Building the bridges between QoS and QoE for network control mechanisms

Adlen Ksentini

N^o d'ordre: 000

HABILITATION À DIRIGER DES RECHERCHES

présentée

devant l'Université de Rennes 1

Mention Informatique

par

Adlen Ksentini

Équipe : INRIA/Dionysos
École Doctorale : Matisse

Titre de la HDR :

Building the bridges between QoS and QoE for network control mechanisms

soutenue le 3 Juillet 2013 devant le jury :

Président		
Guy	Pujolle	Professeur, Université de Paris 6
Rapporteurs		
André Luc	Beylot	Professeur, INPT/ENSEEIHT Toulouse
Raouf	Boutaba	Professeur, Université de Waterloo
Ahmed	Toufik	Professeur, Université de Bordeaux
Examinateurs		
Andrzej	Duda	Professeur, ENSIMAG Grenoble
Gerardo	Rubino	Directeur de Recherche, INRIA Rennes
César	Viho	Professeur, Université de Rennes 1

Contents

Chapter 1

Introduction

1.1 Research context

This last decade has known an explosion of the wireless access connectivity, mostly dominated by Wireless Local Area Networks (WLAN), 3GPP-based cellular networks (3G and 4G), and to a lesser extent by Digital Video Broadcast (DVB) networks. WLANs ensure a high data rate with short coverage (about a hundred of meters). It is mainly used for domestic accesses (through an ADSL Box), in companies and few large public deployments (e.g. airports, shopping malls, and train stations). Cellular networks are operated networks. They offer higher coverage (in the range of kilometers) supporting high data rates (mainly with Long Term Evolution – LTE or 4G [1]). Besides the radio access, known as Radio Access Network (RAN), a cellular network consists of a wired core network or Evolved Packet Core (EPC) that connects RAN to Packet Data Networks, e.g., Internet.
Combined with the emergence of smartphones and tablet PCs, wireless connectivity has changed the users' way to connect to the Internet, where we clearly observe that most of the connection to video platforms (e.g., Youtube, video on demand and IPTV) and social network applications are originated from wireless as well as mobile networks. However, massive accesses to these applications from wireless networks introduce several issues related to the increasing traffic, and the real-time property of some of these applications, which require a guarantee of Quality of Service (QoS) and Quality of Experience (QoE). These issues give rise to several challenges:

- The need for network operators to have efficient mechanisms and protocols to avoid congestion and to support QoS;
- The need for content providers to consider the user context, in terms of bandwidth, user terminal, resolution, etc.), when creating and encoding an audio/video content.

Each wireless network has its own conception, which makes it different from the other. WLANs (or IEEE 802.11 [2]) are based on a distributed mechanism to share the wireless channel among the stations. Each station, independently from the other, uses the Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA) algorithm to access the wireless channel. The main principle of CSMA/CA is that before sending a packet, a wireless station senses the medium for a fixed duration. If the medium is free, it then transmits its packet and awaits an acknowledgment from the receiver. Otherwise, it backs off its transmission for a random duration.
Unlike WLAN, the RAN part of the cellular network is centralized around the Base Station (BS), where this latter selects stations (or User Equipment – UE in the 3GPP terminology)

having the possibility to transmit in the Uplink frame. The BS, through the Downlink frame, forwards packets destined to UEs. RAN is known as a collision-free access as there is no contention to handle since UEs are attached to the network. Moreover, cellular networks are composed by a wired core network or EPC, which manages the mobility of UEs and forwards the IP packets to the destination.
In the following sections, we will detail the research context of our work for each wireless network, by including the case of wireless terminals having both connectivity (WLAN and 3G) at the same time. Independently from the used network technology, we were also interested in adapting the content format regarding users' context. The ultimate goal of our contributions is to efficiently handle network resources to ensure QoS for real-time applications and to therefore maximize user's QoE.

1.1.1 WLAN or IEEE 802.11

Initially developed to support only best-effort applications, the IEEE 802.11 standard used in WLAN does not include indications to differentiate the channel access between flows from different traffic classes. Indeed, all flows contend for the channel access with the same priority, which is not acceptable in case of real-time applications that usually require strict temporal constraints. Furthermore, it is difficult to guarantee QoS for real-time applications in WLAN without taking in consideration its characteristics, in terms of: (i) topological changes (due to mobility of users); (ii) the dynamic nature of the wireless channel conditions; (iii) the unavailability of information on the network configuration (e.g., the number of stations). In this context, the IEEE 802.11e group has established principles and mechanisms to ensure QoS in WLAN [2]. These mechanisms enable service differentiation between Access (traffic) Classes (AC) and flows. However, there are still some challenges to address in order to efficiently support QoS in WLAN. Indeed, the Medium Access Control (MAC) procedures, proposed by 802.11e group, are static and do not adapt well to the network dynamics. When the number of nodes increases, the QoS for high priority classes cannot be guaranteed. This also highlights another issue that remains unsolved by the 802.11e group, which is admission control procedure. Usually, an admission control algorithm is used to accept or reject a new incoming flow. The main goal of using admission control is to accept enough flows to maximize the network resources utilization without degrading the QoS performances of admitted flows. An effective resource allocation in IEEE 802.11 is difficult to achieve due to the intrinsic nature of the CSMA/CA scheme. The difficulty lies in estimating the value of the achievable QoS performance in WLAN; this value depends on several time-varying factors including the number of active flows, the active traffic volume for each AC, etc.
Efficient multicast support is another challenge to address when deploying multimedia applications over WLAN. We recall that a number of real-time applications are based on multicast communications, such as IPTV, visio-conference, and network gaming. Currently, each wireless station has the choice to use a physical data rate depending on its wireless channel quality. When the channel quality degrades, in terms of Signal Noise Ratio (SNR), the wireless station uses robust modulations (to the Bit Error Rate – BER), which achieves low data rate to transmit data packets. But, when the channel is good, the wireless station can use less robust physical modulations, ensuring high data rate to transmit data packets. For instance, the 802.11b [4] allows using a data rate ranging from 1 Mbps (basic rate with the most robust modulation) to 11 Mbps (lowest robust modulation). Besides avoiding using data acknowledgement to avoid overloading the network when the group size is large, the 802.11 standard proposes using the basic rate for transmitting the multicast packets. In fact, multicast communications in 802.11 are transmitted with the

lower physical data rate to address the lack of reliability in the wireless channel. Accordingly, using basic rate for multicast communication is not optimal for bandwidth-intensive real-time applications, such as IPTV and network gaming. It is worth noting that there is a new 802.11 group, 802.11aa [5], which aims at proposing new MAC procedures for an efficient handling of multicast-based video streaming applications in WLAN. However, the proposed solutions are still in their infancy and need to be yet tested and validated.
Last but not least, energy consumption is a critical issue, which needs to be solved in WLAN, particularly for real-time application. The Power Save Mode (PSM), proposed in the 802.11 standard, aims at minimizing energy consumption when the wireless station is not active. When PSM is enabled, a wireless station goes to sleep for a fixed duration (i.e., multiple times of the Beacon interval) once that there is no data to send or to receive. In this case, packets, destined to this station, are buffered at the Access Point (AP). Since the sleep duration expires, the station wakes up and checks if there are any pending packets at the AP. If so, the station requests these packets from the AP. Otherwise, it goes to sleep for another duration. PSM performs well for best-effort applications, but achieves worst performance in case of real-time applications. In fact, the increase of the end-to-end delays, caused by the fact that data packets can remain for an unknown duration at the AP, represents the main drawback of activating PSM in case of real time applications.

1.1.2 Cellular networks (3G and LTE)

Unlike WLAN, 3GPP-based cellular networks efficiently guarantee QoS for real-time applications. Through the centralized architecture of RAN around BS, cellular networks easily solve the problems cited in the previous WLAN-relevant section. It supports service differentiation by implementing efficient scheduling algorithms at BS, which can give high priority to the real-time applications over the best-effort ones for accessing the channel. Moreover, 3GPP has recently defined new mechanisms around the Multimedia Broadcast and Multicast Services (MBMS) [6] concept, which can handle and manage multicast and broadcast communications in a more efficient way at RAN. The MBMS concept is now gaining great attention; it is even proposed to replace DVB for broadcasting television programs for mobile stations.
Nevertheless, the massive traffic generated by the emerging applications, such as Machine to Machine (M2M) and social networks, has begun to introduce congestion in the EPC part of cellular networks. This kind of applications exhibits a traffic pattern (the uplink traffic is higher or equal to the downlink traffic) highly different from the usual one (downlink traffic is higher than the uplink traffic) handled in the cellular network, such as http and video streaming. It is worth noting that EPC nodes that may be particularly affected by congestion, are the Mobility Management Entity (MME), the Serving Gateway (S-GW) and the Packet Data Network Gateway (PDN-GW).
Cellular-based Machine-to-Machine (M2M) or Machine Type Communications (MTC) are about enabling automated applications that involve machine or device communication without any human intervention over cellular networks. MTC will enable an endless number of applications in a wide range of domains impacting different environments and markets. It will connect a huge number of MTC devices to the Internet and the networks. Depending on the use case, a MTC device transmits or receives a determined amount of data at a determined frequency, e.g., a smart meter sending measurement results every day at 23:00h. MTC devices can be either fix installed (e.g., implemented in a factory's machine, gas meters, etc.) or mobile (e.g., fleet management devices in trucks). Congestion in EPC, particularly at its control plane, occurs when a potential number of MTC devices attempt attaching/connecting to the network all at once, after detecting an event.

