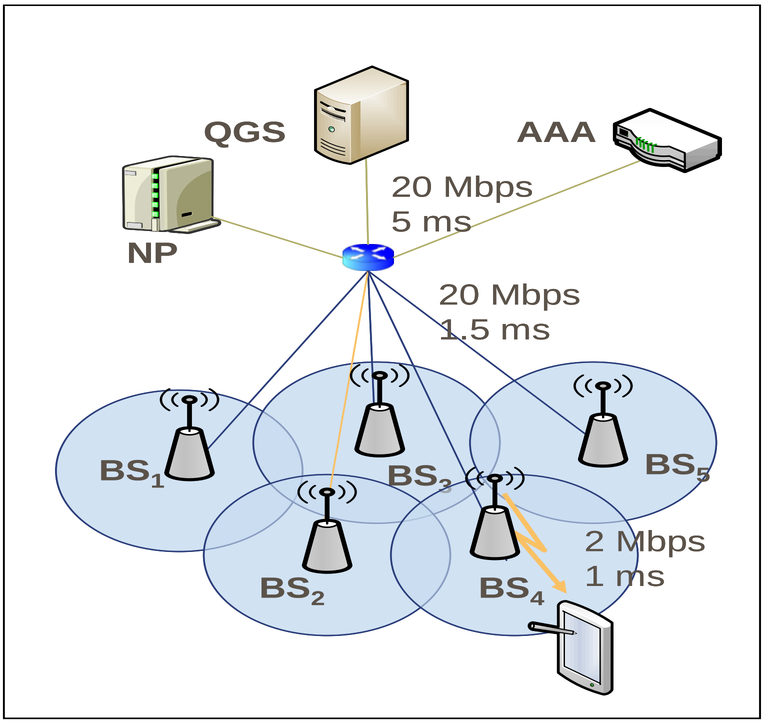

Figure 3.9 Simulation Topology

Fig. 3.10 shows that both the proposed extended encrypted SLS (EESLS) and DSNP exhibit close handoff negotiation delays associated to the round trip delays from the MSs to the BSs. The slight difference among them is attributable to the decryption time of the handoff negotiation message in case of the EESLS method. On the other hand, the BS-collaboration method shows the highest handoff negotiation delay because of the fact that the new BS gets the SLS from the previous BS, which requires communication between the two BSs. This increases the overall negotiation delay, whereas in DSNP the new BS already has the SLS or receives the SLS from the MS in case of the EESLS scheme.

Fig. 3.11 shows that the proposed EESLS scheme has the lowest signaling overhead as the MSs deliver their own SLS to only the next point of attachment. Fig. 3.12 demonstrates that both the BS-collaboration approach and the EESLS method require the storage of a lower number of SLSs at the table of BSs than that of DSNP. The small difference between the BS-collaboration method and the EESLS method is because when an MS performs handoff in the BS-collaboration method, the previous BS is asked to deliver the SLS of that MS, and right after that, the previous BS erases that SLS from its table. On the other hand, in the EESLS method, the previous BS erases the SLS of an MS when it realizes that the MS is not active in its coverage area.

We proposed an enhanced version of encrypted SLS mechanism that addresses its security limitations and makes it robust enough to prevent malicious users from stealing the service levels of legitimate users. EESLS highly increases the scalability of the system by minimizing the signaling overhead and the required data storage. This reduces the size of the tables of BSs and also the time required to search into these tables. EESLS also achieves low handoff negotiation delay, which is essential to provide seamless handoff and to ensure the continuity of the service.

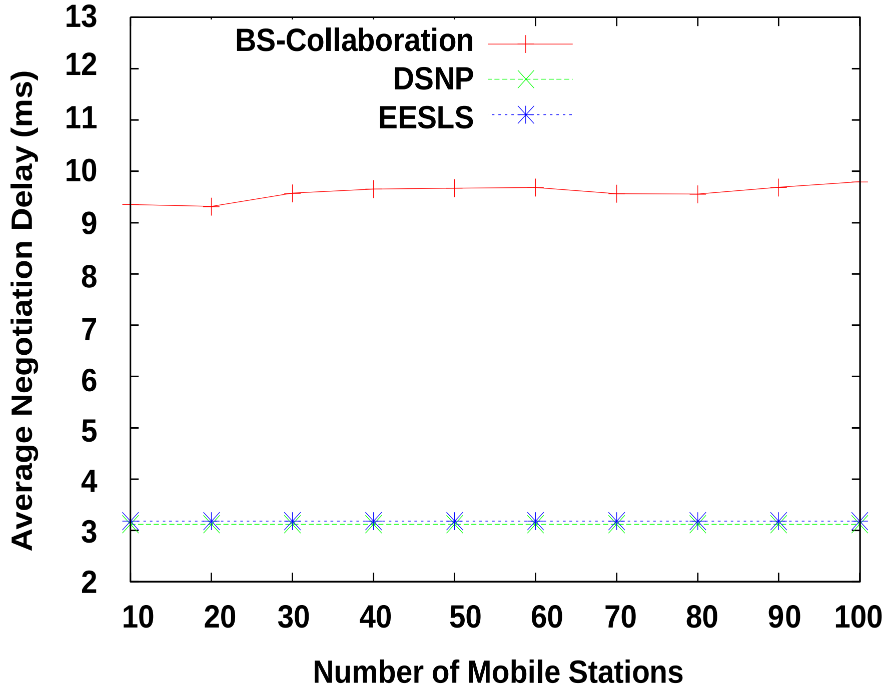

Figure 3.10 Handoff negotiation delay

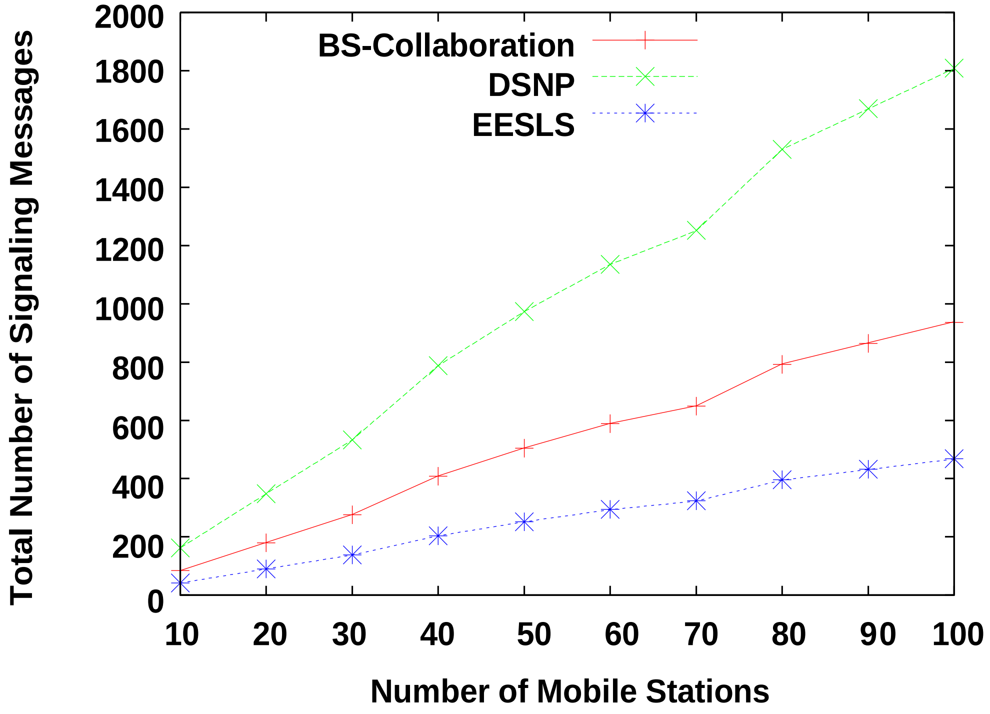

Figure 3.11 Handoff negotiation signaling

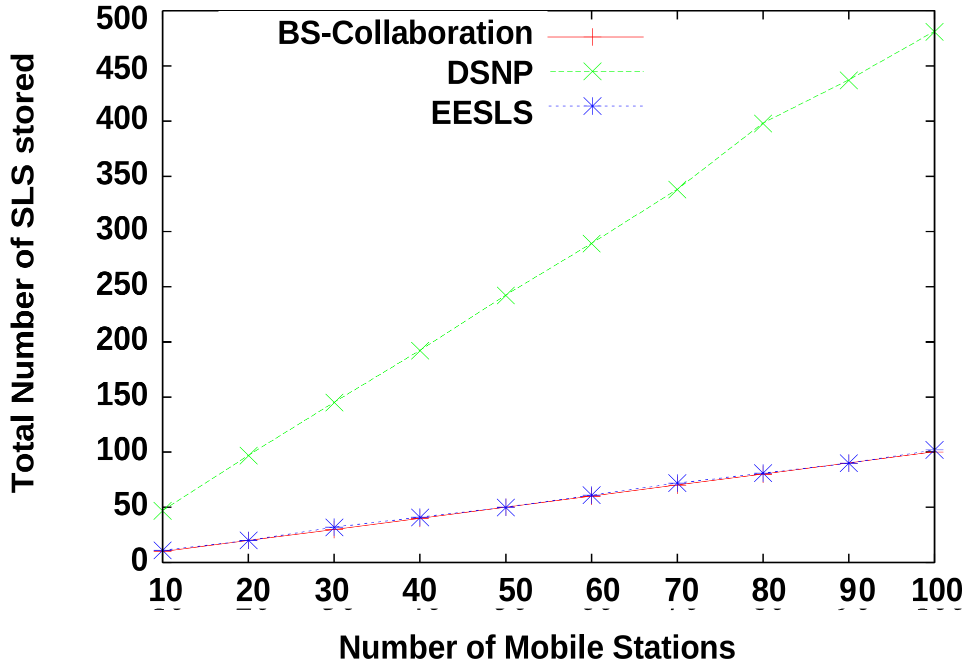

Figure 3.12 Handoff negotiation data storage

3.4 Results Analysis

- DSNP yields the lowest handoff negotiation delay, as all possible new BSs have already knowledge on the SLSs. Thus, the BS can perform traffic conditioning almost immediately when the MS performs handoff. However, DSNP presents problems related to scalability in terms of the data storage and signaling overhead. When a MS performs handoff to a new BS, all neighboring BSs receive the SLS of the MS from the QGS, even if some of these BSs will never serve the MS. Thus, BSs maintain huge state tables for storing SLS of MSs that may never visit their coverage areas. Another issue of DSNP is its perpetual storages of data, as there is no mechanism to inform the BSs when to erase the SLS of MSs. In addition, the QGS needs to know about the network topology to identify the neighbors of each BS.
- BS-Collaboration approach addresses the scalability problem of DSNP, the signaling overhead is much smaller than that of DSNP as SLSs are delivered to only the required BS. However this mechanism presents the longest handoff negotiation delay.
- The propose EESLS shows a very short handoff negotiation delay, similar to that of DSNP. The small difference is due to SLS decryption time. EESLS has the lowest signaling overhead as the signal required is just from the MS to the new BS and from the BS to the QGS. The data storage is also very low, as EESLS introduces a mechanism to erase the SLSs of departing MSs. In this way, each BS maintains the SLS of only MSs that currently exist in its coverage area.

3.5 Summary

In this chapter we proposed a mechanism to manage the mobility of users in the dynamic service level negotiation system. The proposed mechanism combines the benefits of both proactive strategies like short handoff negotiation delay, and reactive strategies like high scalability. The proposed method is worthy of evaluation as an indicator of the direction of the new mobility management technology.

The propose EESLS is an enhanced version of Encrypted SLS mechanism that addresses its security limitations and makes it robust enough to prevent malicious users from stealing the service levels of legitimate users. EESLS highly increases the scalability of the system by minimizing the signaling overhead and the required data storage. This reduces the size of the tables of BSs and also the time required to search into these tables. EESLS also achieves low handoff negotiation delay, which is essential to provide seamless handoff and to ensure the continuity of the service level perceived by users while they are on the move performing handoff between different base stations.

Chapter 4

Bandwidth Aggregation

This chapter presents a new functionality for the dynamic service level negotiation system to allow users to obtain higher transmission rate by aggregating the available bandwidth of multiple wireless networks.

In next-generation wireless networks, users are expected to be able to receive the same services as they do over wired networks, including high bandwidth demanding services like real-time multimedia applications, which are highly sensitive to delay, jitter, and bandwidth restrictions. These characteristics become more significant in wireless mobile networks as the bottleneck for most wireless communications is the last hop, from the Base Station (BS) to the Mobile Station (MS). Furthermore, such services should be provided over a variety of wireless technologies that exhibit different data rates. Thus, the next-generation wireless networks are expected to provide constant high bandwidth for real-time multimedia applications to be successful. Two research topics, as follows, emerged recently to deal with such requirements:

- **BAG Bandwidth aggregation:** MSs equipped with multiple interfaces using different wireless technologies are able to obtain simultaneous connections through these interfaces when the coverage areas of these technologies partially overlap as in Fig. 4.1. Such a capacity allows mobile terminals to increase the streaming bandwidth by distributing the load over multiple network paths.
- **Dynamic QoS negotiation:** A MS in a dynamic QoS negotiation system is able to negotiate with the network the desired service level for its traffic, which should be guaranteed by the network during the entire course of the session.

By combining the benefits of these two research areas, a user who wants to execute a real-time video application should negotiate the amount of bandwidth required by this application. If the bandwidth of a single interface is not enough to meet the required one, the user may consider two or more interfaces to ensure the quality of the video application. However, the transmission of packets of a single application via multiple paths, with varying characteristics in terms of capacity and propagation delay, makes those packets arrive to the final destination in an out-of-order manner, which results in packet reordering and increase the delay. Accordingly, the packet loss rates increase due to packets' timer expiration.

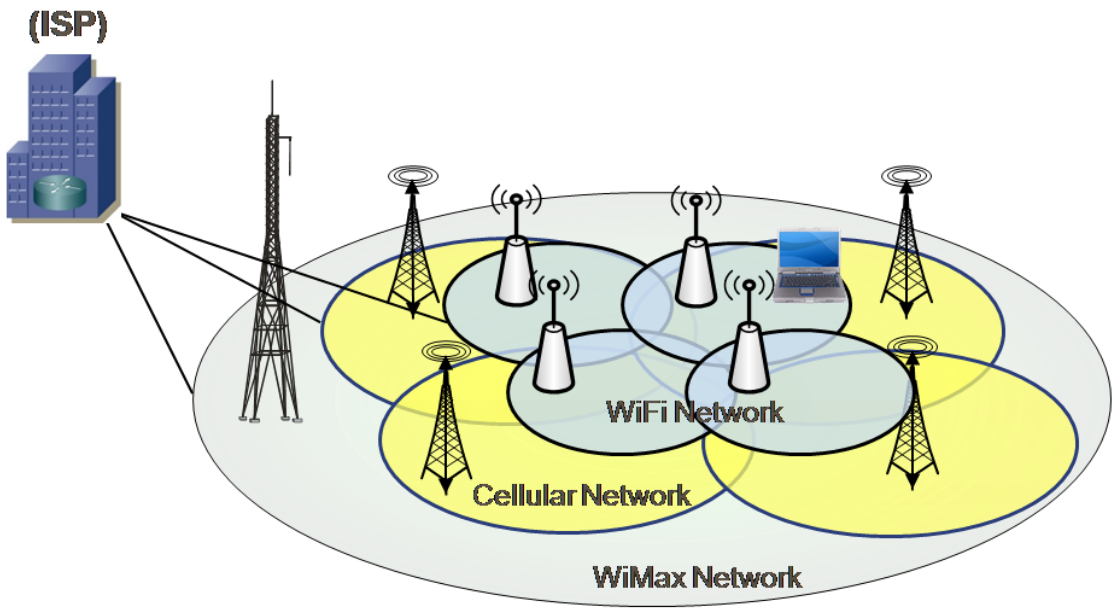

Figure 4.1 Overlap coverage areas for different wireless technologies

4.1 Related works

To enable MSs to simultaneously access two or more different BSs, MSs are equipped with multiple wireless interfaces. To allow a MS to simultaneously register multiple Care-of-Addresses (CoAs), Mobile IP (MIP) simultaneous binding option [52] [53] is used. On the other hand, to keep senders always informed of these CoA registrations directly from the MS, the route optimization option [54] is used.