As mentioned earlier, this congestion particularly impacts the MME, S-GW and PDN-GW nodes, which ultimately results in the degradation of the network performance and hence affects the QoS support of non-MTC traffic.
Another type of traffic that can affect the EPC nodes and may incur congestion is the one generated by social network applications. Indeed, many social network platforms (e.g., Twitter and Facebook) or news tickers (e.g. CNN and sport events) are based on a one-to-many communication paradigm, i.e., one entity posts a message of the same content which is then received by many users that have "subscribed" to this "news feed". For example, in Japan there is a popular application, called Bijin-Tokei [7], that enables users to receive a photo of a "beautiful girl" holding a black board that shows the current time. These photos are sent to all subscribers (including mobile ones) every one minute as photos need to be updated at the same frequency. Other mobile web applications that involve the delivery of the same content to multiple users being in the same location are location-based "check in" services such as Foursquare (1 million users), Facebook places, Gowalla, Brightkite, Yelp, and Google's Latitude. These applications allow users, particularly mobile users, to check in at locations they visit as a way to find other friends, coordinate gatherings and exchange content of common interest among a "social network" of users. There are further many emerging mobile games, allowing users to play a game relevant to their current location and with other users in the same location, e.g., SCVNGR and Zynga, resulting therefore in a frequent and dynamic exchange of content among a group of mobile users in the same neighborhood. The problem today is that every user establishes a point-to-point communication to the Web server to request the HTML/XML data. While this solution works fine for low-interest information (i.e. where only few users are interested), for high-interest feeds (i.e. information that are "followed" by many users in real-time) this solution introduces a significantly high, and above all unnecessarily duplicate load on the mobile network and the Web servers.

1.1.3 Wireless Heterogeneous access

In today's wireless networking domain, diverse wireless technologies are utilized for sharing data and providing data services. Among the available technologies, the leading example is the widely-deployed 3GPP cellular networks (including the 3G and LTE) and WLANs. This opens the opportunity for: (i) network operator to offload part of the 3G/4G traffic through the WLAN; (ii) users to choose the best connection regarding some criteria (e.g. bandwidth, cost of communication, security, etc.). However, the selection of the network access has to consider attributes and criteria defined by the operator as well as the user. Criteria defined by the user aims to maximize its QoS (data rate, security, delays) and minimize communication cost. Those defined by the network operator are mainly related to the network resources optimization (e.g. load balancing between wireless networks). There are many works that addressed the problem of network selection in heterogeneous wireless networks. These solutions are based either on mathematical techniques (e.g., stochastic programming to model the random aspect of user connection) or on multi-criteria optimization techniques. Usually, the network selection mechanism is called when the mobile station enters into a zone covered by different wireless access technologies.
There are also standardisations activities that facilitate the handover procedure when user is roaming among heterogeneous wireless. IEEE 802.21 group has defined the Media Independent Handover (MIH) [8] in order to allow mobile terminals to select the wireless network in a transparent way to the user by avoiding service disconnection. MIH specifications are complementary procedures to the above-mentioned network selection mechanism.

1.1.4 Hierarchical video coding

Independently from the network access technology, it is possible to increase the QoS support for real-time applications and especially for video services by taking in consideration the user context when creating the content. In fact, user context (e.g. used terminal, network access technology, and used bandwidth) gives interesting information, which can help to adapt the content format to the undelrying environment. For example, there is no need to send a High Definition (HD) video to a user connected from a smartphone or a low-resolution terminal. But, encoding the same content in different formats is costly in terms of data storage. Thus, one solution is to use hierarchical coding such as the Scalable Video Coding (SVC) introduced in the H.264 standards [9]. Scalability in SVC is achieved by taking advantage of the layered approach already known from former video coding techniques. Three fundamental types of scalability could be used in SVC, namely spatial, temporal, and quality. Usually, a SVC stream includes one base layer and one or several enhancement layers. The removal of an enhancement layer still leads to reasonable quality of the decoded video.
By employing SVC, it is possible to constitute a set of layer combinations to create the video streams, which allows targeting several spatial, temporal and quality scales, depending on the users' context. Furthermore, SVC optimizes the network resources as it avoids the usage of additional bandwidth for sending useless content for the receiver.
It is worth noting that there are two ways in SVC to adapt to a user environment: (i) at the server level, by selecting the number of SVC layers to send to the receiver; (ii) at an intermediate node in the network (also known as Media Aware Network Element or MANE [10]), which withdraws useless layers. But, the main challenge associated to this adaptation is how to map (define) the number of SVC layers to a user context.

1.2 Contributions

In order to address the challenges described above, the novelty of our contributions are in proposing dynamic and adaptive mechanisms to control network functionalities. The adaptive approaches are useful to tackle issues related to: (i) the network dynamics and changes in terms of congestion and contention level; (ii) the wireless channel dynamics and changes in terms of signal quality; (iii) the application dynamics and changes in terms of generated traffic; (iv) the users experience dynamics and changes in terms of QoE degradation/enhancement. Generally speaking, the conducted research work considers the network (or an application in the network) and the network functionality to control as a closed loop system. The objective is to modify (control) the system inputs (e.g., accept or reject a new flow in the network, increase the sleep duration for a wireless station, control congestion, etc.) and monitor the system outputs (e.g., delays, QoE, Queue size, etc.) in order to use these outputs for the next decision epoch of the controlled network mechanism. The monitoring processes could be implemented at different levels, namely network, user terminal or application. The principle of this controlled system is depicted in Figure 1.1.
Similar to the monitoring process, the mechanism or function used to control the network could be located at different levels: network node (e.g. router, AP, BS, etc.), a user terminal (e.g. MAC, network or transport layer) or at the application. The control mechanism has a direct impact on the network (or an application in the network). For instance: (i) accepting a new flow in the network may reduce the available resources or increase the network load; (ii) increasing the sleep duration for a station may reduce QoE perceived at the application level. The control mechanism does not influence the network only when it

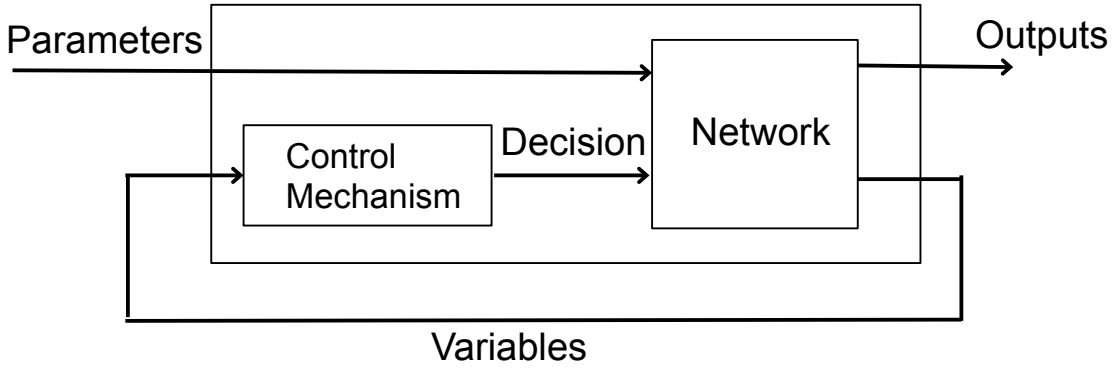

Figure 1.1: Principle of our contributions.

is used in unidirectional-based network such as DVB. In order to control the network, it is necessary to take measures (monitoring process) of the network or the application variable (system outputs) in a real-time fashion. These variables will be used as input of the network control mechanism. Note that these variables could represent the network performance (e.g. traffic load, the Queue size of a router, etc.), or the application performance (e.g. data rate, delays, QoE, etc.). Our contributions are classified according to the type of monitored variable (system output) used by the network control mechanism. The first class of our research work pertains to QoS/Network centric (e.g. delays, the network load, the queue size of a MME, delays, data rate) information, while the second class is based on human centric information by monitoring users' QoE. From hereunder, we refer to the first class as network-centric contributions, and to the second class as human-centric contributions. To the best of our knowledge, human-centric contributions are the first that put user at the hear of the of network control mechanism. The only work that considered such information is the one dedicated to the adaptation of VoIP [11] data rate over Real Time Protocol (RTP), whereby the QoE information is sent back to the receiver through Real Time Control Protocol (RTCP) packets.
The rest of this document is organized in the following fashion. Chapter 2 presents the network-centric contributions; particularly focusing on two representative works: admission control in WLAN and congestion control in LTE. In Chapter 3, we introduce research work belonging to the human-centric class, where the focus is on three representative contributions: QoE metric, QoE-based in-network adaptation of SVC flows in DVB-T2 and multicast optimization in WLAN. Concluding remarks and future research perspectives are provided in Chapter 4.

Bibliography

[1] 3GPP TS 23.401, "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access"

[2] IEEE, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, IEEE Standard 802.11, June 1999.

[3] IEEE 802.11e Standard – Part 11 Medium Access Control (MAC) Quality of Service (QoS) Enhancements, 2005.

[4] IEEE 802.11b: Enhancements to 802.11 to support 5.5 and 11 Mbit/s (1999)

[5] IEEE P802.11aa/D6.0 Draft Standard – Amendment 3 : MAC Enhancements for Robust Audio Video Streaming, 2011.

[6] 3GPP TS 23.246, "Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description".

[7] Bijin Tokei Application, URL: http://www.bijint.com/en.

[8] The Institute of Electrical and Electronics Engineers (IEEE). Media Independent Handover Services IEEE Standard 802.21, 2007. Rev.4.0.

[9] Scalable Video Coding, Joint ITU-T Rec. H.264 – ISO/IEC 14496-10 / Amd. 3 Scalable Video Coding, Nov. 2007

[10] S. Wenger, Y. Wang and T. Schierl, "RTP Payload Format for SVCVideo", IETF Internet Draft, draft-ietf-avt-rtp-svc-20.txt, Dec 2009.

[11] I.-H. Mkwawa, E. Jammeh, Lingfen Sun, E. Ifeachor, "Feedback-Free Early VoIP Quality Adaptation Scheme in Next Generation Networks", in proc. IEEE Globecom 2010, Miami, USA.