The use of multiple interfaces in wireless devices has been studied for different purposes. For instance, Stream Control Transmission Protocol (SCTP) [55] uses multiple interfaces to define different paths for communication between two end-terminals. To ensure high reliability, SCTP uses one of the available paths as the primary path for data transmission. The other paths are used for the retransmission of the lost packets or as a backup in case of the failure of the primary path. For this purpose, some interfaces should belong to the same technology to guarantee a certain level of fault-tolerance on any interface. Some variations of the original SCTP have been presented in [56]-[59], which are able to distinguish among losses due to congestion and radio channel failures to better select the path for data transmission. The work presented in [60] provides mechanism to monitor the one-way delay variation throughout the available paths. A variation of SCTP was developed to provide bandwidth aggregation, particularly for the provision of real-time applications to wireless mobile users. Multiple interfaces are also used for bandwidth aggregation, particularly for the provision of bandwidth-intensive real-time applications to MSs. Load-Sharing SCTP (LS-SCTP) [61] introduces a new functionality to SCTP by involving all the active transmission paths in data communication and aggregating their bandwidths to share the data load between two end-points. The bandwidth of MSs with multiple interfaces is aggregated at the transport layer in pTCP (parallel TCP) [62], which is a wrapper that interacts with a modified TCP called TCP-virtual (TCP-v). A TCP-v connection is established for each interface. pTCP sends its whole buffer contents through each TCP-v pipe.

The stripping is performed by pTCP and is based on congestion window-size of each TCP-v connection. When congestion occurs on a certain pipe, pTCP performs data reallocation via another pipe with a larger congestion window. pTCP achieves aggregate bandwidth for stripped connections even in the presence of disparities among the different paths. It also accommodates fluctuations in path characteristics. Multimedia Multiplexing Transport Protocol (MMTP) [63] is a protocol designed for transferring multimedia data on mobile systems. It makes simultaneous use of every communication channel available to send data. Data transmission in MMTP is performed by two mechanisms. The first is a set of rate control protocols associated with each outgoing channel. The second is a scheduling algorithm that places incoming packets on the appropriate channel.

MMTP is a link-layer aware protocol. Indeed, it refers to the link layer to estimate the available bandwidth, relays this information to the application layer for rate adaptation, and accordingly performs congestion control.

A network layer solution based on tunneling was proposed in [64] and [65] and performance of TCP has been evaluated. Though similar in spirit to our architecture, this work does not look into real-time application support or address in depth the architecture components that enable diverse services. The Reliable Multiplexing Transport Protocol (RMTP) [66] is a reliable rate-based transport protocol that multiplexes application data onto different channels.

In a bandwidth aggregation scenario, packets of the same flow are transmitted over multiple interfaces. While this operation has many advantages, it makes packets of same application experience different latencies and delay jitter, resulting in out-of-order delivery to the final destination.

For connectionless-oriented protocols such as User Datagram Protocol (UDP), addition of buffering capabilities to end terminals can ensure coherent reception and recover the original timing relationships between the transmitted data. However, when the interfaces exhibit significantly different channel conditions, a significant jitter can be experienced and the use of a small buffer will not be efficient enough. Some solutions for this issue have been proposed in [67]-[82].

For applications based on connection-oriented protocols (e.g., TCP), such disorder in packet reception results in the transmission of unnecessary duplicate acknowledgments (DupAcks). Indeed, current implementations of TCP work on the assumption that out-of-order packets indicate network congestion. TCP senders mistakenly halve their congestion windows when packets are reordered. Some modifications to the TCP to make the communication more robust have been proposed in [83]-[90].

To cope with packet reordering in multipath environments, various scheduling strategies have been proposed. Most of them are based on round robin scheme that is suitable for environments with paths homogeneous in terms of bandwidth and delay. For heterogeneous paths, Weighted Fair Queuing (WFQ), Weighted Round Robin (WRR) and Surplus Round Robin (SRR) [91] are notable scheduling schemes.

The concept of QoS negotiation in wireless networks has simplified the scheduling operation as the access network guarantees certain amount of bandwidth to MSs during their connection. Thus knowledge on the bandwidth of each path is available. The Earliest Delivery Path First (EDPF) scheme presented in [67] exploits such characteristic and focuses its operations on finding out the best path for the delivery of each packet.

As previously mentioned the use of multiple interfaces is highly beneficial in wireless networks as it increases the system reliability and enhances the quality of real-time applications. However, it may cause excessive energy drainage and may consequently reduce the lifetime of the scarce battery of mobile users. To cope with this issue, Multinet [92] proposed a software-based approach that virtualizes the simultaneous connections (to multiple BSs) into a single virtual wireless card which is constantly switched across the multiple networks.

4.2 Bandwidth Aggregation-Aware Dynamic SL Negotiation

In Chapter 2, we presented a dynamic service level negotiation system that allows users to negotiate dynamically the specific amount of bandwidth that their traffic need. Such a negotiation system implements a time-slot approach for bandwidth allocation at the BSs to guarantee the negotiated service level to the MSs.

To exploit the advantages of Internet services providers offering services via several wireless technologies and mobile computers equipped with multiples wireless interfaces, we included bandwidth aggregation in that service level negotiation system to allow MSs to obtain higher data rate.

Fig. 4.2 deploys the architecture for bandwidth aggregation-aware SL negotiation System. The component are the same than that for the service level negotiation explained in chapter 2, however, in this case the network operator deals with three different wireless networks.

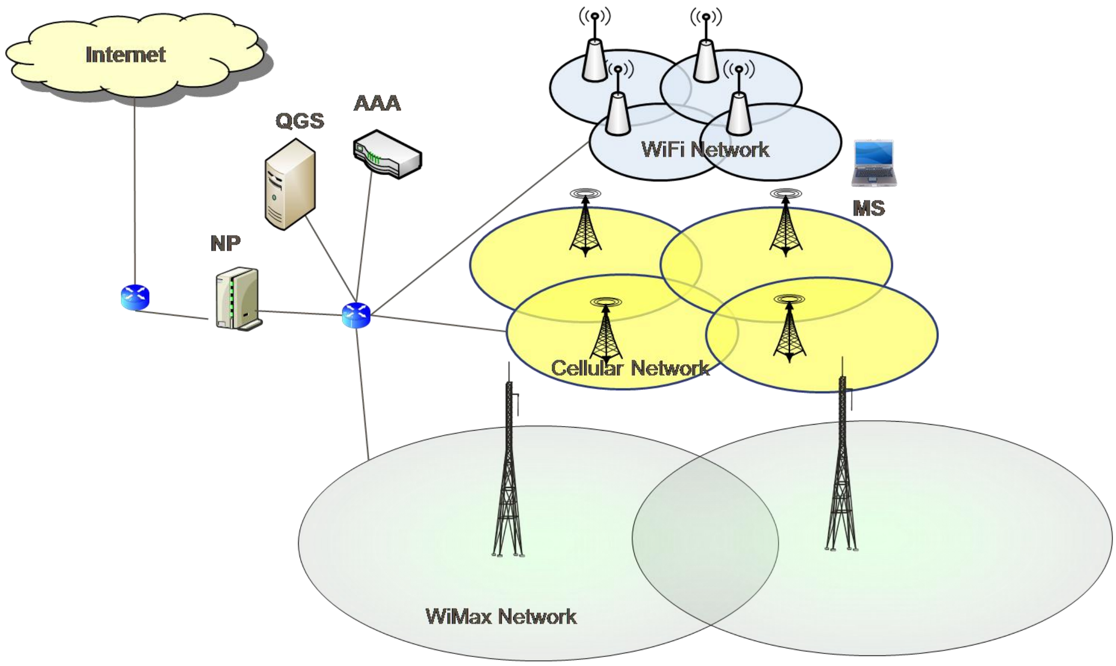

Figure 4.2 Architecture for the Bandwidth Aggregation-aware SL Negotiation System.

Some important characteristic of the SL negotiation system are as follows.

- Even when a MS is able to negotiate SLSs through all its interfaces, each SLS is associated to one specific interface. Thus, MS should perform an initial negotiation through each interface it attempts to use to connect to the network.

- A MS handoff refers to the event where MS changes its point of attachment in one specific wireless network technology. It is unlikely that a MS performs handoff through more than one interface at the same time.
- MSs are able to perform handoffs between BSs of the same wireless technologies (horizontal handoff) only.

Fig. 4.3 shows the general procedure that MSs follow to negotiate the required bandwidth for their applications. The MSs attempt to get the whole required bandwidth through any of the available interfaces, starting from the interface with the strongest signal and following a descending order of the signal strength. Recall that every time a MS's negotiation request is rejected by QGS, QGS informs the available bandwidth to the user. In case of the user could not obtain the requested bandwidth through any interface, it evaluates whether the sum of the available bandwidth of each interface satisfies the required bandwidth. In affirmative case, the user negotiates the available bandwidth through each interface until reaching the requested bandwidth. Otherwise, the user should consider downgrading the requested bandwidth or waiting for better network conditions.

After each successful SL renegotiation, NP is informed of the new path to transmit the packet for the MS. In case of renegotiation due to handoff, NP simply redirects the traffic from the previous BS to the new BS.

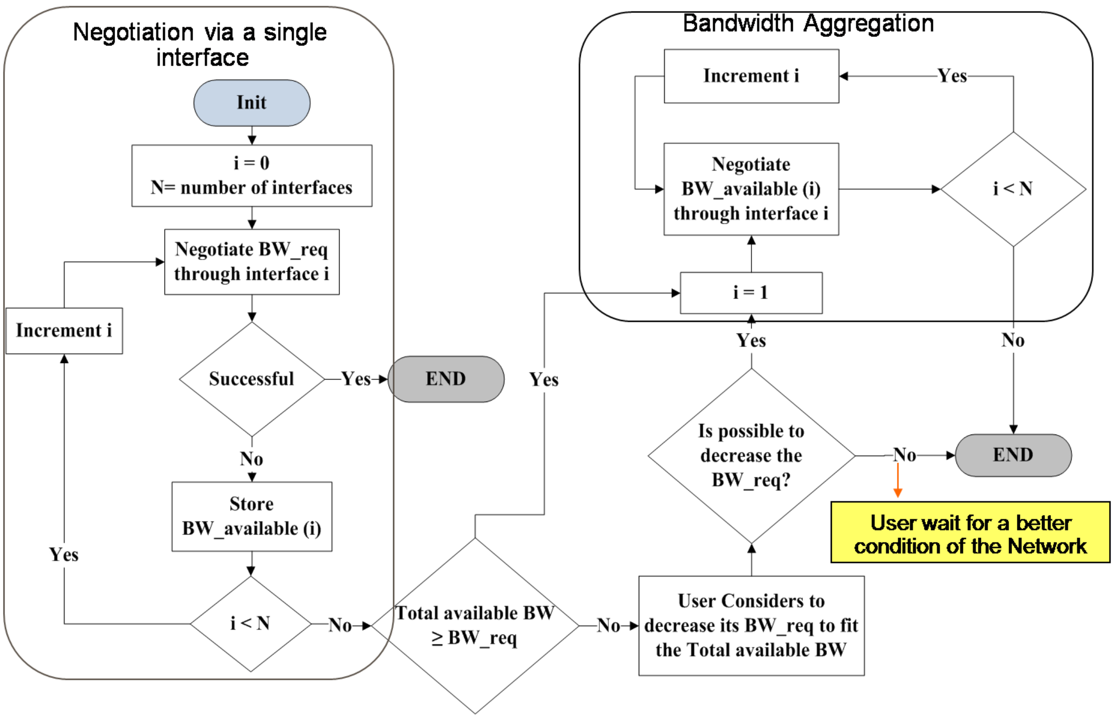

Figure 4.3 Initial SL negotiation process.

4.3 Bandwidth Aggregation Issues

In the context of Service Level Agreement (SLA) management, BAG may introduce new issues to operators of wireless network systems. Indeed, since the possibility of having the coverage areas of two or more BSs overlapped and the use of multiple wireless Interfaces (IFs) have not been considered in the design of current service level negotiation mechanisms, a MS may receive or send data at bandwidths higher than what it deserves, in other words higher than what it subscribed for. This will result in an unfair service as it will damage the service quality perceived by other unfortunate subscribers. Another issue of BAG consists in the delivery of data via multiple paths, which lead packet reordering. In case of video streaming, packet reordering may result in extra delay in playback at the receiver side.

The two issues mentioned above can be resolved by the addition of an effective bandwidth aggregation control mechanism to the SL negotiation system and the development of an efficient multipath scheduling strategy, respectively. The former should ensure that each MS does not receive or send data at a bandwidth higher than what is indicated in its SLA. The latter will be discussed next in this chapter.

4.4 Bandwidth Aggregation control mechanism

Mobile stations able to connect to the network via multiple interfaces simultaneously introduce a new issue related to SLA management. When an MS negotiates the service level for its traffic, the QGS confirms from the AAA server that the MS is allowed to receive the requested service level. The AAA server verifies if the requested service level exceeds the agreed SLA of the MS or not. Since, in bandwidth aggregation scenarios, MSs are able to negotiate SLSs through several interfaces, such a verification method is not suitable, as the MS can negotiate SLSs through all its interfaces, and the network separately verifies each SLS as in Eq. 4.1

$$BW_i \leq BW_{SLA}; \qquad 1 \leq i \leq n \tag{4.1}$$

where BW_i, BW_{LSA}, and n denote the bandwidth negotiated through the interface i, the bandwidth specified in the SLA of the MS, and the number of interfaces of the MS, respectively. Thus, a MS may obtain up to n times the bandwidth indicated in its SLA as shows Eq. 4.2.

$$\sum_{i=1}^{n} BW_i \leq n \times BW_{SLA} \tag{4.2}$$

On the other hand, some other MSs may get their SLS requests rejected due to the unfair service level assignments. To guarantee an efficient and fair use of the network resources among all competing MSs, in a bandwidth aggregation system, the network operator should consider using some or all available interfaces to ensure the service quality in case a single SLS (provided by a single interface) does not meet the pre-agreed SLA. In the same manner, if the aggregate SLSs provided by multiple interfaces exceed the pre-agreed SLA, the network operator should hinder the user from using some of the interfaces to ensure a fair utilization of network resources among all active MSs. Thus, the network should ensure that the total bandwidth assigned to an MS, via its available interfaces, does not exceed that of the agreed SLA, as shown in Eq. 4.3.

$$\sum_{i=1}^{n} BW_i \leq BW_{SLA} \tag{4.3}$$

In the envisioned architecture, the AAA server performs the bandwidth aggregation control mechanism. Indeed, the AAA server keeps track of the SLSs negotiated by each MS.

4.5 Multipath Scheduling Strategy

The successful transmission of data belonging to a single application via multiple paths depends on the appropriate scheduling strategy [93][94]. Various scheduling strategies have been proposed; the Round Robin scheduling mechanism was first proposed. It is suitable for environments with paths homogeneous in terms of bandwidth and delay. For heterogeneous paths, Weighted Fair Queuing (WFQ), Weighted Round Robin (WRR), Weighted Interleave Round Robin (WIRR), Surplus Round Robin (SRR), and Earliest Delivery Path First (EDPF).

EDPF [67] is the most notable scheduling algorithm that bases its scheduling on a prior knowledge of the available bandwidth at each interface. The key idea behind EDPF algorithm lies on the estimation of the delivery time of the next packet through each path. Using this estimation, EDPF transmits the packets via the path with the earliest delivery time. However, this mechanism presents some issues to be adopted by the dynamic service level negotiation system. Indeed, EDPF uses the estimation of the end-to-end delay from NP to the MSs, by keeping track of the queuing delay at any intermediate node including the BSs as shows Fig. 4.4. Thus, EDPF assumes a common queue at the BS for MSs' traffic, and also assumes that BS can transmit packets of any MS at any time.

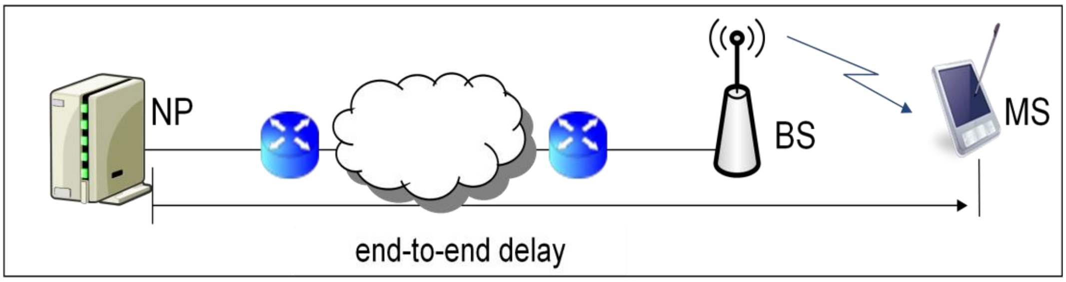

Figure 4.4 The end-to-end delay from NP to MS

As previously mentioned, our SL negotiation system implements a time-slot division approach for bandwidth allocation at the BSs. Thus, each MS is allocated a specific period of time to access the wireless channel. At any given time-slot, only one MS is allowed to transmit/receive data through a particular BS. The time-slot size varies from an MS to another, because the length of the time-slot depends on the amount of bandwidth negotiated by MSs. Thus, BSs transmit packet of each MS in a specific period of time (the time-slot exclusively assigned to that MS). Additionally, the BSs implement the individual queue, where there is a different queue for each MS that negotiate for SL better than best-effort as shows Fig. 4.5. Thus, the queuing delay of each queue is independent from other queues.

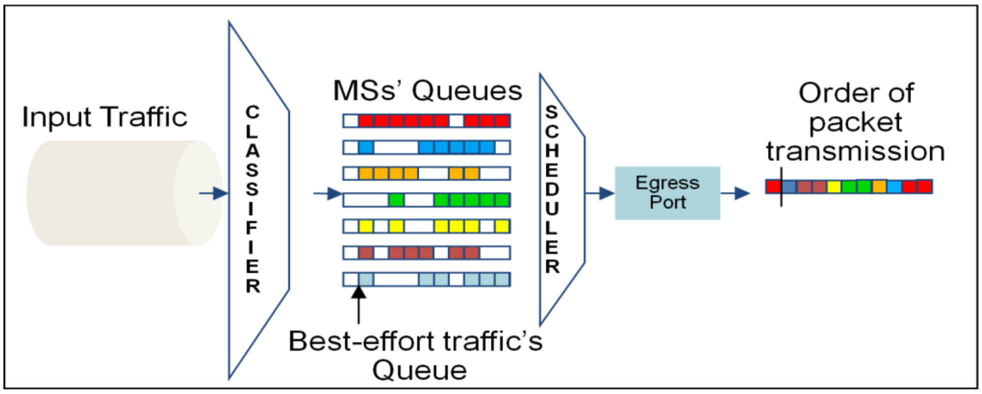

Figure 4.5 Individual queue implemented at the BSs

As shown in Fig. 4.6, in the dynamic SL negotiation system, if a packet of a MS, let say MS_4, arrives to the BS after the time-slot assigned to MS_4, it will remain in the BS' queue until the next round to server MS_4. The waiting time may be up to Δ ms, where Δ represents the time interval in which all queues will be served. Therefore, a multipath scheduling algorithm that takes into consideration the resource allocation scheme implemented in our SL negotiation system is required.

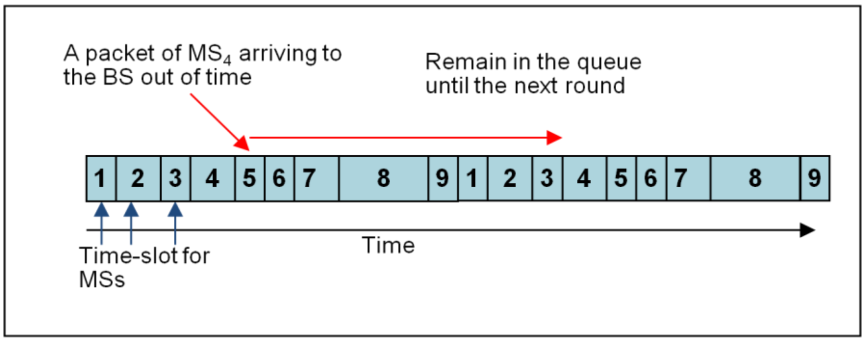

Figure 4.6 Time-slot division strategy

We developed an enhanced version of EDPF called Time-Slotted Earliest Delivery Path First (TS-EDPF). TS-EDPF uses the time-slot assigned to the MS through each available path for an accurate computation of the delivery time of the next packet.

After each successful SL negotiation or SL renegotiation, the BS assigns a specific time-slot to the MS, after that the BS should inform the NP of the specific beginning and ending times of the time-slot assigned to the MS. Using these two parameters, NP can make an accurate estimation of the delivery time of the next packet for the MS through each available path, as shows Fig. 4.7.

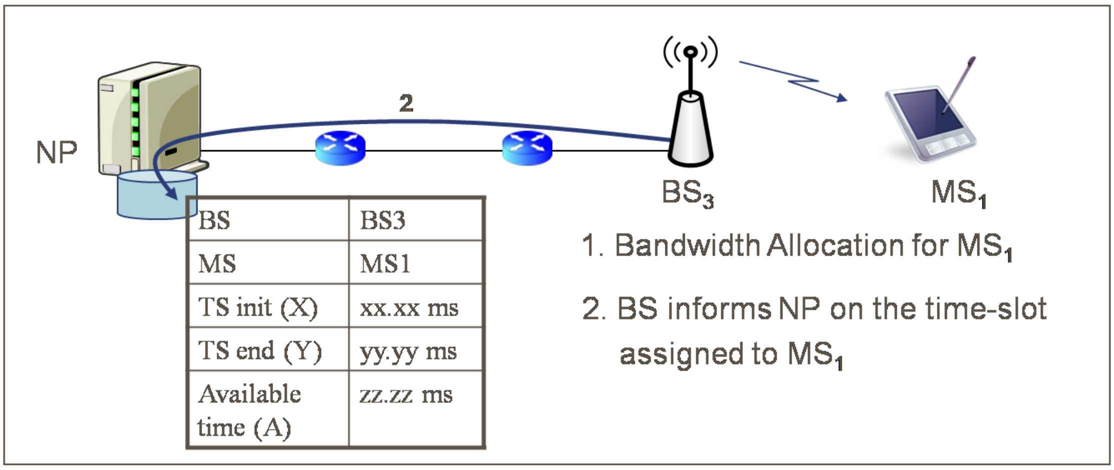

Figure 4.7 Table of available paths for MS

To estimate the delivery time of a packet via a specific path, TS-EDPF computes the time at which the packet arrives at the BS by computing the time at which the transmission can begin at the BS on the path. Then, it adjusts this time so it is within the time-slot assigned to the MS. By adding the transmission delay, we obtain the delivery time of the packet which should be within the time-slot of the MS.

The time at which the transmission can begin at the BS is denoted as

$$S_i^l = MAX(a_i + D_l, A_l) \tag{4.4}$$

where a_i and D_l denote the time at which packet i arrives at the NP and the delay from the NP to the BS along path l, respectively. A_l denotes the time instant when path l will be available for the next transmission.

To adjust S^l_i to be within the time-slot assigned to the MS, let $[X_l, Y_l]$ be the time-slot period for the MS through path l and $X_i^{'}$ be the starting time of the subsequent time-slot. Furthermore, let $\Gamma(S^l_i,l)$ be the function that returns the next valid time at which the transmission can commence at the BS on path l based on the time-slot $[X_l, Y_l]$

$$\Gamma(S_i^l, l) = \begin{cases} S_i^l, & if \quad S_i^l \in [X_l, Y_l] \\ X_l^{'}, & otherwise \end{cases} \tag{4.5}$$

To compute the transmission delay for packet i via link l, denoted by T^l_i, let Li be the size of packet i and let β_l denote the bandwidth of the wireless link on path l. It should be reminded that in a time-slot division system, each MS uses the total bandwidth of the link during a short period of time:

$$T_i^l = \frac{L_i}{\beta_l} \tag{4.6}$$

Then, the algorithm computes the time at which the transmission of packet i can be completed at the BS on path l, denoted by E_i^l:

$$E_i^l = \Gamma(MAX[a_i + D_l, A_l], l) + T_i^l \tag{4.7}$$

Finally, it should be ensured that the transmission of packet i is completed within the time-slot assigned to the MS. Let $\Theta(E_i^l,l)$ be the function that returns the next valid time at which the transmission of packet i can be completed at the BS on path l based on the time-slot $[X_l, Y_l]$:

$$\Theta(E_i^l, l) = \begin{cases} E_i^l, & if \quad E_i^l \in [X_l, Y_l] \\ X_l^{'} + T_i^l, & otherwise \end{cases} \tag{4.8}$$

The delivery time of packet *i*, through path *l*, can be then computed as follows:

$$d_i^l = \Theta(\Gamma(MAX(a_i + D_l, A_l), l) + T_i\ l, l) \tag{4.9}$$

TS-EDPF estimates the delivery time of a packet through each available path and then schedules the packet via the path with the earliest delivery time.

4.6. Efficient Bandwidth Utilization

Supporting VBR video along with voice and data over bandwidth-constraint networks continues to be formidable problem. The difficulty arises because VBR video is unpredictably bursty and because it requires performance guarantees from the network. While resource reservation schemes work best for CBR traffic, there is no consensus on which strategy should be used for VBR traffic. On one hand, since real-time VBR traffic is delay sensitive, a resource reservation scheme seems to be the right choice, on the other hand, because VBR video is bursty, if resources are reserved according to peak rates, the network may be under-utilized if the peak-to-average rate ratios are high. These two opposing characteristics have resulted in a common belief that it is unlikely that performance guarantees can be provided to such bursty sources with very high network utilization. The main design objective of emerging broadband networks is to provide high speed transmission of a wide range of quality of services.

Video is becoming the major component of broadband network traffics and, therefore, an efficient video traffic transmission mechanism is important to network operators. The IP-based Internet was not designed to support QoS guarantees; it offers “best effort” services, in which the network allocates bandwidths among all of the instantaneous users as best as it can, and attempts to service all of them without any explicit commitment on the rate and other service qualities. Real time applications especially for MPEG videos with bursty characteristics often do not work well across the best effort service because of variable queuing delays and congestion losses. In addition, it may not be possible to retransmit information for a real time service, and so any loss in the network results in lost information at the decoder, rather than just increased delays in the case of file transfer.

Much work has been focused on provisioning QoS support to the Internet service model such as Integrated Service and Differentiated Service. Many mechanisms, architecture, and algorithms are proposed to transport video over IP networks based on these two models [96]-[98]. While asynchronous transfer mode ATM networks are more suited to real time and guaranteed QoS communications, the bursty nature of VBR video, combined with the diversity of its QoS requirements, make it difficult to transport video traffic in a cost effective manner. The ratio between the peak and average bit rate may be as large as 10 for some video sequences. If the bandwidth is allocated according to the peak rate of the such videos, no packet loss occurs, but a substantial amount of bandwidth is wasted during most of transmission; on the other hand, if the bandwidth is allocated based on the mean rate, the video service will suffer from unacceptable losses and delays, especially those with hard real time constraints. It is very difficult to meet the QoS parameters such for such kind of traffic while keeping high network utilization.

Dynamic bandwidth allocation approach is an alternative for improving network utilization that allows users to dynamically reserve or adjust network resources. When there is not enough reserved resource for the user to transmit its traffic, a renegotiation is initiated to ask for more. If the reserved resource is more than enough, some bandwidth can be released. In such a way, network utilization can be improved significantly.

Emergence of multimedia communications has inspired a number of dynamic bandwidth allocation schemes. The works [99]-[101] address the problem of transporting pre-recorded videos for video on demand (VoD). These papers consider transmission of stored videos from a server to the client across a network, and explore how the client buffer space can be used most efficiently toward reducing the variability of the transmitted traffic. A number of cells should be built up in the viewer’s buffer before the commencement of playback; the buildup, cell transmission rate, and set-top memory size must be chosen so that there is no starvation or overflow.

For real-time video transmission, one major class of dynamic resource management algorithms is based on parameter measurements. Several measurements based dynamic bandwidth allocation algorithms, which initiate their renegotiation processes based on the actual measurement of Cell Lost Rate (CLR) or user parameters (Ups), have been proposed [102]-[104]. In [102], the CLR is calculated up to the current period, and the service rate for the next period is adjusted based on the current CLR. Owing to the difficulty in assessing the CLR on line and the indirect relationship of the current CLR and UPs with future bandwidth requirements, these approaches are not effective enough to enhance QoS and improve network utilization. In [103], the UPs, such as the peak rate and sustained rate, are adjusted for every GOP (group of picture). UPs could be inherently inaccurate because they are calculated from previous GOPs. To reduce the buffer size, the source quantization step is adjusted on line. The major drawback of this algorithm is that the user parameters (peak rate, sustained rate, and burst length) need to be renegotiated for each GOP, which is a big burden to network management.

Another major class of algorithms is based on prediction. User parameters or bit rates are predicted on-line based on the available information [104],[105], and resources are allocated based on the predicted results. In [106], a traffic model called the Deterministic Bounding Interval (D-BIND) is used. The allocation algorithm stores the currently reserved D-BIND parameters and calculates the D-BIND parameters for the last M frames. A renegotiation takes place when a difference exists between the reserved and measured D-BIND parameters. The calculation of the D-BIND parameters may be problematic since it is done for each frame. So the estimation of the bounding rate over the interval is computationally expensive. [107] approached the problem in the frequency domain. They proposed a method to dynamically allocate the bandwidth based on the predicted low frequency band of the video traffic. The low frequency band represents the slow variation of the VBR rate. It is important to predict these parameters accurately so that network resources can be used efficiently.