Chapter 2

Network-centric contributions

Contents

2.1 Introduction

In the network-centric contributions, the network control mechanism is based on the information measured from the network state and/or the QoS performance at the application level. These contributions allow adapting to the network or application changes by monitoring in real-time the network or application state. This kind of solutions was used to address the following issues:

- Admission control in WLAN ([1][2]) (Details in section 2.2)
- WLAN operating in a noisy channel ([3]):
  In 802.11-based wireless networks, a packet error means transmission failures between a pair of wireless stations due to: (i) collision with other packets; (ii) Bit Error Rate (BER)-corrupted packet. When detecting an erroneous packet, receiver station must automatically reject this packet, and no ACK is transmitted. Accordingly, the sender station assumes that packet loss is an effect of collision and takes measures to avoid further collision in the network by delaying the retransmission of this packet (i.e. increase its Contention Window - CW). This is obviously sub-optimal in case of BER corrupted packets: contention window should not be increased to avoid collisions when loss is due to noise, so it is important for the sender to differentiate between the origins of the lost packet.
  To tackle this issue we introduced a new mechanism that allows wireless station to differentiate between collision and BER loss through adapting the RTS/CTS handshake mechanism to noisy channel environments. The main contribution was to

respond to a packet loss following the RTS/CTS exchange by invoking the retransmission routine; however instead of increasing the CW, we proposed to maintain its current value.

- VoIP over WLAN (VoWLAN) ([4]):
  In this contribution we optimized the support of VoWLAN service, by proposing a cross-layer mechanism that adapts the voice coder's rate according to the MAC layer feedbacks, which reflects the network state in term of network load and physical rate. Indeed, we proposed to use different voice coders at the application layer; one high quality coder such as G.711 and one medium quality coder such as G.728 or G.729. The application layer is allowed then to switch between these two voice coders according to the triggers announced by the MAC layer. That is, if the network is congested and/or the physical rate is reduced (until a certain threshold) then the G.729 coder is used, guarantying thus an acceptable VoIP quality while reducing the network overload. If the network is in a relax situation and the physical rate is using higher bit rate, then the G.711 coder is used, ensuring thus a high VoIP service quality.

- Congestion control in LTE in case of
  - MTC traffic ([5][6]) (Details in section 2.3)
  - social network traffic ([7]):
    The proposed contribution is based on content detection systems such as Deep Packet Inspection (DPI) to identify traffic belonging to a group of users (sharing the same content) of a social network. Upon detecting the type of traffic, we proposed to control it by creating a multicast group. This would reduce the amount of traffic exchanged by switching from unicast communications to multicast communications.
    Another solution is to cache, at the geographically nearest base station, the shared content among users. Here we positioned ourselves in the case where the social network traffic comes from the same geographical region.

In this chapter, we will concentrate further on two representative works belonging to this class of contributions: (i) admission control in WLAN; (ii) congestion control in LTE: the case of MTC.
As stated in the Introduction chapter, admission control is crucial in 802.11-based WLAN. It allows regulating the wireless traffic by maximizing network resources utilization while ensuring acceptable QoS for admitted flows. Our contribution is two-fold: (i) a new 802.11 MAC protocol that derives the CW value according to the network state (contention level) and the application tolerated delay; (ii) based on this new protocol, a distributed admission control was proposed.
The second contribution presented in this chapter is about supporting MTC in LTE, and particularly controlling the congestion that can occur in this case. We proposed a congestion-aware admission control solution, which selectively rejects signaling messages from MTC devices at the RAN following a probability that is set based on a Proportional Integrative Derivative (PID) [27] controller reflecting the congestion level of a relevant EPC node.

2.2 Admission Control in IEEE 802.11

2.2.1 Research context and related work

In its basic form, the IEEE 802.11 Distributed Coordination Function (DCF) provides a simple and flexible mechanism for sharing the medium, but lacks the ability to guarantee service levels to meet the demands of multimedia applications. As a consequence, there has been a considerable effort to improve the MAC's ability to serve and interact with higher level QoS mechanisms. IEEE's 802.11e Task Group has worked towards designing and developing a framework for QoS support. Based on the basic DCF, the 802.11e proposals focus primarily on providing differentiated access to individual traffic classes (TCs). In particular, the Enhanced Distributed Control Access (EDCA) uses priority concepts to alter the existing MAC scheme. During initialization, EDCA assigns static MAC parameters for each TC. Based on these parameters, the MAC protocol provides different service levels to different TCs. It is readily realized that EDCA parameters do not accommodate all network configurations in terms of relative (per-class) network load [8]. Particularly, EDCA is unable to absorb a high number of multimedia flows due to a too narrow backoff range (0, 31) assigned to high-priority flows, which lead to high intra-TC contention level. This situation entails high collision rate, poor medium utilization, and increased medium access delays.
Furthermore, there is a key trade-off between fully filling the network capacity and maintaing acceptable QoS, even when using the 802.11e specifications. Admission control is mandatory to achieve this goal in 802.11. Indeed, if there are no restrictions to limit the volume of traffic being introduced to the service set, performance degradation will result due to higher backoff time and collision rate.
There are several works that addressed the admission control issue in WLAN. We can classify them into two categories:

- Model-based admission control, where performance metrics are constructed to estimate the network status,

- Measurement-based admission control, where the admission control decisions are based on the continuous measured network conditions.

Model-based mechanisms are mainly derived from the analytical models introduced in ([9][10]), which evaluated the 802.11 performance in terms of throughput and delay. Based on the throughput model developed in [9], work in [11] predicts the achievable throughput of each flow. If the predicted throughput satisfies the new flow's need, it is accepted. Otherwise, it is rejected. However, this admission control mechanism is not realistic, as it is based on analytical results derived in saturation condition (i.e. each station always has packets to transmit). In [12], the authors use an analytical model to estimate an average delay for the traffic of different priorities in the unsaturated 802.11e WLAN. Based on the delay criterion the authors propose an admission control mechanism. This latter considers the effect of admitting a new real-time flow on the channel utilization and the delay experienced by existing real-time flows, ensuring that the channel is not overloaded and the delay requirements are not violated. The authors affirm that the used analytical model overestimates the access delay and hence the decisions are made with a certain security margin. However, this overestimation results in rejecting new flows even if the network can satisfy their requirement in term of delay.
Measurement-based mechanisms, on the other hand, represent most works on admission control in WLAN. This is due to the flexibility and low computation requirements of these

approaches. In what follows, we will details three representative solutions in this category. Here, we focus more on measurement-based mechanisms as our proposed admission control belongs to this category.

Based on local network measurements, authors in [13] propose to control the arrival rate at each station to achieve a given objective such as, maximum throughput, maximum delay, jitter or loss rate in the network. The developed analytical model is able to assess the capability of the 802.11 for supporting major QoS metrics. The model is further extended in [14] to control the admission of network flows based on a new metric (channel busyness ratio) as a good indicator of the network state; channel busyness ratio is used to derive a rate control algorithm (CARC - Call Admission and Rate Control). Besides not being applicable to 802.11e-like protocol where several traffic classes (having several requirements) may simultaneously operate in the network and even coexist at a single station, CARC tries to find the optimal network utilization (maximize the throughput), while barely considering delay fluctuations.

Distributed Admission Control (DAC) and Two-level Protection and Guarantee Mechanisms [15][16] are combined to propose an efficient admission control mechanism for 802.11e-based WLAN. DAC is a measurement-based admission control mechanism that was considered by the 802.11e working group. In this algorithm, the resource budget for each TC is periodically announced by the AP in the beacon frame, so that each station may decide whether to accept or not new flows. A new stream to be admitted tries first to access the network and it rejects itself after a certain period if its requirements are not met. With this algorithm, the residual network resources are fairly distributed among the competing streams (streams seeking for acceptance) at different stations in the sense that different TC's (in different stations) compete to accommodate their new entering streams; the stream is then locally accepted if it reaches its targeted throughput. This situation may cause spectrum waste because there may be enough resources to admit one additional stream, but due to the algorithm fairness and absence of coordination none of the competing streams is accepted and the available bandwidth remains unfilled.

Another shortcoming of DAC algorithm resides in the lack of protection to existing flows only when the network load is not too heavy. If the network resources are not sufficient to admit the new stream, the performance degradation will affect all TC's streams (as much as it does for other TCs active in the network). This is due to the fact that entering streams are aggregated with other active streams in the same TC queue. The above mentioned phenomenon is usually referred to as "spill over" effect in WLAN – when traffic is overloaded in a TC, performance in other TCs will also be affected. Still, the major problem with DAC-based approaches consists in the fact that the overall network bandwidth is statically allocated among different TCs, so each TC receives a fixed share of bandwidth that cannot be exceeded. This may severely affect the flexibility of the admission control mechanism since it is very difficult to beforehand forecast the per-TC traffic volume in realistic multimedia-dedicated WLANs. Therefore, streams from a given TC may be rejected while some bandwidth still unfilled in other TCs, which means bandwidth wasting or additional revenue loss for network operator. Another side effect is that the admission decision depends only on local measurements collected at the admitting station level. However, the stream admission may have different impacts at different stations (resp. flows) depending on the load of each active station. The stream admission may actually cause QoS violation at certain stations while not effecting at all other active stations in the network; this is particularly prevalent for high-bit-rate stations, which usually cannot carry the load in a sufficiently timely manner as the load (resp. medium access delay) increases.