In [108], a fast convergent algorithm was proposed. This algorithm not only incurs small prediction errors but, more importantly, achieves fast convergence. All these bandwidth allocation algorithms measure or predict parameters on line. The source information, or a priori information, is not exploited. For videos, the source information is available to the network, and can be exploited to ease network management. Based on the belief that the source information should be used to manage network resources, a new approach, in which a renegotiation process is initiated only when a scene change occurs, is proposed in [109][110], it is applicable to pre-recorded videos. It is well known that the bit rate changes dramatically only when a scene change occurs, and thus, the renegotiation is necessary only at that time.

In the propose SL negotiation system, MSs negotiate for the peak rates of the VBR applications to ensure high video quality, and BSs allocate the required bandwidth for the entire course of the sessions. Such a characteristic of the propose system made unsuitable the former approach to mitigate the inefficient bandwidth reserved for VBR applications. On the other hand, the individual queues implemented at the BSs (see Fig. 4.5) along with the time-slot assigned to serve each queue allow BSs to detect when the allocated bandwidth is underutilized. The ideal scenario is where a BS serves the queue associate to a MS during the time-slot assigned to that MS, and during the serving time there are packets into the queue. In contrast, if the serving queue becomes empty, the BS has not packet to transmit until any packet arrive to the serving queue or the current time-slot finish and the BS goes to serve the queue associate to the next time-slot.

To tackle the inefficient use of the allocated bandwidth for VBR traffic we use the concept of priority queue as follow.

- Designate the queue to be served as a Priority queue; when a time-slot of a MS starts, the queue associated with this MS becomes the priority queue
- The priority queue is exclusively served during the time-slot while there are packets into queue.
- When the priority queue becomes empty, BS serves the best-effort traffic queue.
- When a packet arrives at the priority queue, the BS completes the transmission of the current packet and goes back to serve the priority queue.

In this fashion, the BS serves best-effort traffic during the empty gaps of the time-slot of MS executing VBR applications.

4.7 Performance evaluation

Haven described the details of the bandwidth aggregation control mechanism, the multipath scheduling algorithm, and the efficient bandwidth utilization scheme, we now direct our focus to evaluating their performances.

4.7.1 Evaluation for the Bandwidth Aggregation Control Mechanism

To illustrate the benefits behind the use of a BAG control mechanism, we conduct some simulations using the Network Simulator (NS2). In the conducted simulations, unless otherwise specified all mobile stations are equipped with three interfaces. The three interfaces are assumed to correspond to different wireless technologies supported by the same ISP in a single domain as shown in Fig. 4.2.

The number of MSs is varied from 20 to 200. The bandwidth level indicated in the SLA of each MS varies from 300Kbps to 2Mbps.

Two SLS negotiation approaches are studied. In the first approach, mobile users negotiate their SLSs with the network through their interfaces. The network verifies only if the requested bandwidth in single SLS requests do not exceed the contracted one in the SLA. This approach is henceforth referred to as Uncontrolled BAG method. In the second approach (dubbed as Controlled BAG), the network ensures that the total bandwidth assigned to a mobile station, via its available interfaces, does not exceed that of the agreed SLA.

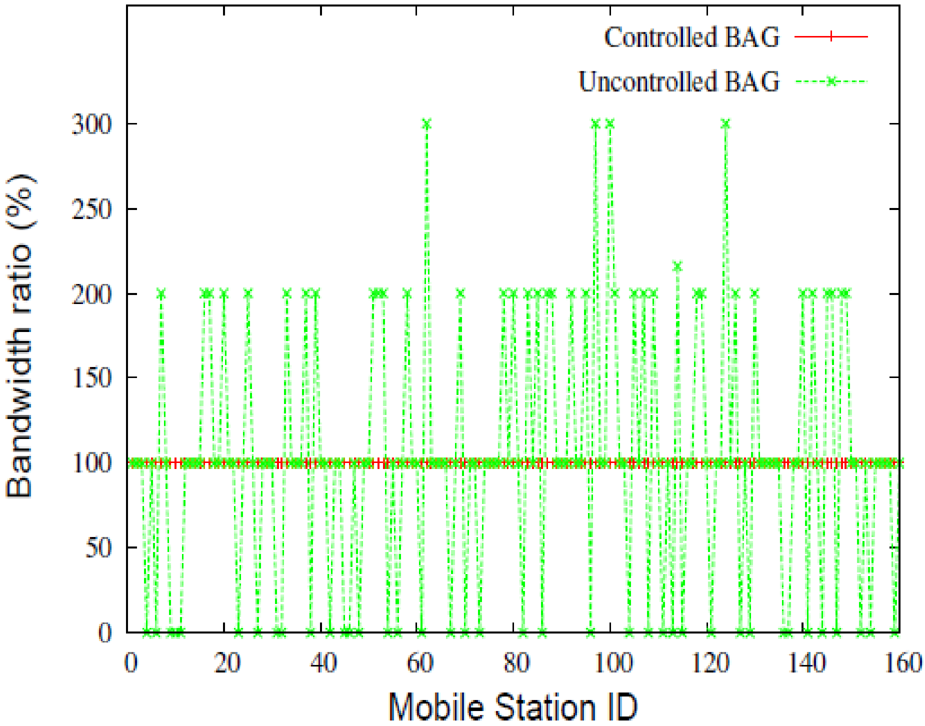

Figure 4.8 Ratio of bandwidth used to that of SLA for 160 mobile stations

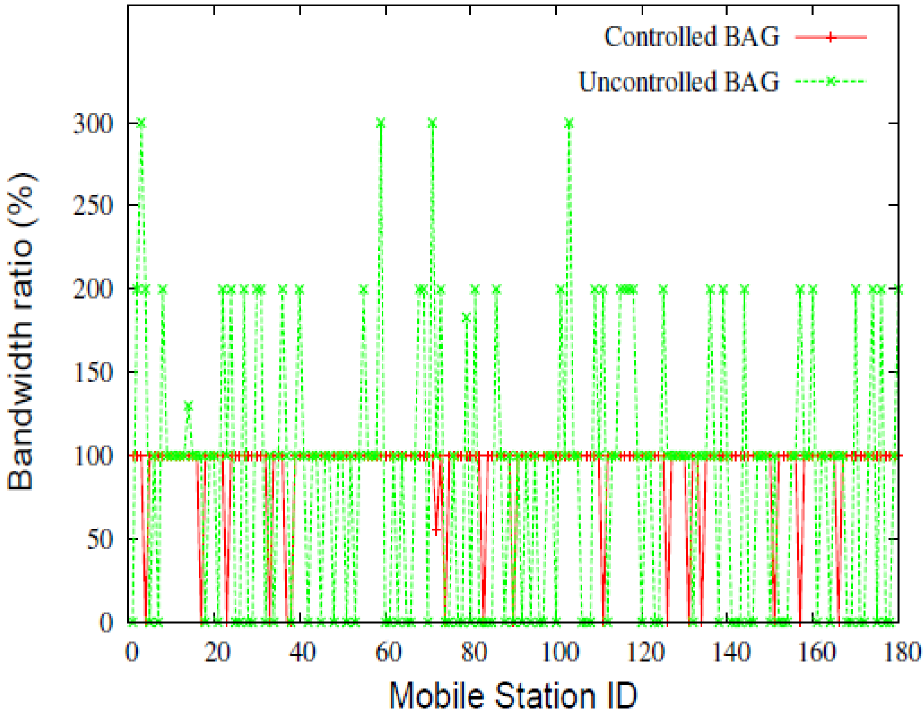

Figure 4.9 Ratio of bandwidth used to that of SLA for 180 mobile stations

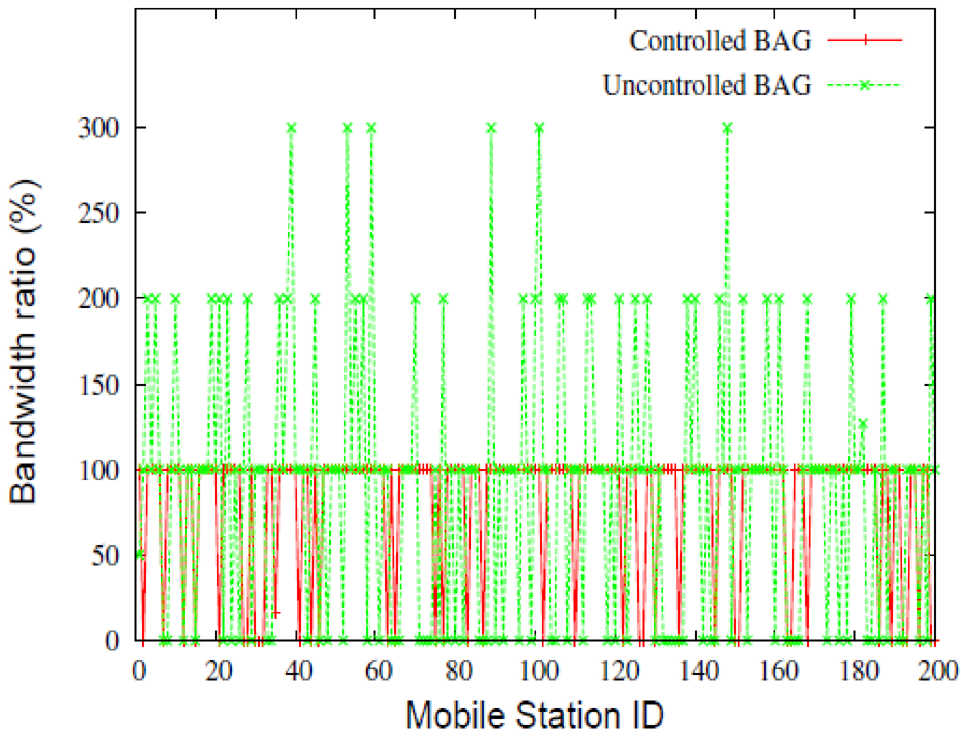

Figure 4.10 Ratio of bandwidth used to that of SLA for 200 mobile stations

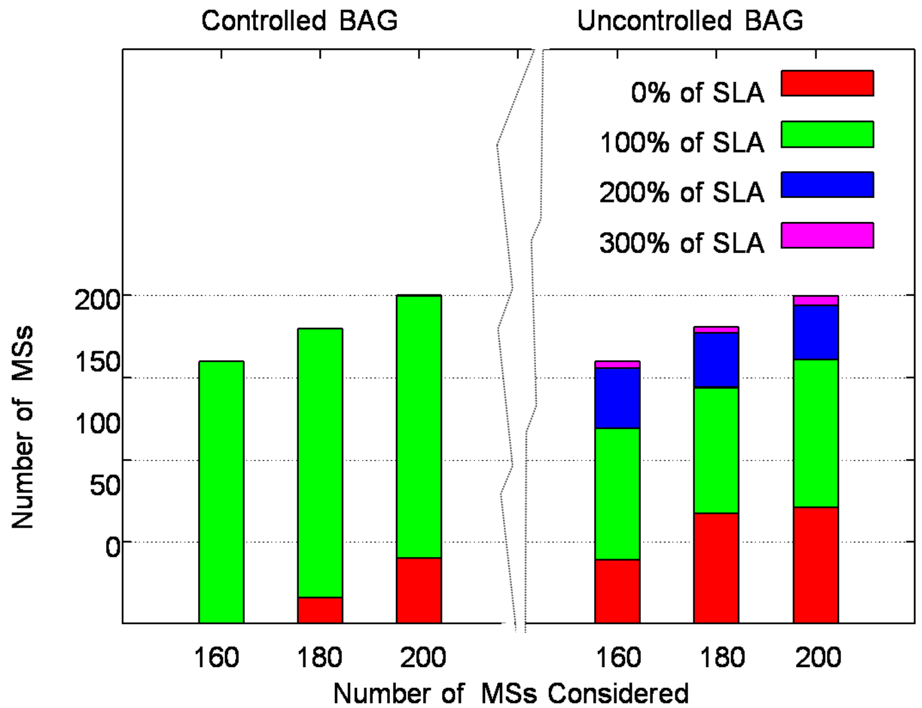

Fig. 4.11 Comparison of ratio of bandwidth used to that of the SLA

Figures 4.8, 4.9, and 4.10 show the ratio of the individual bandwidth actually used by each mobile station to that of its SLA. The figures consider three populations of mobile users (160, 180, and 200 MSs respectively). The figures show that when BAG is not controlled, some mobile stations get up to three times their agreed bandwidth depriving others from having access to the bandwidth they subscribed for. This intuitively results in an unfair service, a fact that is illustrated in the variation of the bandwidth ratio from 0% to 300%. When the BAG control mechanism is in use, each MS receives a bandwidth in the range of 0% to 100% of its SLA. In case of 160 MSs shown in Fig. 4.8, all mobile nodes are provided with bandwidth equal to that of their SLA. This demonstrates that the BAG control mechanism makes efficient use of the aggregate bandwidth of the three simulated bases stations. In the absence of such BAG control mechanism, the system ends up by allocating 300% of SLA to few MSs, 200% of SLA to others MSs and 0% to many MSs. This obviously puts both the scalability and fairness of the system in question. When the network is visited by a high number of mobile nodes (180MSs Fig. 4.9, and 200 MSs Fig. 4.10) and the network resources become scarce, the BAG control mechanism rejects requests of some mobile nodes but its performance remains comparatively much better than that of the uncontrolled BAG approach.

Fig. 4.11 shows the comparison between Controlled Bag and Uncontrolled BAG in term of the number of MSs that obtained zero, one two and three times the bandwidth indicated in their SLAs. When there are 160 MSs the bandwidth is enough to provide each MS with the bandwidth indicated in its SLA as Controlled BAG did. However, Uncontrolled BAG assigns to some MSs up to 3 times the bandwidth indicated in their SLA, as a result, the number of MS that cannot access the network is increase.

Fig. 4.12 plots the utilization rate of the BS resources when the two approaches are applied. The figure shows that when the BAG control mechanism is used and the number of visiting MSs is 60, there is still room for BSs to serve other mobile nodes as their bandwidth utilization rate is still lower than 50% (around 40% in average). However, as is appreciable in Fig. 4.13, when the BAG control mechanism is not used, the utilization rate of $BS1$ and $BS3$ reach 100% and that for $BS2$ is almost 100%. Thus the total network resources are consumed by merely 60 MSs. However, in case of the BAG control mechanism, the scalability of the system goes up to 160 MSs.

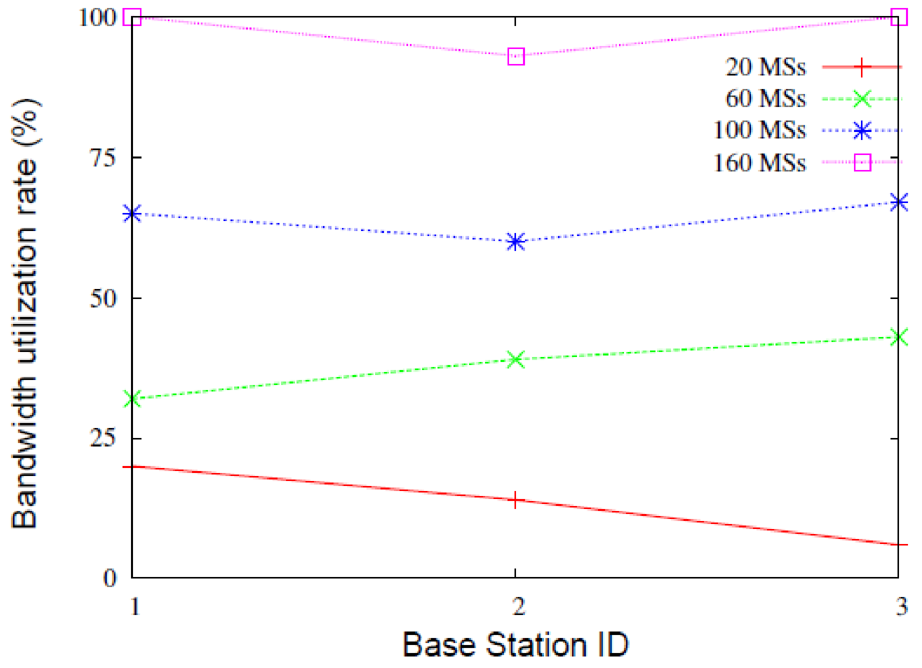

Figure 4.12Bandwidth utilization rate for Controlled BAG

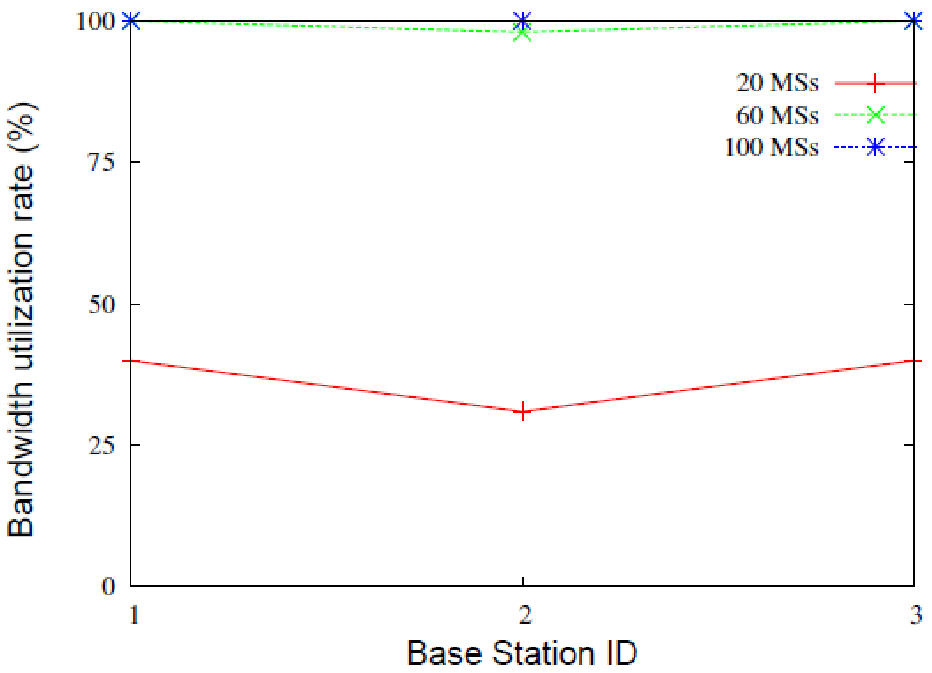

Figure 4.13 Bandwidth utilization rate for Uncontrolled BAG

To illustrate the idea with more clarity, we plot the blocking probability of MSs for different number of mobile stations. The results are shown in Fig. 4.14. Based on the number of wireless technologies in use, two scenarios are considered. Firstly, use of two interfaces and use of three interfaces. The goal behind this experiment is to investigate the impact of the number of deployed interfaces on the system scalability. The figure shows that in case of the BAG control mechanism, the system starts blocking requests when the number of mobile stations exceeds 100 and 160 when two and three IFs are used, respectively. In the absence of such BAG control mechanism, the blocking probability gets non-null values earlier, in the presence of few MSs (i.e., 60 MSs when three IFs are used). Based on the above results, it can be concluded that in the absence of a BAG control mechanism, MSs are allocated bandwidths exceeding that of their SLA. This renders the ISP unable to control its own resources. This ultimately results in an unfair service and high blocking probability.

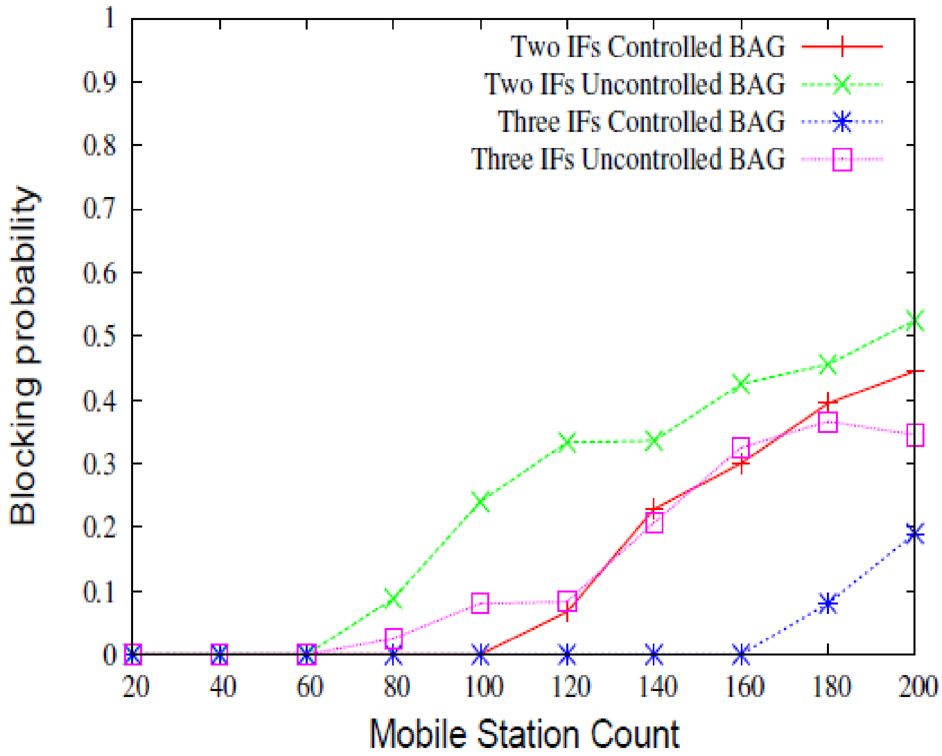

Figure 4.14 Blocking probabilities for different number of interfaces

4.7.2 Evaluation of TS-EDPF scheduling algorithm

In order to verify the effectiveness of the proposed multi-path scheduling algorithm TS-EDPF, we conducted several simulations. As comparison terms, we used the three most suitable algorithms for the proposed QoS negotiation system, namely weighted round robin (WRR), weighted interleaved round robin (WIRR), and earliest delivery path first (EDPF) scheduling algorithms.

WRR and WIRR are able to use the knowledge of the negotiated bandwidth through each available path for an accurate and effortless distribution of packets among them. On the other hand, the EDPF scheduling algorithm additionally makes use of the delays between the network proxy and the BSs to estimate the delivery times of packets via each available path.

As previously mentioned, in the SL negotiation system, MSs negotiate the amount of bandwidth required for their traffics. A MS with multiple interfaces should negotiate the bandwidth to use through each of these interfaces. After each successful bandwidth negotiation, the network proxy is informed of the time-slot assigned to the MS. An important feature of TS-EDPF algorithm is that the network proxy does not need to know the quantity of the negotiated bandwidth for the MS because the architecture uses a time-slot division strategy to guarantee the SL to MSs. Thus, the bandwidth of each MS through the wireless link l used to calculate the transmission time of a packet (variable B_l in Eq. 4.6) is equal to the total bandwidth of this link, which depends on the wireless technology of the BS associated to this link. The network proxy has knowledge of the link bandwidth. It should be reminded that in a time-slot division system, each MS uses the total bandwidth of the link during a short period of time.

Several simulations were performed using the network simulator (NS2). For all the evaluations we consider one MS equipped with three interfaces that correspond to different wireless technologies supported by the same service provider in a single domain as shown in Fig. 4.15.

4.7.2.1 Finding the best value of Δ

For this evaluation the MS has an aggregate bandwidth of 640 Kbps to receive a video streaming from a video server. The maximal transmission delay is set to 300 ms The MS negotiated bandwidth proportionally to the channel capacity for each wireless network, as show Table 4.1. First, we evaluate the scheduling algorithms over two scenarios to find out the best size of the time-slot interval;

- **Scenario 1: Time-slot interval Δ = 1s.**

Time-slot interval of 1s means that the MS will receive the service from each BS once in one second, the specific time of the service are given by the time-slot assigned to the MS during the negotiation process.

Table 4.1 Bandwidth negotiated out of the required one

Network	Bandwidth
Cellular	3%
WiFi	17%
WiMax	80%

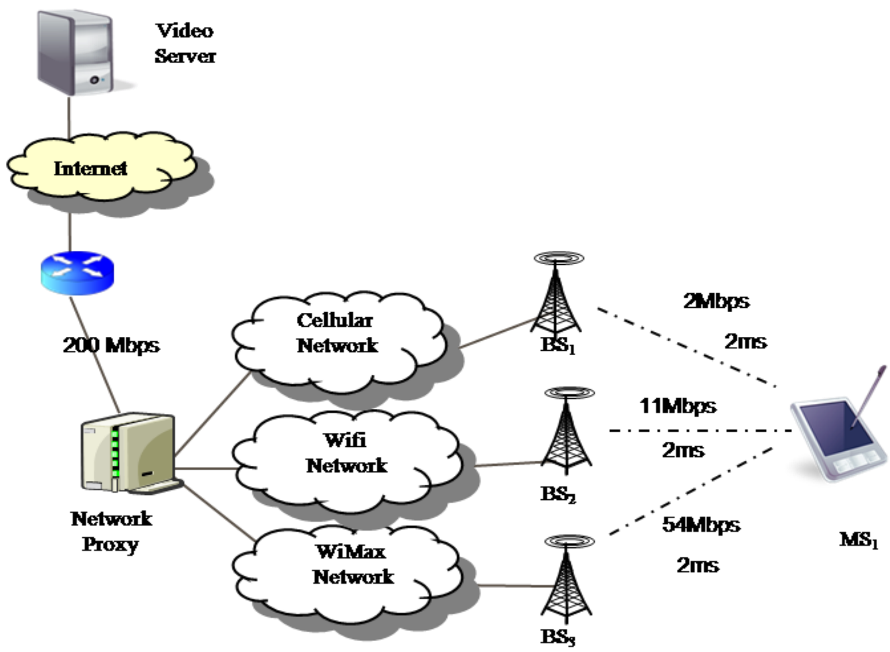

Figure 4.15 Topology for evaluation of TS_EDPF

Fig. 4.16 shows the actual playback time of the first three hundred packets delivered by the evaluated algorithms. The TS-EDPF scheme outperforms clearly all the other schemes. Fig. 4.17 shows the playback delay experienced by packets that indicates the time for which a packet resides in the buffer awaiting the arrival of preceding packets. Notice that among the first three hundred packets only five packets arrived in out-of-order manner in case of TS-EDPF. However, in case of the original EDPF, a quite number of packets arrived out of order and this resulted in a longer reordering delay compared to TS-EDPF.

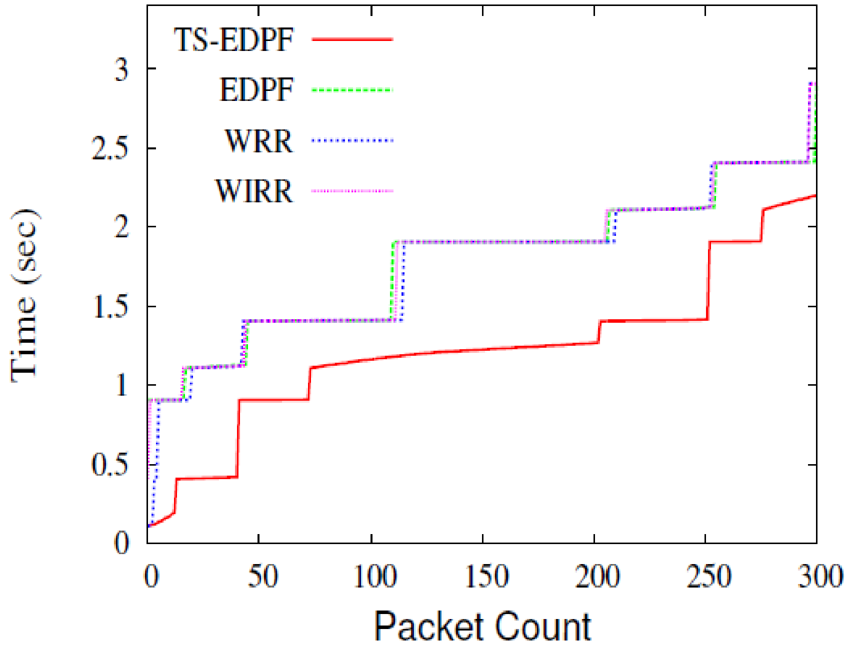

Figure 4.16 Playback time of packets.

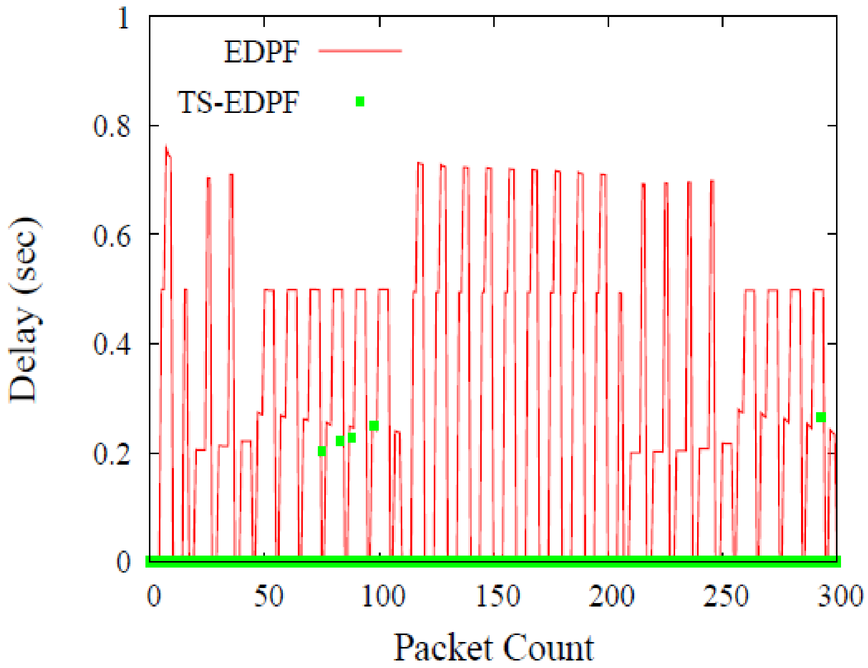

Figure 4.17 Playback delays of packets.