Virtual MAC and Virtual Source Algorithms [17][18] propose a fully distributed VMAC

(Virtual MAC) algorithm that operates in parallel to the real MAC in the mobile host but the VMAC does not handle real packets; rather, it handles "virtual packets". Each station runs a VMAC instance that monitors the capability of the wireless channel and passively estimates whether the channel can support new service demands (e.g., delay and loss). Unlike the case of real packets, VMAC doesn't transmit anything but estimates the probability of collision. When a collision is "detected", the VMAC enters a backoff procedure, just as a real MAC would do. The virtual source (VS) algorithm consists of a virtual application; an interface queue, and the VMAC. The virtual application generates virtual packets like a real application. Packets are time-stamped and placed in a virtual buffer. After a virtual packet has been processed in the VMAC, the total delay is calculated.
VMAC's main criterion to make an admission control decision is based only on delay and collision estimates. It does not provide any achievable throughout information, which is also useful to multimedia applications. The achievable QoS is estimated only at the admitting station, although flow admission may unevenly affect the different backlogged flows, provoking delay violation at certain flows while other flows in the network still experience acceptable delays. As mentioned earlier, the outcome of stream admission should be beforehand assessed at all active stations. In fact, flows belonging to the same TC use roughly the same CWs' ranges, and thus they more or less experience the same packet-service times (i.e., the time needed to successfully transmit the frame located at the front of the queue). Hence, depending on the volume of their offered load, different flows may suffer from widely different enqueuing delays. In other words, admission of a new flow means a slightly increased packet-service time with different outcomes on different active flows. The impact of a stream admission should be therefore assessed at all active stations.

2.2.2 Contribution

We proposed a per-flow based admission control, where the flow's constraints in term of QoS are taken in consideration for the admission decisions. Of course, the proposed admission control considers also the impact of a stream admission on the network state in order to protect the admitted flows' QoS. However, to implement this admission control mechanism, there is a need to translate the flow's requirement into MAC parameters. Actually, the current 802.11e MAC protocol is not able to ensure such per-flow differentiation. In fact, the 802.11e specifications consider the same MAC parameters for two flows belonging to the same TC, even they have different constraints. To address this issue, we proposed a new MAC 802.11 protocol featuring delay-driven CW adjustment. By analyzing all factors that influence the medium access delay, we derived a distributed model able to accurately predict the achievable delay at each network flow using different network measurements. Delays bounds associated to each traffic class are assumed to be communicated to MAC layer through a top-down cross-layer interaction. Based on the latter model, we derived an admission control algorithm that allows to "a-priori" assess the achievable throughput before admitting new incoming streams taking into considerations their QoS requirements. This could contribute in improving network utilization; the objective being to preserve the QoS of already active flows while maximizing the volume of QoS-enabled services, providing to network operators an improved resource control mechanism (i.e., allows to generate more revenues).

Figure 2.1 represents the global overview of the proposed admission control, and the interaction with the new 802.11 MAC protocol.

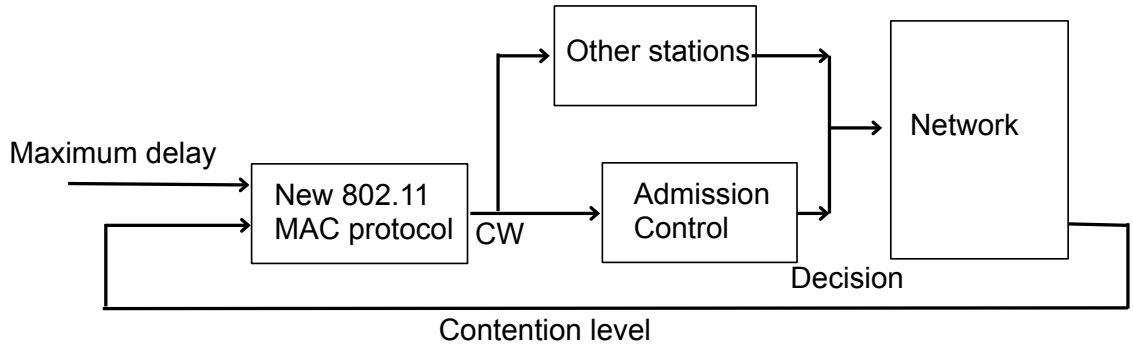

Figure 2.1: Overview of the proposed admission control mechanism.

New IEEE 802.11 MAC protocol featuring Delay-driven CW adjustment

Conventional IEEE 802.11 backoff schemes have many shortcomings that make it difficult to provide deterministic guarantees. The exponential CW increasing is more likely to produce probabilistic service assurances and high oscillations in delays (throughput) since the CW is reinitialized to its minimum value (CW_{min}) after each successful transmission. In order to limit the effect of high inter-TC contention, different Arbitrary Inter Frame Space ($AIFS[i]s$) may be assigned to different traffic classes $TC[i]$; this would defer transmissions of low-priority flows only when their respective transmission attempts coincide with high priority flow transmission. At this point, managing the contending flows through appropriate CW scheme is a key component to effectively maintain acceptable QoS level for multimedia flows.

At MAC layer, packets are serviced with a variable latency that depends on the current CW size, the mean frame size ($E[P]$), and the mean number of transmission attempts before effectively gaining access to the medium. Besides, the network load (i.e., transmission volume from other nodes) may strongly affect the end-to-end communication latency as a substantial amount of time slots is occupied, which ends up provoking frequent backoff freezing. Actually, each new packet selects a random backoff interval ($E[CW]$) that is more or less quickly decremented depending on the number of time slots where the medium was observed as busy. The packet transmission deferring period depends on the selected backoff interval as much as it does depend on the degree of network load.

We define PST (Packet Service Time) as the time needed to successfully transmit a packet; this delay is defined as the time interval elapsed between the time when a packet arrives at the front of the queue and the time when it is received by the receiver. The delay considers only channel access delay, transmission delay, and associated overhead (i.e., queuing delay is not included).

Let $B(T) = \frac{B}{I}$ be the number (B) of busy time slots over the number (I) of idle slots observed during the last T time slots ($T = B + I$). The total deferring time for a packet can be approximated by $E(CW) * (1 + \frac{B}{I})$; this delay takes into account both the backoff interval and the freezing period. Compared to the technique that achieves direct measurement of the freezing period at each flow [10], our technique is based on continuous monitoring of the overall network load, which could be better exploited to predict network load trends. Measuring the freezing period for each transmitted packet may exhibit high oscillations, not to mention the involved complexity. Using the overall network occupancy ($\frac{B}{I}$) to estimate the access delay leads to inherent measurements coordination between

different active flows as they observe the same network activities.
We define $E[P]$ as the mean number of time slots occupied by a single packet transmission including PHY/MAC overhead, Short Inter Frame Space (SIFS), and Acknowledgment (ACK) when considering the DCF basic mode. It is worth mentioning that within DCF basic method (without Request To Send (RTS)/ Clear To Send (CTS) handshaking), each failing transmission (due to frame collision or bit alteration) occupies roughly the same number of slots as a successful transmission [8]. In the following we assume a DCF MAC protocol operating without RTS/CTS handshaking, and that packet loss provoked by wireless link interferences (BER) is negligible. The overall packet service time (PST) may be quantitatively estimated as follows

$$PST = [E(CW) * (1 + B(T)) + E(P)] * E[TransAtt] \tag{2.1}$$

Here, $E[TransAtt]$ is the mean number of transmission attempts needed to successfully access the medium; this parameter depends on the PER (Packet Error Rate) and the automatic retransmission (ARQ) scheme being used at MAC layer. Generally speaking, a packet is kept in the transmitter queue until either a timer times out (i.e., after 7 failed transmission attempts), or the packet is successfully received and acknowledged by the receiver. Since the backoff process have a geometric distribution with probability of success p, the mean number of transmission attempts $E[TransAtt]$ would be $1/p$. At this point, the probability of transmission success, p, can be approximated as the fraction of the number of transmitted frames over the number of transmission attempts. Thus, the mean number of transmission attempts $E[TransAtt]$ can be estimated as

$$E[TransAtt] = \frac{1}{1 - \frac{Collisionbs}{TaransmissAttempts}} = \frac{TransAttempts}{SucceedTransmissions} \tag{2.2}$$

Note that $E[TransAtt]$ may return different values depending on the flow's traffic class and its associated AIFS. Obviously, inter-TC collisions are most of the time avoided since flows with the highest priority seize the medium while other flows enter in differing state. As $B(T)$ is calculated based on the overall network load, it is inherently coordinated between stations. Each station averages the measurements over the period T required to sense "CW_{max}" idle time slots. By choosing the frequency of measuring $\frac{B}{I}$ in this way, we are ensured that all backlogged flows (regardless of priority) would have attempted to access the medium at least once within this period. Thus, $\frac{B}{I}$ measurement is more accurate by considering all active flows, and also more stable as they are averaged over a long-enough period. Thus, the value of T is set to 1024 "*idle*" slots. For the same reasons, $E[TransAtt]$ values are also averaged over the period needed to sense 1024 idle time slots. As apparent from Figure 2.2, each station in the network may have different traffic classes with different requirements in terms of QoS metrics performances. Several MAC queues are indeed implemented within a single station. Each queue supports one TC, behaving similar to a single DCF entity within the 802.11 standard. In this context, the last packet in the queue (packet $\#N$) should not exceed the maximum delay tolerated by the traffic class (TC) to which it belongs. By considering that both the arrivals (λ) and the service (μ) are exponential, the PST will be therefore constrained by

$$PST \leq \frac{MaxDelay}{N} \tag{2.3}$$

The formula above generalize our PST estimation model to estimate the enqueuing time by taking into account the number of packets (N) currently in the MAC queue (the N packets ahead of the last packet entering the queue). From the formulas (2.1) and (2.3),