Table 4.2 shows more detailed results. The buffer size reflects the largest number of packets that were queued in the buffer awaiting playback. The bandwidth ratio indicates how much bandwidth the MS could indeed use out of the agreed bandwidth. The disorder delivery radio indicates the proportion of packets that arrived in an out-of-order manner. The results in the table demonstrate that the proposed TS-EDPF scheme outperforms the three other schemes in terms of the overall quantifying parameters. Indeed, the proposed scheme ensures high utilization of the network resources while minimizing the number of packets received out of order and thus reducing the associated reordering delay. However, due to the time-slot approach for bandwidth allocation at the BSs, packets that arrive to the BS in a time different than the time-slot assigned to the destination node are buffered to wait for such time-slot in the next interval of time. Thus their transmission time is dramatically increased, and as a result, many packets are discarded as it is indicated by the high packet loss ratio in the table.

Table 4.2 Comparison among scheduling algorithms in case Δ = 1s

Algorithm	Buffer size (pkts)	Longest playback delay (ms)	Average playback delay (ms)	BW ratio (%)	Disorder delivery ratio (%)	Longest transmission delay (ms)	Average transmission delay (ms)	Packet loss ratio (%)
TS-EDPF	1	205	0.02	99.75	0.01	500	268	80.2
EDPF	78	758	276	98.99	38.67	1003	768	98.03
WRR	83	737	278	98.94	36.21	1001	770	99.04
WIRR	82	787	287	98.75	54.77	1008	788	100

- **Scenario 2: Time-slot interval Δ = 0.1s.**

Time-slot interval of 0.1s means that the MS will receive the service from each BS once in one hundred milliseconds, ten times in one second. Table 4.3 shows the results when the time-slot interval is set to 0.1s. The results of the table indicate that all schemes achieve fairly high throughput. TS-EDPF shows the best performance: the disorder delivery ratio is 0.6%, the average playback delay is 0.1ms, the average buffer size is only one packet, and zero packet loss. This good performance is attributable to the time-slot based policy enforcement strategy adopted by TS-EDPF and lacking in the other three schemes.

Table 4.3 Comparison in case $\Delta = 0.1s$

Algorithm	Buffer size (pkts)	Longest playback delay (ms)	Average playback delay (ms)	BW ratio (%)	Disorder delivery ratio (%)	Longest transmission delay (ms)	Average transmission delay (ms)	Packet loss ratio (%)
TS-EDPF	1	22	0.1	99.89	0.6	253	224	0
EDPF	14	71	13	98.79	43	310	268	6.2
WRR	15	79	7	99.82	12	306	262	7.4
WIRR	15	77	16	99.8	41	305	271	8.5

The Time-slot interval of 0.1s showed to be the best choice; therefore for now on we use that value for Δ.

4.7.2.2 Evaluating the proposed TS-EDPF

Having obtained the best value for the time slot interval Δ, we now proceed to evaluate TS-EDPF for Constant Bit Rate (CBR) applications and also Variant Bit Rate (VBR) applications.

Table 4.4 Different applications used to evaluate TS-EDPF

Application	Bit rate	Peak rate	Mean rate
CBR_1	2.6 Mbps	-	-
VBR_1	-	2.6 Mbps	790 kbps
VBR_2	-	3.2 Mbps	1.14 Mbps

Three video applications are used in this simulation as shows Table 4.4. CBR_1 is a constant bit rate application with a data rate of 2.6 Mbps. VBR_1 and VBR_2 are variable bit rate video traces collected from [95]. VBR_1 corresponds to the MPEG-4 trace of the movie Jurassic Park-I, generated at high quality with peak rate equal to 2.6 Mbps and mean rate of 790 kbps that represents 30.38% of the peak rate. And VBR_2 Corresponds to the MPEG-4 trace of a soccer game also generated at a high quality with peak rate of 3.2 Mbps and mean rate of 1.14 Mbps that represents 35.63% of the peak rate. The duration of the three video applications is 3600s. We consider one MS equipped with three wireless interfaces of different technologies supported by the same service provider. The MS executes one by one the three video applications in different sessions. Additionally, we use three CBR applications to generate background traffic between the NP and the BSs with bit rate of 15 Mbps, 20Mbps and 25 Mbps respectively.

For each session, the MS negotiates an aggregate bandwidth of 100% of the video application bit rate for CBR or peak rate for VBR traffic, respectively. The required bandwidth is aggregated as shows Table 4.1. The maximum transmission delay for packets of the simulated applications is set to 300ms.

Table 4.5 summarizes the results for the three video applications. The buffer size indicates the largest number of packets that were queued in the buffer awaiting playback. The bandwidth ratio indicates the effective use of the aggregated bandwidth.

The disorder delivery radio indicates the proportion of packets that arrived in an out-of-order manner. The results in the table demonstrate that the proposed TS-EDPF scheme outperforms the three other schemes in terms of the overall quantifying parameters. Indeed, TS-EDPF shows the best performances for CBR_1 application with a disorder delivery ratio of 1.3%, maximum buffer size of only one packet, and the packet loss of 0.002%.

Table 4.5 Evaluation of the scheduling algorithms

	CBR_1			VBR_1(Jurassic park I)			VBR_2 (Soccer game)		
Algorithm	**BW ratio (%)**	**Disorder delivery ratio (%)**	**Packet loss ratio (%)**	**BW ratio (%)**	**Disorder delivery ratio (%)**	**Packet loss ratio (%)**	**BW ratio (%)**	**Disorder delivery ratio (%)**	**Packet loss ratio (%)**
TS-EDPF	97.2	1.3	0.002	30.8	3.3	0.001	35.8	3.1	0.002
EDPF	97.1	61.8	2.84	30.7	51	4.2	35.8	54.9	4.7
WRR	97.1	52.6	4.53	30.7	37.3	7.1	35.8	43.7	7.7
WIRR	97.1	61.8	2.79	30.7	53.5	5.1	35.8	56.9	4.9

For VBR_1, the disorder delivery ratio was 3.3%, the maximum buffer size was three packets, and the packet loss was found to be 0.001%. As for VBR_2, the disorder delivery ratio was 3.1%, the maximum buffer size was 12 packets, and the packet loss was 0.002%. This good performance of TS-EDPF is attributed to the adoption of time-slot based policy that was not considered in the other three schemes. It should be noted that a value of 0.002% as packet loss rate means that the scheduling algorithm is accurately delivering data packets to the MS.

The results indicate that all schemes use efficiently the aggregated bandwidth for the CBR application, achieving fairly high throughputs. On the other hand, for VBR_1 and VBR_2, the bandwidth utilization rates are around 30% and 35%, respectively. This means that around 70% and 65% of the negotiated bandwidth for VBR1 and VBR2 remain unused during the applications' running times (i.e., 3600s).

In general, the results in Table 4.5 demonstrate the effectiveness of the proposed TS-EDPF scheduling algorithm as it achieved by far the smallest buffer sizes, smallest disorder delivery ratios, and the lowest packet loss rates for the three considered video applications. The results also demonstrate that the aggregated bandwidth was used up to 97.16%. That is because the video packets cannot be fragmented. Thus if the time needed to transmit the next packet is larger than the remainder time of the time-slot, the packet will be transmitted during the time-slot in the next round. This implies that there may be an unused time at the end of each time-slot. The proposed TS-EDPF aims to minimize that amount of unused bandwidth by scheduling later-arriving smaller packets to be transmitted during that remaining time. Thus, TS-EDPF schedules a few packets to arrive in out-of-order at the receiver to maximize the bandwidth utilization. These packets arriving out of order do not affect the performance of the scheduling process, as they arrive earlier than when they are scheduled in a strict order. They only affect the buffer size, as they have to wait at the MS' buffer until the preceding packets arrive.

The huge difference between the propose TS-EDPF and other considered algorithms in terms of packet loss rate is attributable to the fact that EDPF, WRR and WIRR do not take into consideration the time-slot assigned to the MS to use the wireless channel; these algorithm assume that the BS can transmit packets of any MS at any time. Thus, these algorithms make packets arrive to the BS at any time, and most of them will remain at the BS' queue for up to the time-slot interval Δ= 0.1s waiting for the time-slot of MS in the next round, as show Fig. 4.18. Then, subsequence packets scheduled via others path arrive early and have to wait at the MS' buffer for reordering, making some packets exceed their timer and get discarded.

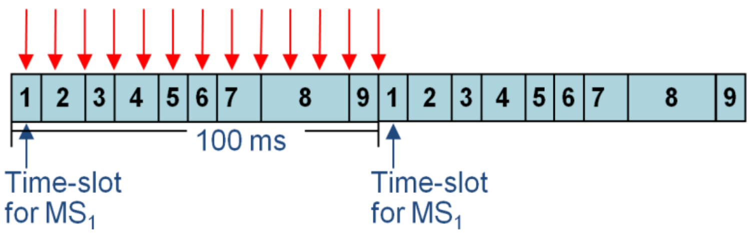

Figure 4.18 Packets arriving out of time at the BS

Another interesting result in table 4.5 is the bandwidth ratio; all considered algorithms achieve similar results for this metric that represent the ratio of bandwidth used out of the negotiated one. That means that the bandwidth utilization of all considered algorithms was very high for CBR (over 97%) and very low for VBR applications (around 30 % and 35% for VBR_1 and VBR_2 respectively).

4.7.2.3. Efficient Bandwidth Utilization

As mentioned earlier, for VBR applications large amounts of the negotiated bandwidth remain unused. This is because users negotiate for the peak rate of VBR applications that is usually reached only once; during the remainder of the transmission, the data rate is much lower than the peak rate. Therefore, huge amounts of the negotiated bandwidth remain unused. Fig. 4.19 deploys the bandwidth utilization of VBR_2 application used in the evaluation of TS-EDPF.

For efficient bandwidth utilization, we implemented a priority queue scheme at BSs. Thus, when a time-slot of an MS starts, the queue associated to this MS becomes the priority queue, which will be exclusively served during the time-slot. When the priority queue becomes empty, BS serves the best-effort traffic queue.

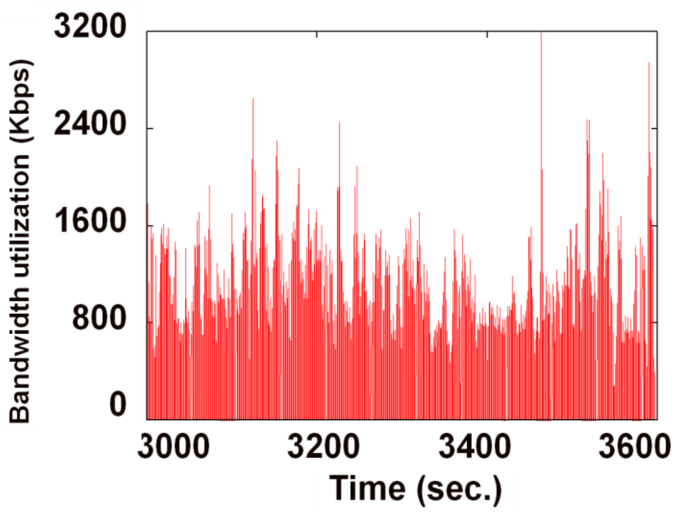

Figure 4.19 Bandwidth utilization for VBR_2

In order to verify the effectiveness of the proposed multi-path scheduling algorithm TS-EDPF executed at the NP, when it is used simultaneously with the propose Priority queue at the BSs, we conducted several simulations and compared the results obtained by the TS-EDPF with the result when both TS-EDPF and Priority Queue are in use.

Fig. 4.20 shows the network topology considered for the simulation. MS_1 is equipped with three interfaces that correspond to different wireless technologies supported by the same service provider in a single domain. MS_2, MS_3, and MS_4 are equipped with one wireless interface corresponding to the wireless technology of BS_1, BS_2, and BS_3, respectively. BSs allocate for Best-Effort (BE) traffic 5% of the channel capacity. MS_2, MS_3, and MS_4 negotiate for best-effort service and each one executes a CBR application of 5 Mbps to keep the Best-effort queue of each BS always full. . Additionally, we use three CBR applications to generate background traffic between the NP and the BSs with bit rate of 15Mbps, 20Mbps and 25 Mbps respectively.

MS_1 executes VBR_1 and VBR_2 applications in different sessions. For each session, the MS negotiates an aggregate bandwidth of 100% of the video application peak rate; 2.6Mbps and 3.2Mbps, respectively. The required bandwidth is aggregated as shows Table 4.1. The maximum transmission delay for packets of the simulated applications is set to 300 ms.

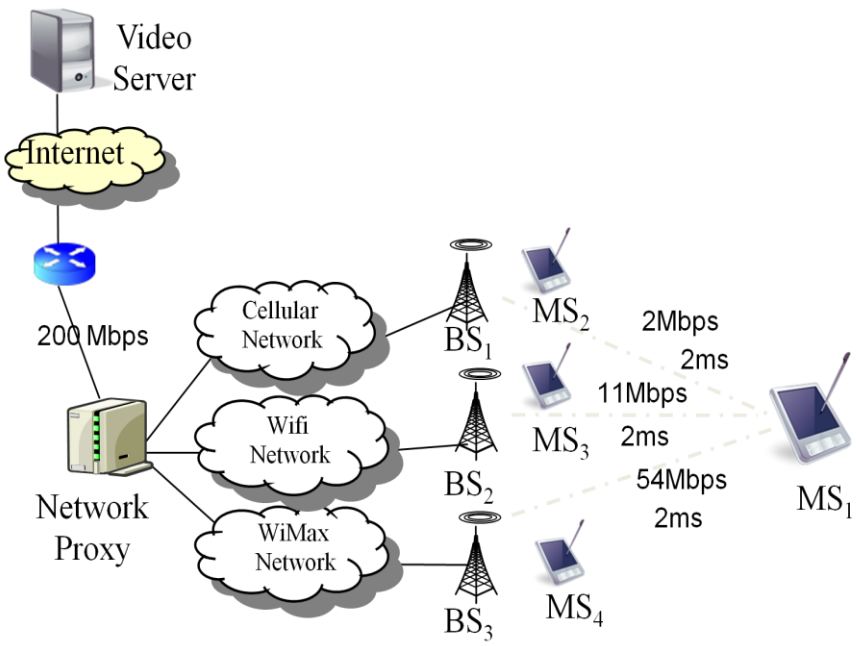

Figure 4.20 Topology for evaluation of TS_EDPF with Priority Queue

Table 4.6 Evaluation TS-EDPF with Priority Queue

	VBR_1(Jurassic park I)			VBR_2 (Soccer game)		
Algorithm	**BW ratio (%)**	**Disorder delivery ratio (%)**	**Packet loss ratio (%)**	**BW ratio (%)**	**Disorder delivery ratio (%)**	**Packet loss ratio (%)**
TS-EDPF	**30.8**	**3.3**	**0.001**	**35.8**	**3.1**	**0.002**
TS-EDPF & Priority Queue	**99.1**	**6.7**	**0.006**	**98.7**	**5.8**	**0.009**

The results in Table 4.6 demonstrates the effectiveness of the incorporated priority queue scheme as it increases the aggregated bandwidth utilization ratio from 30.8% to 99.1% for VBR_1 application, and from 35.8% for to 98.7% for VBR_2 application. However, the priority queue scheme slightly affects the transmission of VBR_1 and VBR_2, by increasing the disorder delivery ratio and the packet loss rate. Thus, the bandwidth utilization during VBR traffics transmission is highly increased using the priority queue at the BSs, at the cost of a slight increment in the packet loss of the VBR application. The priority queue scheme also mitigates the unused bandwidth at the end of the time-slots (as mentioned in Section 4.6.2.2) by serving best-effort traffic during these times.

4.8 Summary

The simultaneous use of multiple interfaces for wireless communication was motivated and included in the QoS negotiation architecture, allowing users to aggregate bandwidth of the multiple interfaces to increase application's throughput. The use of multiple paths with varying characteristics introduces issues in the form of SLA management and packet reordering at the receiver side. To overcome the issues related to SLA management, the addition of bandwidth aggregation control mechanism to the QoS negotiation System was argued. Indeed the evaluation results demonstrated that in presence of such bandwidth aggregation control mechanism, the system tends to be much more scalable and fair. To cope with packet reordering, an enhanced version of EDPF scheduling algorithm termed Time-Slotted Earliest Delivery Path First (TS-EDPF) was developed, which use the information of the time-slot assigned to the MS through each wireless link to estimate the delivery time of packets. Extensive simulation was performed and the results show that TS-EDPF outperforms the most popular scheduling schemes and represents an efficient strategy to deliver real-time video packets under the considered QoS system.

We also propose a mechanism to mitigate the effect of the bursty nature of VBR applications in the dynamic SL negotiation system. Performed evaluations demonstrate the efficiency of the propose scheme, which highly increases the bandwidth utilization rate.

Chapter 5

Concluding remarks

In this thesis, focus was on improving the user experience while getting service from wireless networks by ensuring the continuity of the service level perceived by users while they are on the move between different BSs. The thesis clarified important technical challenges and discussed related issues. With respect to QoS negotiation over wireless networks, the thesis proposed a dynamic service level negotiation system that allows users to dynamically negotiate their desire service levels for their respective traffic, based on the resources required by the applications that they attempt to execute. In order to guarantee the negotiated service levels to users, BSs implement a time-slot division approach, which allows each MS to utilize the total bandwidth of the channel during the time-slots exclusively assigned to the MS. Thus, the BSs avoid collisions in the wireless channel and provide users with the specific amount of bandwidth indicated in their SLSs (thereby providing strict QoS). Users are allowed to renegotiate their current service levels when the resource-requirements of their applications change. In addition, the QGS may require users to degrade their service levels when the resource becomes scarce in the network. However, this is a request only, to establish a new SLS both MS and QGS should agree upon.

In order to provide continuity of the service level perceived by a user upon handoff, this thesis summarized the mobility management strategies and proposed an enhanced version of the most suitable scheme dubbed as ESLS, which presents the desired characteristics in term of scalability and handoff negotiation delay while raising some security issues. An improvement over ESLS was introduced to deal with these security issues. The proposed mobility management mechanism called Extended Encrypted SLS (EESLS) is robust enough to prevent malicious users from stealing the service of legitimate users, and avoid legitimate users altering their own SLSs to obtain better service than that they are allowed to receive. Performance evaluation of the proposed schemes relied on computer simulations and appropriate sets of scenarios were taken into account. The obtained results demonstrated the high scalability of EESLS and also its efficiency in providing seamless handoffs.

Having the underlying resource constraints in wireless networks and the ever-growing demand for bandwidth intensive applications, the dynamic SL negotiation system allows user to increase their data transmission by negotiating available bandwidth of multiple wireless networks. This feature aids the wireless users to reach their desire service levels for their respective traffic flows. However, such a bandwidth aggregation scheme implies transmission of data belonging to a single application via multiple paths with different characteristics, which may result in an out-of-order delivery of data packets to the receiver and introduce additional delays for packets reordering. To distribute the data packets through multiple paths efficiently, we proposed a multi-path scheduling algorithm called TS-EDPF, which makes use of the information related to negotiated bandwidth and the time-slot assigned to the MSs at each BS for an accurate scheduling. Conducted simulations proved the efficiency of TS-EDPF in minimizing the reordering delay and the associated packet loss rate.

The evaluation of TS-EDPF pointed out the under utilization of the negotiated bandwidth for VBR applications. Users generally negotiate for the peak rate of VBR applications to ensure high quality during the whole session. Since the mean rate for those applications is much lower than the peak rate, a big amount of the reserved bandwidth remains unused. To deal with this issue, we use the concept of priority queue to fill up the empty gaps of the time-slots with best-effort traffic. The evaluation demonstrated the effectiveness of these schemes.

The proposed service level negotiation system will allow users to be aware of the service level that the network can provide them with. Thus, users may decide whether to execute their desired applications or to keep waiting further for a better network condition. ISPs ensuring continuity of the negotiated service level to users will greatly improve the Quality of Experience (QoE) as perceived by the users.

Finally, it should be emphasized that the scalability of the system was considered at every phase of its design. Given the centralized nature of QGS and NP, these are the more susceptible components to scalability. However, if it is required, the population of BSs can be distributed in several geographic groups controlled by different QGSs, this change does not affect in any way the functionality of the system. With respect to the NP, it is the entity to execute the scheduling algorithm based on packet by packet estimation of the delivery time through each available path, the improvement of its scalability deserves further study. This forms the basis of our future work.

Bibliography

[1] M. Van Der Schaar, et al., "Cross-layer wireless multimedia transmission: challenges, principles, and new paradigms", IEEE Wireless Communications Magazine, vol. 12, no. 4, pp. 50-58, Aug 2005.

[2] M. Malli, et al., “Adaptive fair channel allocation for QoS enhancement in IEEE 802.11 wireless LAN”, in Proc. of IEEE ICC’04, vol. 6, pp. 347 –3475, Paris, France, Jul 2004.

[3] I. Aad , C. Castelluccia, "Differentiation mechanisms for IEEE 802.11", in Proc. of IEEE INFOCOM’01, vol. 1, pp.209–218, Anchorage, Alaska, Apr 2001.

[4] J. Deng and R.S. Chang, “A priority scheme for IEEE 802.11 DCF access method”, in IEICE Transactions on Communication, vol. E82-B, no. 1, pp. 96-102, Jan 1999.

[5] L. Romdhani, et al., “Adaptive EDCF: enhanced service differentiation for IEEE 802.11 wireless ad-hoc networks”, in Proc. of IEEE WCNC’03, vol. 2, pp. 1373-1378, New Orleans, Lousiana, Mar 2003.

[6] J. Deng, et al., “A new backoff algorithm for the IEEE 802.11 distributed coordination function”, in Proc. of CNDS '04, San Diego, USA, Jan 2004.

[7] Z.J. Haas and J.Deng, "On optimizing the backoff interval for random access schemes", in IEEE Transactions on Communications, vol. 51, no. 12, pp. 2081-2090, Dec 2003.

[8] IEEE 802.11e, Wireless LAN Medium Access Control (MAC) Enhancements for Quality of service (QoS), 802.11e Draft 8.1, May 2005.

[9] A. Nafaa, Y. Hadjadj-Aoul, and A. Mehaoua, “On interaction between loss characterization and forward error correction in wireless multimedia communication”, in Proc. of IEEE International Conference on communication, ICC 05, vol. 2, May 200.

[10] J. Zou and D. Zhao, “Real-time CBR traffic scheduling in IEEE 802.16-based wireless mesh networks”, in Wireless Networks, vol. 15, no. 1, pp. 65-72, Jan 2009.

[11] Y. Iraqi and R. Boutaba, “Supporting MPEG video VBR traffic in wireless networks”, in Computer Communications, vol. 24, no. 12, pp. 1188-1201, Jul 2001.

[12] T. Nguyen, N. Boukhatem, Y. Doudane, and G. Pujolle, "COPS-SLS: A service-level negotiation protocol for the Internet," IEEE Commun. Mag., vol. 40, no. 5, pp. 158–165, May 2002.

[13] D. Durham, Ed., J. Boyle, R. Cohen, S. Herzog, R. Rajan, and Sastry, "The COPS (common open policy service) protocol," IETF RFC 2748, Jan 2000.

[14] X. Wang and H. Schulzrinne, "An integrated resource negotiation, pricing, and QoS adaptation framework for multimedia applications," in IEEE J. Select. Areas Commun, vol. 18, no. 12, pp. 2514–2529, Dec 2000.

[15] Tequila Consortium, "SrNP: Service Negotiation Protocol", 2001. [Online]. Available: http://www.ist-tequila.org/deliverables.

[16] QBone Signaling Design Team, "SIBBS: Simple Inter-domain Bandwidth Broker Signaling." http://qbone.internet2.edu/bb/index.shtml (last accessed on April 20, 2007), Jul 2002.

[17] S. V. den Bosch, G. Karagiannis, and A. McDonald, "NSLP for Quality-of-Service signaling." IETF Internet Draft, [Online]. Available: http://tools.ietf.org/html/draft-ietf-nsis-qos-nslp-17

[18] J. A. Colas, et al., "Connecting Ambient Networks - Architecture and Protocol Design (release 1)," Ambient Networks Consortium, Deliverable D 3.2, Mar 2005.

[19] J.C. Chen, et al., "Dynamic Service Negotiation Protocol (DSNP) and Wireless DiffServ", in Proc of ICC 2002, Vol. 2, USA, pp 1033-1038, Apr 2002.

[20] T. F. Abdelzaher, E. M. Atkins, and K. G. Shin, "QoS negotiation in real-time systems and its application to automated flight control", in IEEE Transactions on Computers, pp. 1170-1183, Nov 2000.

[21] A. Hafid, G. von Bochmann, and B. Kerherve, "A quality of service negotiation procedure for distributed multimedia presentational applications", in Proc. of 5th IEEE International Symposium on High Performance Distributed Computing (HPDC-5), (Syracuse, NY), pp. 330. 339, Aug. 1996.

[22] T. Plagemann, K. A. Saethre, and V. Goebel, "Application requirements and QoS negotiation in multimedia systems," in Proc. of 2nd International Workshop on Protocols for Multimedia Systems (PROMS), (Salzburg, Austria), Oct. 1995

[23] V. Sarangan and J.C. Chen, "Comparative study of protocols for dynamic service negotiation in next generation Internet", in IEEE Communications Magazine, pp. 151.156, Mar 2006

[24] 3rd Generation Partnership Project (3GPP), Technical Specification Group (TSG) RAN, "RAB quality of service negotiation over Lu", Version 4.0.0, Mar 2001.

[25] 3rd Generation Partnership Project (3GPP), Technical Specification Group, "Services and System Aspects, .General packet radio service (GPRS); service description; stage 2 (release 6)", 3GPP TS 23.060, Version 6.5.0, Jun 2004.

[26] J.C. Chen and T. Zhang, "IP-Based Next-Generation Wireless Networks", Wiley, Jan 2004.

[27] P. Gevros, J. Crowcroft, and P. Kirstein, "Congestion control mechanisms and the best effort service model", in IEEE Network, vol. 15, no, 3, pp. 16-26, 2001

[28] A. Talukdar, B.R. Badrinath, and A. Acharya, "Integrated services packet networks with mobile hosts: Architecture and performance", in Wireless Networks, vol. 5, no. 2, pp 111-124, Mar 1999

[29] J. Chiu, Z. Huang, C. Lo, W. Hwang, and C. Shieh, "Supporting End-to-End QoS in DiffServ/MPLS Networks," in Proc. International Conference on Telecommunications, ICT 2003, vol. 1, pp 261- 266, 2003

[30] T. Taleb and A. Nafaa, "A fair and dynamic auction-based resource allocation scheme for wireless mobile networks," in Proc. IEEE ICC'08, Beijing, China, May 2008

[31] C.-C. Chang, H. Sardesai, A. Weiner, "Code-division multiple-access encoding and decoding of femtosecondoptical pulses over a 2.5-km fiber link", Photonics Technology Letters, IEEE vol. 10, no. 1, pp 171 – 173, Jan 1998

[32] M. Dohler, A. Gkelias, and H. Aghvami, "Resource allocation for FDMA-based regenerative multihop links," in IEEE Transactions on Wireless Communications, vol. 3, no. 6, pp 1989 – 1993, Nov. 2004.

[33] Q. Dai, L. Rong, H. Hu, and G. Su, "Resource allocation using time division multiple access over wireless relay networks," The Journal of China Universities of Posts and Telecommunications, vol. 15, no. 3, Pages 69-74, Sep 2008

[34] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks - part I: theoretical framework," in IEEE Trans. Wireless Comm., vol.4, no. 2, Mar 2005.

[35] W. Fernando and R. Rajatheva, "Performance of COFDM for LEO satellite channels in global mobile communications," In Proc. Vehicular Technology Conference, VTC 98, 1998.

[36] L. Milstein, "Wideband Code Division Multiple Access," in IEEE Journal on Selected Areas in Communications, vol. 18, no. 8, Aug. 2000

[37] P. Cardieri and T. Rappaport, "Channel allocation in SDMA cellular systems" in Proc. from Vehicular Technology Conference, VTC 2001, Brazil 2001

[38] G. Kramer, B. Mukherjee, and G. Pesavento, "Ethernet PON (ePON): Design and analysis of an optical access network," Photonic Network Communication, vol. 3, no. 3, pp. 307–319, Jul. 2001.

[39] J. C. Fernandez and N. Kato, "Multi-path video streaming in wireless networks using time-slot based scheduling," Int. J. Bus. Data Commun. Netw. (IJBDCN), no. 4, vol. 4, pp. 13-23, Oct. 2008.

[40] F. Cuervo, et al., "Megaco protocol" version 1.0. IETF RFC 3015, Nov. 2000.

[41] P. Srisuresh, J. Kuthan, J. Rosenberg, A. Molitor, and A. Rayhan, "Middlebox communication architecture and framework", IETF RFC 3303, Aug. 2002.

[42] I. F. Akyildiz, J. Xie, and S. Mohanty, "A survey on mobility management in next generation all-IP based wireless systems," in IEEE Wireless Commun., vol. 11, no. 4, pp. 16–28, Aug. 2004.

[43] N. Banerjee, W. Wu, S. Das, S. Dawkins, and J. Pathak, "Mobility support in wireless Internet," IEEE Wireless Commun., vol. 10, no. 5, pp. 54–61, Oct. 2003.

[44] R. Ramjee, K. Varadhan, L. Salgarelli, S. R. Thuel, S. Y.Wand, and T. L. Porta, "Hawaii: A domain-based approach for supporting mobility in wide-area wireless networks," IEEE/ACM Trans. Netw., vol. 10, no. 3, pp. 396–410, Jun. 2002.

[45] M. Liu, Z. Li, X. Guo, and E. Dutkiewicz, "Performance analysis and optimization of handoff algorithms in heterogeneous wireless networks," IEEE Trans. Mobile Comput., vol. 7, no. 7, pp. 846–857, Jul. 2008.

[46] S. Mohanty and I. F. Akyildiz, "Performance analysis of handoff techniques based on mobile IP, TCP-migrate, and SIP," IEEE Trans. Mobile Comput., vol. 6, no. 7, pp. 731–747, Jul. 2007.

[47] I. F. Akyildiz and W. Wang, "A predictive user mobility profile for wireless multimedia networks," IEEE/ACM Trans. Netw., vol. 12, pp. 1021–1035, Dec. 2004.

[48] J. C. Chen, A. McAuley, V. Sarangan, S. Baba, and Y. Ohba, "Dynamic service negotiation protocol (DSNP) and wireless DiffServ," in Proc. ICC 2002, New York, Apr. 2002.

[49] J. C. Fernandez, T. Taleb, N. Ansari, K. Hashimoto, N. Kato, and Y. Nemoto, "Dynamic QoS negotiation for next generation wireless communications systems," in Proc. WCNC 2007, Hong Kong, China, Mar. 2007.