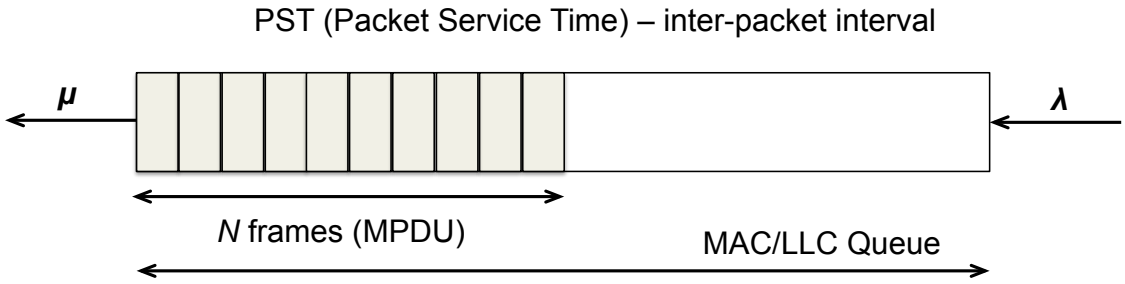

Figure 2.2: MAC layer queue for TC_i.

and given the queue length (N), the appropriate maximum CW size (CW_{max}) that would satisfy the delays constraints associated with each service class (regardless its bit-rate) is obtained as follows

$$CW_{max} = 2 * E[CW] \leq \frac{2 * (MaxDelay - N * E[TransAtt] * E[P])}{N * (1 + B(T)) * E[TranAtt]} \tag{2.4}$$

It is commonly accepted [10] that WLAN capacity (i.e., channel utilization) decreases with an increasing number (M) of active flows. This is caused by high contention level in which case the medium is often occupied by collisions. In this situation, the mean number of attempts to successfully transmit a frame would grow resulting in additional delays at active flows. The contention window size (CW) should be continuously adapted, thereby reacting to changing network conditions while meeting QoS constraints. Actually, when M increases, the CW size is increased to absorb the increasing number of contending flows, and hence minimizing the collision probability for these flows. On the other hand, when M becomes small, the CW size is decreased, which reduces the spacing between successive frame transmissions; large values of CW size may indeed strongly limit the throughput of fewer backlogged flows. As a matter of fact, the current CW_{size} in use should be always larger than a certain variable threshold (CW_{min}) to avoid network performance collapse. From [19], and [20], the minimum CW_{size} that maximizes network performances with M contending flows is given by

$$CW_{min} \geq \lfloor M * \sqrt{2T_c} \rfloor \tag{2.5}$$

$$M = \frac{E[O(T)] * (E[oldCW] + 1)}{T} \tag{2.6}$$

Here, T_c is the average time (in time slots) of channel unavailability upon a collision. T_c is dependent on the physical layer, and is equal to $PHYhdr + E[F] + DIFS$ when RTS/CTS mode is disabled. $E[oldCW]$ is the current mean backoff value. $O(T)$ is the number of slots where the medium was observed as busy out of the previous T slots (B). Like all other network measured parameters (i.e., $E[TransAtt]$ and $B(T)$), $O(T)$ is weighted in respect to past measures using EWMA (Exponential Weighted Moving Average).
Although not accurate (i.e., much incertitude still exists due to different flows' priorities and bit rates), the estimate of the number (M) of active flows is quite pertinent since it still precisely reflects the overall trends of the network contention level, which allow readjusting the CW to optimize the network performance. In fact, constraining the contention window by CW_{min} helps to keep a low collision rate, and hence an acceptable mean transmission attempts (i.e., $E[TransAtt]$) lower than 1.5, which means 3 transmission attempts for 2

successful transmissions). The new CW to be maintained by each TC is given by

$$newCW_{size} = \frac{CW_{min} + CW_{max}}{2} \tag{2.7}$$

with $CW_{max} \geq CW_{min}$
If CW_{max} is smaller than CW_{min}, we assign CW_{min} to CW_{max}. In this case $newCWsize$ is simply reinitialized with CW_{min} value. This situation does not guarantee $MaxDelay$; instead, it keeps network collisions within an acceptable level. Using the above introduced CW_{size} adjustment model, a given flow would use the interval $[0, newCWsize]$ to randomly draw a backoff interval. Note that the parameter CW_{min} is not necessarily coordinated between flows since its value is, in part, based on current CW size that is maintained by the flow. Accordingly, flows calculate different CW_{min} values depending on their class of service ($MaxDelay$ constraints) and their offered load as well.

Multimedia services admission and protection

Since delay estimation is based on inter-packet interval assessment, the achievable throughput together with potential degradations (mean loss rate) may be predictable as well. Using the packet arrival rate (λ), which is a-priory known for a given traffic class (TC), it is possible to capture the queue dynamics based on instantaneous network activities; the packet arrival rate may be for example provided by pre-established Service Level Agreements (SLA). The objective is to predict the impact of new stream's acceptance on the overall network performance. In other words, we assess the consequences resulting from increasing the arrival rate of a given TC/station (i.e., stream admission) before actually admitting any new entering service. As illustrated in Figure 2.2, we consider a MAC queue with a buffer size k. Service is exponential with parameter μ and inter-arrival times are exponential with parameter λ. A loss occurs whenever an arriving packet finds the queue full. The queue occupation rate is thus

$$\rho = \frac{\lambda}{\mu} = \lambda * E[PST] \tag{2.8}$$

The queue model is assumed to be a single-server queue with finite waiting room (M/M/1/K). Certainly, the Poisson assumption for the arrivals of packets is not the most realistic, but considering the exponential case reveals essential features of the system and is a fairly appropriate assumption for an aggregate of different streams (TC). The mean loss rate (L_r) of an $M/M/1/K$ queue is given by

$$Lr_i = \frac{(1-\rho)\rho^k}{1-\rho^{k+1}} \tag{2.9}$$

Since the maximum tolerated loss rate ($MaxDrop = L_r$) is a-priori known for each TC i, we can numerically fix ρ since the MAC queue size (K) is as well known. In fact, the network operator may propose different levels of QoS guarantees, where each level is characterized by maximum QoS metrics performances bounds (MaxDelay and MaxLoss). Table 2.1 illustrates an example of traffic classes when using Diffserv classes mapping.
For instance, assuming a queue length of $k = 30$ packets and with a maximum tolerated loss rate of MaxDrop=1%, the queue occupation rate ρ should be lower than 0.935. In the same manner, $\rho = 0.97$ for a maximum tolerated loss rate of MaxDrop=2%. In this contribution, we aim to categorize the traffic into service classes where each service class has a maximum delay and a maximum loss rate to not violate.

Table 2.1: A Cross-Layer QoS Mapping.

	Conversational services	Streaming services	Best Effort services
Type of application	Interactive voice and video gaming	Streaming audio/video, Multimedia broadcasting	Web browsing, E-mail, Telnet
IP Diffserv class	EF	AF11	Best effort
IP transmission Delay	600 ms	800 ms	Unspecified
Loss Percentage Resiliency (a_i)	1%	2%	Unspecified

Based on the delay analysis (i.e., PST) and the mean tolerated loss rate, we can now determine the appropriate μ (i.e., $\frac{1}{E[PST]}$) that satisfies the relation (2.9). Thus, we analytically figure out the appropriate CW that provides a mean inter-packet transmission interval ($E[PST]$) necessary to maintain a queue occupation rate at the desired level (ρ). By combining formula (2.1) and formula (2.8), we obtain the appropriate contention window size that satisfies the loss requirements associated to a given TC

$$NewCW_{size} = 2 * E[CW] = 2 * \frac{\frac{\rho}{\lambda} - E[P] * E[TransAtt]}{(1 + B(T)) * E[TransAtt]} \tag{2.10}$$

with $CW_{min} \geq newCW_{size}$ and $newCW_{size} \leq CW_{max}$
While the contention window (CW_{size}) given by formula (2.7) ensures an acceptable delay with regards to TC's requirements, formula (2.10) allows to avoid TC's queue overflow by each time checking if the current PST (i.e., $NewCW_{size}$) is able to absorb the packet arrival rate (λ). More precisely, the new CW size ensures that the TC's flow in which the entering stream will be aggregated will not violate its maximum tolerated loss rate. The new calculated CW size ($NewCW_{size}$) should be also larger than CW_{min}. This means that the network is able to accommodate the new stream's offered load while still meeting delays guarantees ($NewCW_{size} \leq CW_{max}$) and keeping an acceptable contention level ($NewCW_{size} \geq CW_{min}$) to avoid network performances collapse.
Combined with the delay-driven CW adjustment introduced in formula (2.7), the above formula may be used to accept new streams in the network. This consists in assessing if a new stream may be serviced while not interfering with already active flows. As highlighted already, an over-admission will unavoidably affect all currently serviced flows as the medium is shared and an increasing in the contention level affects all flows regardless their bitrates or priorities. On the other hand, different active flows may simultaneously maintain widely different CW sizes due to different values of CW_{min} and CW_{max}. The maintained CW contention window depends, actually, as much on the flow's offered load

as it does on the flow's traffic class. In certain circumstances, an over-admission may cause certain flow to violate its CW_{min} limit, while other flows still use CW sizes larger than their calculated CW_{min}; flows with high bitrates are generally the first flows to reach their CW_{min} limits. At this point, it is readily realized that the impact of new stream admission should be estimated at all stations.
At new stream admission, each flow in the network recalculates the values of CW_{min}, CW_{max}., and $NewCW_{size}$ according to formulas (2.4), (2.5), and (2.10). The new values of these parameters should take into account changes in network availability entailed by admitting a new stream. Accordingly, certain determinant measurement-based parameters such as $B(T)$, $O(T)$, and $E[TransAtt]$ should be reconsidered. While $E[TransAtt]$ fluctuations are limited by using an appropriate CW_{min}, both $B(T)$and $O(T)$ exhibit significant changes that should be considered to accurately re-estimating the new achievable QoS performances. Again, it is worth mentioning, that λ is actually the arrival rate of streams' aggregate belonging to the same TC. At new stream admission, the overall arrival rate at the TC's queue would increase as follows $\bar{\lambda} = \lambda + \Delta\lambda$, where $\Delta\lambda$ is the packet arrival rate of the new entering stream. In this case, the network load should be updated to reflect the additional load induced by the new stream.