[50] M. Barry, A. T. Campbell, and A. Veres, "Distributed control algorithms for service

differentiation in wireless packet networks," in Proc. IEEE Infocom 2001, Anchorage, AK, Apr. 2001.

[51] CB/LBNL/VINT: Network Simulator ns (Version 2). [Online]. Available: http://www.isi.edu/nsnam/ns/.

[52] W. Fritsche and F. Heissenhuber, "Mobile IPv6: Mobility support for Next Generation Internet," IPv6 Forum, White Paper, 2000.

[53] C. Perkins, "IP Mobility support for IPv4," RFC 3344, Aug. 2002.

[54] R. Vadali, J. Li, Y.Wu and G. Cao, "Agent-Based Route Optimization for mobile IP," IEEE Commun. Mag., vol. 43, no. 12, pp. 156-163, Dec. 2005.

[55] R. Stewart et al., "Stream control transmission protocol," in RFC 2960, Oct. 2000.

[56] S. Fu and M. Atiquzzaman, "SCTP: State of the art in research, products, and technical challenges," IEEE Commun. Mag., vol. 42, no. 4, pp. 64–76, Apr. 2004.

[57] L. M. F. Yu and V. C. M. Leung, "A new method to support UMTS/ WLAN vertical handover using SCTP," IEEE Wireless Commun., vol. 11, no. 4, pp. 44–51, Aug. 2004.

[58] R. Fracchia, C. Casetti, C. Chiasserini, and M. Meo, "WiSE: Best-path selection in wireless multihoming environments," IEEE Trans. Mobile Comput., vol. 6, no. 10, pp. 1130–1141, Oct. 2007.

[59] Balk, M. Sigler, M. Gerla, and M. Sanadidi, "Investigation of MPGE4 Video Streaming over SCTP, Sixth World Multiconference on Systemics, Cybernetics, and Informatics (SCI 2002), Orlando, Fl. USA, Jul. 2002.

[60] M. Jain and C. Dovrolis, "End-to-end available bandwidth: Measurement methodology, dynamics, and relation with TCP throughput," IEEE/ACM Trans. Netw., vol. 11, no. 4, pp. 537–549, Aug. 2003.

[61] A. Abdelal, T. Saadawi, and M. Lee, "LS-SCTP: A bandwidth aggregation technique for stream control transmission protocol," Comput. Commun., vol. 27, no. 10, pp. 1012–1024, Jun. 2004.

[62] H. Hsieh and R. Sivakumar, "pTCP: An End-to-End Transport Layer Protocol for Striped Connections," in Proc. of 10th IEEE Int. Conf. on Network Protocols (ICNP 2002), Paris, France, Nov. 2002.

[63] L. Magalhaes and R. Kravets, "MMTP—Multimedia multiplexing transport protocol," in Proc. 1st ACM Workshop Data Communications in Latin America and the Caribbean (SIGCOMM-LA 2001), San Jose, Costa Rica, Apr. 2001.

[64] K. Chebrolu and R.R. Rao, "Communication Using Multiple Wireless Interfaces," Proc. IEEE Wireless Comm. and Networking Conf., Mar. 2002.

[65] D.S. Phatak and T. Goff, “A Novel Mechanism for Data Streaming across Multiple IP Links for Improving Throughput and Reliability in Mobile Environments,” Proc. IEEE INFOCOM ’02 Conf., pp. 773-781, Jun 2002.

[66] L. Magalhaes and R. Kravets, “Transport Level Mechanisms for Bandwidth Aggregation on Mobile Hosts,” Proc. IEEE Int’l Conf. Network Protocol, Nov. 2001.

[67] K. Chebrolu and R. Rao, “Bandwidth aggregation for real-time applications in heterogeneous wireless networks,” IEEE Trans. Mobile Comput., vol. 5, no. 4, pp. 388–403, Apr. 2006.

[68] T. Taleb, T. Nakamura, and K. Hashimoto, “Multi-source streaming in next generation mobile communication systems,” in Proc. IEEE ICC’08, Beijing, China, May 2008.

[69] J.C. Fernandez, T. Taleb, K. Hashimoto, Y. Nemoto, and Nei Kato, “Multi-path Scheduling Algorithm for Real-Time Video Applications in Next-Generation Wireless Networks”, in Proc. of The International Conference on Innovations in Information Technologies (ITT’07), Dubai, UAE. Nov. 2007.

[70] T. Taleb, J.C. Fernandez, K. Hashimoto, N. Kato, and Y. Nemoto, “A Bandwidth Aggregation-aware QoS Negotiation Mechanism for Next-Generation Wireless Networks”, in Proc. of IEEE Global Telecommunications Conference (Globecom'07), Washington DC, USA, Nov. 2007

[71] R. Stewart et al., “Stream control transmission protocol,” RFC 2960, Oct. 2000.

[72] D. S. Phatak and T. Goff, “A novel mechanism for data streaming across multiple IP links for improving throughput and reliability in mobile environments,” in Proc. IEEE INFOCOM’02, New York, pp. 773–781, Jun 2002.

[73] D. Jurca and F. Pascal, “Media flow rate allocation in multipath networks,” IEEE transactions on multimedia,” vol. 9, No. 6, pp. 1227-1240, 2007.

[74] D. Jurca and P. Frossard, “Video packet selection and scheduling for multipath streaming,” IEEE Trans. Multimedia, vol. 9, no. 3, pp. 629–641, Apr. 2007.

[75] S. Vutukury and J. J. Garcia-Luna-Aceves, “MDVA: A distance-vector multipath routing protocol,” in Proc. IEEE INFOCOM, 2001, vol. 1, pp. 557–564.

[76] W. Wei and A. Zakhor, “Multipath unicast and multicast video communication over wireless ad hoc networks,” in Proc. IEEE/ACM Broad-Nets, pp. 496–505, Oct. 2004.

[77] S. Tao and R. Guerin, “Application-specific path switching: A case study for streaming video,” in Proc. ACM Multimedia, pp. 136–143, Oct. 2004.

[78] Z. Ma, H.-R. Shao, and C. Shen, “A new multi-path selection scheme for video streaming on overlay networks,” in Proc. IEEE ICC, 2004.

[79] K. C. Leung and V. O. K. Li, "Flow assignment and packet scheduling for multipath routing," J. Commun. Networks, vol. 5, no. 3, pp. 230–239, Sep. 2003.

[80] J. Chen, S.-H. G. Chan, and V. O. K. Li, "Multipath routing for video delivery over bandwidth-limited networks," IEEE J. Select. Areas Commun., vol. 22, no. 10, pp. 1920–1932, Dec. 2004.

[81] S. Mao, S. Lin, S. S. Panwar, Y.Wang, and E. Celebi, "Video transport over ad hoc networks: Multistream coding with multipath transport," IEEE J. Select. Areas Commun., vol. 21, no. 10, pp. 1721–1737, Dec. 2003.

[82] A. C. Begen, Y. Altunbasak, O. Ergun, and M. H. Ammar, "Multipath selection for multiple description video streaming over overlay networks," Signal Process.: Image Commun., vol. 20, pp. 39–60, 2005.

[83] H. Balakrishnan, V. Padmanabhan, S. Sheshan, and R. Katz, "A comparison of mechanisms for improving TCP performance over wireless links," IEEE/ACM Trans. Networking, vol. 5, no. 6, pp. 756–769, Dec 1997.

[84] K, Chebrolu, B. Raman, and R. Rao, "A network layer approach to enable TCP over multiple interfaces," Wireless Networks, vol. 11 , no. 5, pp 637 - 650, Sep. 2005

[85] H. Hsieh and R. Sivakumar, "A transport layer approach for achieving aggregate bandwidths on multi-homed mobile hosts," in Proc. ACM MOBICOM'02, Atlanta, Sep. 2002.

[86] L. Magalhaes and R. Kravets, "Transport level mechanisms for bandwidth aggregation on mobile hosts," in Proc. IEEE ICNP'01, Riverside, Nov. 2001.

[87] K. Chebrolu and R. R. Rao, "Communication using multiple wireless interfaces," in Proc. IEEE WCNC'02, Orlando, Mar. 2002.

[88] J. Aweya, M. Ouellette, and D. Y. Montuno, "A self-regulating TCP acknowledgment (ACK) pacing scheme," International Journal of Network Management, vol. 12, no. 3, pp. 145–163, May/June 2002.

[89] E. Blanton and M. Allman, "On making tcp more robust to packet reordering," Computer Communication Review, vol. 32, no. 1, pp. 20–30, Jan 2002

[90] M. Zhang, B. Karp, S. Floyd, and L. Peterson, "Improving TCP's performance under reordering with DSACK," International Computer Science Institute, Berkeley, Tech. Rep. TR-02-006, Jul 2002

[91] H. Adiseshu, G. Parulkar and G. Varghese, "A Reliable and Scalable Striping Protocol," ACM Computer Communication, vol. 26, no. 4, pp. 131-141, Oct. 1996.

[92] R. Chandra, P. Bahl and P. Bahl, "Multinet: Connecting to Multiple IEEE 802.11

Networks Using a Single Wireless Card," in Proc. of IEEE INFOCOM, Hong Kong, Mar 2004.

[93] T. Taleb, D. Mashimo, K. Hashimoto, N. Kato, and Y. Nemoto, "On how to mitigate the packet reordering issue in the explicit load balancing scheme," in Proc. IEEE Global Information Infrastructure Symp. (GIIS), Marrakech, Morocco, Jul. 2007

[94] J.C. Fernandez, T. Taleb, M. Guizani, and N. Kato, "Bandwidth Aggregation-aware Dynamic QoS Negotiation for Real-Time Video Streaming in Next-Generation Wireless Networks", IEEE Transactions on Multimedia, no. 6, vol. 11, pp. 1082-1093, Oct. 2009

[95] MPEG-4 and H.263 Video Traces for Network Performance Evaluation, 2006. [Online]. Available: http://www.tkn.tu-berlin.de/research/trace/trace.html.

[96] H. Zhao, N. Ansari, and Y. Q. Shi, "Transmission of real time video over IP Differentiated services", IEE Electronics Letters, vol. 38, pp. 1151 – 1153, Sep. 2002.

[97] S. Bakiras and V. Li, "Efficient resource management for end-to-end QoS guarantees in DiffServ Networks," ICC02, vol. 2, pp. 1220 – 1224, 2002

[98] M. Furini and D. F. Towsley, "Real-time traffic transmissions over the Internet," IEEE Transaction on Multimedia, Vol. 3, No.1, pp. 33 – 40, Mar. 2001

[99] J. Salehi, Z. Zhang, J. Kurose and D. Towsley, "Supporting stored video: Reducing rate variability and end to end resource requirements through optimal smoothing," IEEE/ACM Trans. on Networking, vol. 15, pp. 1148–1166, 1998.

[100] J. M. Mcmanus and K. W. Ross, "Video on demand over ATM: constant-rate transmission and transport," IEEE J. Select. Areas Commun., vol. 14, pp. 1087–1098, Aug. 1996.

[101] L. Zhang and H. Fu, "A novel scheme of transporting pre-stored MPEG video in ATM networks," Computer Communications Elsevier Science, vol. 23, pp. 133–148, 2000

[102] E. W. Fulp and D. Reeves, "Dynamic Bandwidth Allocation for VBR Sources," in Proc. IEEE RTSS Workshop on Resource Allocation Problems in Multimedia Systems, Washington, DC, Dec. 3, 1996.

[103] D. Reininger, M. Ott, G. Michelitsch, and G. Welling, "Dynamic Bandwidth Allocation for Distributed Multimedia with Adaptive QoS", In Proc. IEEE RTSS Workshop on Resource Allocation Problems in Multimedia Systems, Washington, DC, Dec. 3 1996

[104] K. Shiomoto, S. Chaki, and N. Yamanaka, "A Simple Bandwidth Management Strategy Based on Measurements of Instantaneous Virtual Path Utilization in ATM Networks," IEEE/ACM Transactions on Networking, pp. 625–634, 1998

[105] H. Zhang and E.W. Knightlly, "RED: A new approach to support delay sensitive VBR

video in packet switched networks,” Proc. 5th Workshop on Networking and Operating System Support for Digital Video, pp. 275–286, Apr. 1995.

[106] A. M. Adas, “Using Adaptive Linear Prediction to Support Real-Time VBR Video Under RCBR Network Service Model,” IEEE/ACM Transactions on Networking, vol. 6, pp. 635–644, 1998

[107] S. Chong, S. Li, and J. Ghosh, “Dynamic bandwidth allocation for efficient transport of real time VBR video over ATM,” IEEE J. Select. Areas Commun. vol. 13, pp. 12–23, Jan. 1995.

[108] H. Zhao, N. Ansari, and Y. Q. Shi, “A fast nonlinear adaptive algorithm for video traffic prediction,” Proc. of International Conference on Information Technology: Coding and Computing (ITCC2002), Las Vegas, pp. 54–58, Apr. 2002.

[109] B. Melamed and D.E. Pendarakis, “Modeling Full Length VBR Video Using Markov Renewal Modulated TES Models,” IEEE/ACM Transactions on Networking, vol. 5, pp. 600–612, Feb. 1998.

[110] H. Liu, N. Ansari, and Y.Q. Shi, “Dynamic Bandwidth Allocation for VBR Video Traffic Based on Scene Change Identification,” Proc. IEEE International Conference on Information Technology: Coding and Computing (ITCC2000), Mar. 27–29, pp. 284–288, 2000.

.

Journal Papers

- J.C. Fernandez, T. Taleb, M. Guizani, and N. Kato, “Bandwidth Aggregation-aware Dynamic QoS Negotiation for Real-Time Video Streaming in Next-Generation Wireless Networks”, IEEE Transactions on Multimedia, No. 6, Vol. 11, pp. 1082-1093, Oct. 2009

- J.C. Fernandez and N. Kato, “Multi-path video streaming in wireless networks using time-slot based scheduling”, International Journal of Business Data Communications and Networking, No. 4, Vol. 4, pp. 13-23, October-December 2008.

International Conference Papers

- J.C. Fernandez, T. Taleb, K. Hashimoto, Y. Nemoto, and Nei Kato, “Multi-path Scheduling Algorithm for Real-Time Video Applications in Next-Generation Wireless Networks”, in Proc. of The International Conference on Innovations in Information Technologies (ITT’07), Dubai, UAE. Nov. 2007.

- T. Taleb, J.C. Fernandez, K. Hashimoto, N. Kato, and Y. Nemoto, “A Bandwidth Aggregation-aware QoS Negotiation Mechanism for Next-Generation Wireless Networks”, in Proc. of IEEE Global Telecommunications Conference (Globecom'07), Washington DC, USA, Nov. 2007.

- J.C. Fernandez, T. Taleb, N. Ansari, K. Hashimoto, N. Kato, and Y. Nemoto, “Dynamic QoS Negotiation for Next-Generation Wireless Communications Systems”, in Proc. of IEEE Wireless Communications and Networking Conference (WCNC'07), Hong Kong, Mar. 2007.

Local Conference Papers

- Z. Fadlullah, J.C. Fernandez, and Nei Kato, “A Secure and Robust Framework of Service Level Specification for Next Generation Wireless Networks”, in Proc. of the Institute of Electronics, Information and Communication Engineers (IEICE) Conference, Okinawa, Japan, Oct. 2008.