$$\bar{B}(T) = \frac{\bar{B}}{I} = \frac{B + \beta}{I - \beta} \tag{2.11}$$

with $\beta = \bar{\lambda} * T * (20 * 10^{-6}) * L$
Here, L is the mean number of time slots occupied by a MAC packet of a given flow, including the overhead involved by acknowledgement. $O(T)$ should be as well updated with the new flow arrival as follows

$$\bar{O}(T) = \frac{\bar{B}}{T} = \frac{B + \beta}{T} \tag{2.12}$$

Given the above-mentioned parameters, all active stations calculate the new values of $CW[i]_{min}$, $CW[i]_{max}$, and $NewCW[i]size$ for each TC i. If the new values satisfy all QoS constraints ($CW[i]_{min} < NewCW[i]_{size} < CW[i]_{max}$) associated to each TC_i, then the station concludes that the entering stream will not affect its already serviced streams. If all stations will not be affected by the entering stream, the AC algorithm may then proceed with stream admission. Otherwise, it means that the stream admission may severely degrade the quality of currently servicing flows, which should lead to rejection of the entering stream.
The first issue to tackle when designing a distributed AC mechanism is the coordination between competing nodes. In fact, besides necessitating a unified admission model for all stations, we further require to harmonize the estimation of achievable QoS at different station in order to achieve a coordinated admission control decision. Particularly, multiple new real-time streams may be simultaneously admitted by individual nodes if not coordinated, causing "over-admission". To mitigate this problem while keeping the distributed feature of our protocol, we divide the time into admission cycles (epochs) where only one single stream may be accepted in an admission cycle. The network is assumed to operate on "slotted" synchronization epochs, where each epoch is actually equal to a beacon period. This way, the admission cycle is long enough to allow network measurements ($E[TransAtt]$), at different stations, to converge towards accurate values reflecting the real network conditions before admitting new stream in the next synchronization epoch.
To completely avoid the over-admission problem, we adopt a coordinator-aided admission control scheme. In other words, all admission decisions are made by a coordinating node (CN), which can record the current number of admitted real-time flows and their occupied

channel bandwidth in the network; clearly, this will prevent over-admission situations. The coordinator node is also in charge of other responsibilities related to service level agreement (SLA).

It is important to note that a coordinator is available whether the WLAN is working in infrastructure or in ad hoc mode. If the network is working in the infrastructure mode, the access point is inherently the coordinator. Otherwise, a mobile node can be elected to act as the coordinator in the network using one of many algorithms in the literature (see [21], and references therein). A natural solution would be to appoint the node in charge of sending the MAC-level beacon as the CN. As in 802.11 ad hoc mode, in case of failure a distributed backoff-based mechanism would designate a new node to periodically send the beacon.

Each time a station has a new stream to admit, it should beforehand evaluate locally its impact using new values of $B(T)$ and $O(T)$ as given by formulas (2.11) and (2.12). Using formula (2.10), the station S should as well assess the risk of having overflow by calculating $NewCW_{size}$, where λ is replaced by $\lambda + \Delta\lambda$; in Figure 2.3, λ_i (i=1 to 3) stands for the rate ($\Delta\lambda$) of a new entering stream.

If the new entering stream doesn't affect the locally active TCs' flows, the station S

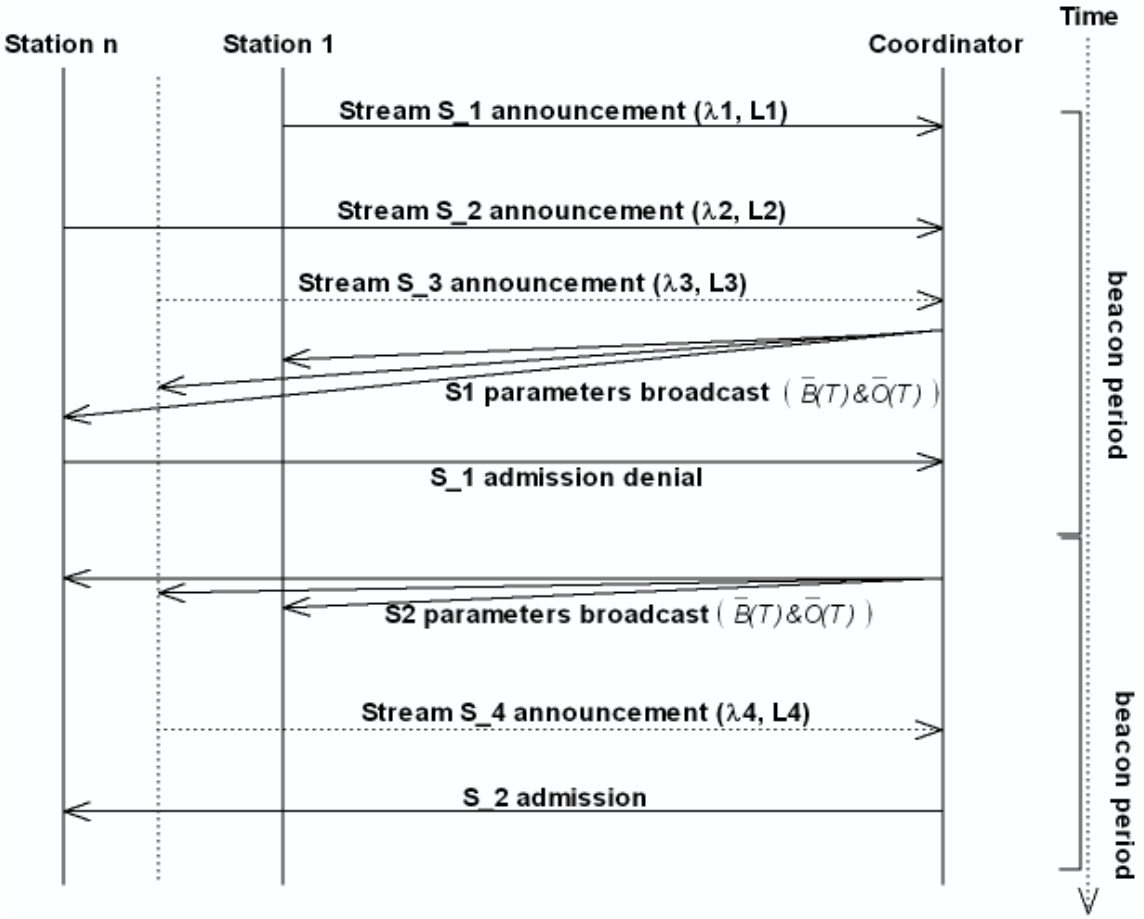

Figure 2.3: Admission control message exchange

announces the stream's bit-rate (λ) and nominal Maximum Service Data Unit (MSDU) size (in terms of time slots) to the CN which, in turn, recalculates the new values of net-

work occupancy parameters ($B(T)$ and $O(T)$) to be broadcasted. Then, all active stations evaluate the impact of new stream admission (i.e., with new $B(T)$ and $O(T)$ changes) on their TC's flows and eventually deny the admission if the QoS of one of its TCs may degrade. Note that each $TC[i]$'s flow in the network calculate $NewCW_{size}$ using its own packet-arrival rate (λ) and maximum queue-occupation-ratio ρ_i corresponding to its traffic class.

Figure 2.3 illustrates a scenario where in the first beacon period the coordinator receives 3 new streams announcements. The coordinator calculates and broadcasts parameters associated to the first stream (S_1). The admission is then aborted by station n the admission of S_1 interferes with its QoS constraints. In the second beacon period, the coordinator broadcasts S_2 parameters and finish by accepting the stream as no active station have denied the acceptation within the current beacon period. Typically, here S_2 should have a lower packet rate than S_1.

For scalability reasons, AC handshake messages are kept to a minimum by broadcasting CN messages (i.e., parameters broadcast and admission messages). Furthermore, response messages (i.e., admission denial message) are sent by an active station only if one of their QoS thresholds, associated to TCs' flows, would be violated with the new stream admission. A single denial message suffices to abort the whole stream admission process, so other stations don't need any more to send denial messages, i.e., all stations overhear AC messages.

To increase the reliability of CN's broadcasted messages, we use efficient basic data rate (1Mbps) usually employed to transmit the beacon, RTS/CTS, and ACK messages. On the other hand, during AC process, all directed messages exchanged between the coordinator node and other stations are fully persistent in the sense that they are retransmitted until successful reception.

Upon a first admission in a given beacon period, the other flows seeking admission in the network should defer the announcement to the next beacon period and additional network measurements are carried out before final admission. This allows all stations to take into account the changes in network availability before accepting new streams (i.e., allows the different competing stations to have a coherent perception of the network availability by carrying out measurements during a long-enough period such as a beacon period).

2.3 Congestion control in LTE: MTC case

2.3.1 Research context and related work

The 3GPP standard is currently working on the standardization process for supporting MTC communication in LTE [22]. The envisioned 3GPP architecture with MTC support is shown in Figure 2.4. It consists of three main domains, namely the MTC device domain, the communication network domain, and the MTC application domain. In the network domain, most important nodes of a 3GPP Evolved Packet System (EPS) network are shown. The MTC application domain consists of MTC servers, under the control of the mobile network operator or a MTC provider. Table 2.2 provides a brief description of the most important EPS nodes, shown in Figure 2.4. Two new entities related to MTC

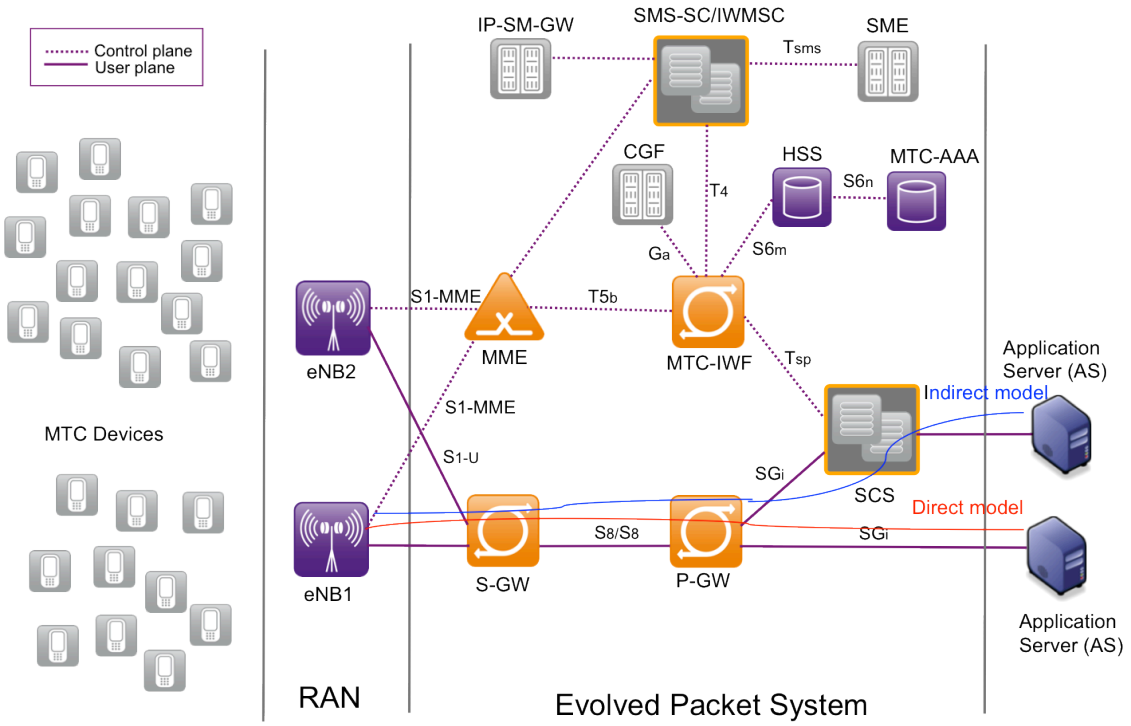

Figure 2.4: Architecture for MTC in 3GPP [22].

recently emerged in the 3GPP architecture. They are namely, MTC-IWF (InterWorking Function) and SCS (Services Capability Server). A MTC-IWF may be a standalone entity or a functional entity of another network element. The MTC-IWF hides the internal PLMN (Public Land Mobile Network) topology and relays or translates signaling protocols used over the Tsp interface to invoke specific functionality in the PLMN. SCS is an entity that connects to the 3GPP network to communicate with MTC devices and the MTC-IWF entity. As depicted in Figure 2.4, there are three ways of establishing connection between MTC servers and MTC devices. In the direct model, a MTC server connects directly to the 3GPP network and gathers data (through the user plane) from the MTC devices. Indirect model involves the services of SCS in order to use for example control plane device trigger. In this case, SCS is either controlled by the MTC provider or by the network operator. The final model is a hybrid model, whereby the MTC server can simultaneously use both direct and indirect models.

Table 2.2: EPS most important nodes.

Node	Description
eNB	Evolved Node B, the LTE's base station.
MME	Mobility Management Entity, a control plane entity for all mobility related functions, paging, authentication, bearer management in EPS.
MTC-IWF	MTC Interworking Function hides the internal Public Land Mobile Network (PLMN) topology and relays or translates signaling protocols used over Tsp to invoke specific functionality in the PLMN.
HSS	Home Subscriber Server, main database containing subscription-related information.
S-GW	Local mobility anchor for intra-3GPP handoffs.
P-GW	Packet Data Network Gateway, interfaces with the Packet Data Network (e.g., Internet).
SCS	Services Capability Server is the entity that connects MTC application domain to the network domain.

Whilst MTC represents an important business opportunity for mobile operators, mobile operators fear the congestion that could come with the deployment of billions/trillions of MTC devices, not to mention millions of smart mobile phones and their associated mobile traffic. Most signaling congestion avoidance and overload control mechanisms proposed in the context of MTC over cellular networks implement one of the following approaches: (i) segregate MTC traffic from the normal UE traffic in order to separate the network access for the two types; i.e., this helps to anticipate the congestion which may happen due to MTC traffic; (ii) when congestion occurs, apply some back-off mechanisms rejecting MTC traffic at the RAN equipment (eNodeB) or at the EPC nodes (e.g., MME, S-GW or even PDN-GW).

To differentiate MTC traffic from the classical traffic, most proposed solutions group the MTC devices into groups or clusters according to different metrics/features (e.g., low mobility, QoS requirement [23][24], belonging to macro or femtocell [25]). After grouping the MTC devices, there are two methods for separating the access to the RAN, and avoid the RACH (Radio Access CHannel) overload. The first one consists of defining "grant time periods" when MTC devices are authorized to connect to the network. The network also defines "forbidden time intervals" during which a MTC device is not allowed to connect to the network, be it the home network or a visited network. Intuitively, a grant time interval does not overlap with a forbidden time interval. Over the grant time, assigned to a MTC device, the communication window is further limited. The access time of MTC devices is also randomized over the communication window/grant time. In case of multiple MTC devices attempting to connect to the network during a specific and short communication window/grant time, to avoid signaling congestion and to cope with possible network overload during communication windows, the communication windows of the different MTC devices can be distributed over the grant time interval, via for example, randomization of the start times of the individual communication windows. This operation assists in reducing peaks in signaling and data traffic from MTC devices. Another way to define the duration of the grant interval is in the case where the network is aware of the period of time the MTC devices have to transmit. In fact, the network can dynamically increase the time grant duration dedicated to MTC devices if it is aware of the scheduling of MTC traffic. In some scenarios the network can predict when access load will surge due to MTC devices.

The second method consists in defining specific low levels parameters for separating Random Access CHannel (RACH) resources for MTC and non-MTC devices. The separation of RACH resources between MTC and non-MTC devices allows the limitation of the number of MTC devices capable to connect to the network, while maintaining normal network access for non-MTC traffic. To implement this separation, a simple way is to define a MTC specific backoff scheme. With this mechanism, the access attempts from MTC devices could be dispersed over a large time interval to prevent contending the RACH resources.

In addition to grouping MTC devices, there are also other solutions to anticipate the system overload by rejecting MTC device attach request if there are not sufficient network resources, or by grouping the signaling messages from a group of MTC in one common bulk signaling message. In [23][24] the authors first group the MTC devices into clusters according to their QoS characteristics and requirements, e.g., the cluster packet arrival rate and the maximum tolerable jitter by the MTC devices composing the cluster. When a MTC device attempts attaching to the network, it sends its QoS characteristics and requirements to the current eNB. If there are enough resources to satisfy the MTC requirements, the MTC device is accepted and added to an existing MTC cluster having the same constraints, or a new cluster is created. Otherwise, the attach request of the MTC device

is rejected. In [20], the authors also show the potential of handling signaling messages common to a group of MTC devices in bulk.
However, regrouping the MTC devices and separating their traffic from the other traffic at the RAN level is not always efficient to avoid congestion. In some situations, there is need to reduce the MTC traffic by a specific amount implementing admission control at eNodeBs or even at MTC devices. Indeed, admission control can be activated at the eNB upon receiving a congestion signal from the EPC nodes (MME/HSS). Or, it can be communicated to the MTC devices level as in the 3GPP Access Class Barring (ACB) solution. ACB is a solution, which effectively reduces the collision probability of transmitting the bulk of preambles at the same RACH resource. Based on the broadcasted parameters by eNodeBs, a UE determines whether it is temporarily barred from accessing the cell. An access class barring factor or access probability (p) determines the probability that access is allowed. If a random number n generated by the UE is equal to or greater than p, then access is barred for a mean access barring time duration. In the legacy ACB scheme, there are 16 access classes. AC 0-9 represent normal UEs, AC 10 represents an emergency call, and AC 11-15 represent specific high priority services, such as security services or public utilities (e.g., water/gas suppliers). A UE may be assigned one or more access classes depending on the particular cell access restriction scheme. Using network simulation, the work in [26] evaluated and compared the performance of ACB and the backoff procedure for reducing RACH contention. The findings of this work showed that using fixed ACB parameters for MTC devices and UEs is not optimal, as high values of p and access barring duration increase the access delay of MTC devices. In contrast, a low access probability p and a short barring duration increase the contention on RACH resources under heavy traffic load. Accordingly, the authors proposed using adaptive ACB parameters, by allowing eNodeBs to periodically adjust these values based on the system load. In case of high loads, higher values of p and access barring duration are assigned to the MTC devices, whereas under low loads these values are decreased. The obtained results clearly show that Adaptive ACB (A-ACB) can increase system performances when it operates under high load while reducing the MTC access latency in case of low loads. In addition, the work showed that A-ACB exhibits better performance compared to the Adaptive Backoff procedure and ACB with fixed values. The authors also point out that ACB parameters can be updated by sending SIB (system information block) that carries physical channel information (such as Random Access Channel information, Random Access parameters Hybrid Adaptive ReQuest (ARQ)). However, SIB broadcasting cycle is around a second to several minutes.

2.3.2 Contribution

Similar in spirit to the Access Class Barring solution, we proposed the Congestion Aware Admission Control (CAAC) scheme. In CAAC, MTC devices are grouped according to their priority classes. In case of congestion, each class is blocked with a probability p as in ACB. In contrary to ACB, where there is no indication on how this probability is computed, in CAAC, the reject probability p is derived for each class by the relevant EPC nodes (e.g., MME, S-GW and PDN-GW) being under congestion. Based on this probability, each eNodeB (at the RAN level) accepts/rejects MTC traffic belonging to a specific class.
In this contribution, we particularly focused on congestion that may happen at a MME. Such congestion can be directly induced from the delay in processing incoming packets at the application layer. This delay may lead to high and variable latencies at the input buffer, which may induce buffer overflow (i.e. packets loss). In contrast, congestion at

EPC nodes, particularly PDN-GWs, concerns the link layer, which may affect the outgoing buffer length. In CAAC, MMEs control their level of congestion by adjusting (through the reject probability p) the amount of incoming signaling traffic from the MTC devices. If the congestion is detected, the probability p is set to a high value. Otherwise, this probability remains low. This probability is then communicated, through dedicated signaling packets or incorporated in existing ones, to the relevant eNodeBs that use it as part of their admission control operation. To achieve the above-mentioned goal, CAAC uses PID controller, a well-known controller in the field of classical control theory [27]. Each MME implements an independent PID controller that uses only locally available information to derive the probability p in order to maintain the MME's queue around the optimal value that avoids both system overloads and underutilization, at the same time. It shall be noted that lots of research work in the literature have indicated that the queue length and the queue length variation represent a good indication to quantify the severity of congestion [28]; thus, our usage of queue length for congestion assessment in CAAC.

Figure 2.5 depicts the envisioned control system using a PID controller. The system in-

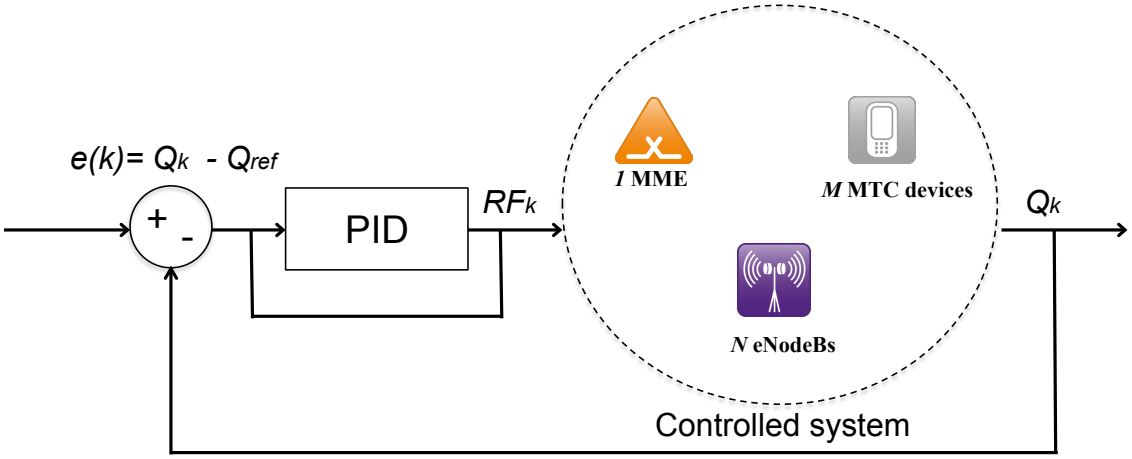

Figure 2.5: The controlled system.

volves all elements impacting its length, including MTC devices, eNodeBs attached to a MME and the MME itself. The PID controller takes as input the error signal $e(k)$ that represents the difference between the current queue length and the queue reference (Q_{ref}) value (i.e., the threshold above which, MME is considered overloaded), and gives as output the Reject Factor (RF) that represents the amount of traffic to reject in order to reduce the error signal. The principal objective is to retrieve the optimal value of the Reject Factor (reject probabilities) considering two metrics in conflict, for instance, the minimization of the queuing delay (i.e., small buffers) and the maximization of the throughput (i.e., avoiding underflows). Note that the loop represents the action of rejecting MTC traffic through the admission control operation enforced at eNodeBs. The envisioned system is based on a PID controller. We therefore use PID control function [27] in order to express the relationship between RF_i for each class of traffic C_i and $e(k)$ as follows:

$$\begin{aligned} RF_i = RF_i(k-1) + k_p(1 + \tfrac{T}{T_i} + \tfrac{T_d}{T})e(k) \\ -k_p(1 + 2\tfrac{T_d}{T})e(k-1) + k_p\tfrac{T_d}{T}e(k-2) \end{aligned} \tag{2.13}$$

where T, $e(k)$, $e(k-1)$, and $e(k-2)$ denote the sampling period, the error signal at the kth sampling instant ($t = k * T$), the error signal at the $(k-1)$th sampling instant and the error at the $(k-2)$th sampling period, respectively. The parameters T_i and T_d depend on the proportional gain k_p, the integral gain k_i, and the derivative gain k_d. They are equal to $(\frac{k_p}{k_i})$ and $(\frac{k_d}{k_p})$, respectively. The three parameters k_p, k_i and k_d represent the positive definite gains of the PID controller (i.e., proportional, integrative and derivative gains, respectively), which have an important impact on the stability and the speed of convergence of the system to the reference value. The proportional action (with k_p gain) is used to reduce the tracking error but does not eliminate it; the integrative action (with k_i gain) allows an asymptotic convergence and error rejection, and the derivative action (with k_d gain) improves the stability and the transient response. It should be noted that these values are obtained from empirical results and do not change the fundamental performance of the controller under variable network conditions. Thanks to the developed controller, the following actions are achieved:

- When the congestion is high, the tracking error is high (positive values): the system triggers a reaction by increasing the RF_i value for each class, which increases the reject probability $p_{i,j}$ to be sent to eNodeBs.
- When the system workload is low, the tracking error takes a negative value (also the RF_i values), i.e., no need for rejecting MTC signalling traffic at eNodeBs.

Now, there is a need to translate the RF_i values into a reject probability to be sent to eNodeBs participating in the congestion alleviation; i.e., eNodeBs that forward MTC signalling to the MME. We defined the following formulas in order to derive the value of $p_{i,j}$ for each traffic class i at an eNodeB j:

$$p_{i,j}(k) = \begin{cases} min(\frac{traffic_j}{\sum_{l=0}^{n} traffic_l}.RF_i(k), 1) & \text{if} \quad RF_i(k) \geq 0 \\ 0 & \text{else} \end{cases} \tag{2.14}$$

where $traffic_l$ denotes the amount of traffic coming from $eNodeB_i$. Indeed, the probabilities $p_{i,j}$ represent the proportion of MTC signalling traffic that need to be rejected, in order that the system reaches the reference value of the queue length at MME. Aiming at ensuring fairness among eNodeBs, this probability was carefully dimensioned to depend mostly on the amount of traffic generated by an eNodeB, which participates in the congestion alleviation. Thus, the higher the traffic forwarded by an eNodeB is, the higher the value of p (to be applied by this eNodeB to reduce the MTC traffic) will be. As mentioned before and following 3GPP standards, the admission control is implemented at the RAN level. Upon receiving an attach request from a MTC device, an $eNodeB_j$ applies admission control based on the reject probability $p_{i,j}$ received from the MME. Requests from MTC devices belonging to a class C_i are, thus, rejected with the probability $p_{i,j}$, by randomly choosing a uniform value between zero and one. If this value is greater than $p_{i,j}$, the MTC attach request is accepted. Otherwise, the request is rejected, and a backoff time value is indicated to the MTC device in order to ensure a good distribution of future incoming attach requests over time.

2.4 Summary of results

To conclude, we can summarize our representative works belonging to the network-oriented contributions as follows. For WLAN, we presented a delay-sensitive scheme and an admission control mechanism that are based on a thorough analysis of the tradeoff between

high network utilization and achieving bounded QoS metrics in operated 802.11-based networks. Firstly, we derived a delay estimation model to adjust the contention window size in real-time considering key network factors, MAC queue dynamics, and application-level QoS requirements. We validated the proposed model through simulation. The aim were to evaluate the accuracy of CW adaptation in monitoring MAC queuing delays. The simulation results indicated that when the network is sufficiently relaxed, there are no violations of the delay threshold for higher and medium priority flows. Also, the proposed MAC protocol ensures roughly the same delays to TC's flows regardless their respective bit rates. But, when the network load is high, there are several delay violations at high priority flows, since it is more difficult to ensure delays below the fixed threshold (0.5 sec for high priority flows).
Furthermore, we compared the proposed protocol to existing QoS-capable protocols (Adaptive EDCF (AEDCF) [29] and EDCA). Thanks to a more careful CW size adjustment, the proposed protocol achieved better results in maintaing bounded delay for high priority traffic, particularly in the case where the network load is high.
Secondly, we derived a fully distributed admission control, which provides protection for existing flows in terms of QoS guarantees and overcomes the problem of delay violation in high network load condition.
For LTE networks, we addressed the problem of congestion and system overload, which occur when MTC communications are deployed in EPS. We introduced a solution that rejects MTC traffic at the RAN using feedback from core network nodes regarding their congestion status. The system was modeled using a PID controller that controls and maintains the congestion level around an acceptable value, by tuning the amount of MTC traffic to be rejected at eNodeBs. Using computer simulation, we compared CAAC solution against the conventional approach whereby no admission control, at eNodeBs, is used for MTC signalling traffic as well as against the ACB solution introduced by the 3GPP standards. We considered three constant values of the reject probabilities p to be used in ACB. Our study showed that the ACB and CAAC mechanisms can alleviate congestion at the MME level and that is thanks to their admission control feature. The worst performances are observed when no-admission control is used, where bursts of MTC signaling traffics affect the queue length (i.e., over-utilization) resulting in packet drops. In case of ACB, we noticed that using high reject probability (p=0.5) reduces considerably the congestion at MME, but at the price of underutilizing the network resources (i.e., low throughput). Furthermore, this situation may impact the QoS expected by the MTC provider, as the MTC traffic to be sent to the remote server is severely decreased. In contrast, a low value of the reject probability (p=0.1) increases the network resources utilization, but causes packet losses at the MME. On the other hand, CAAC ensured a stable behaviour of the system, since no packet drops due to congestion are noticed, even when receiving highly bursty traffic from MTC devices.
Research work carried out in this category of contribution were done in collaboration with: Abdelhamid Nafaa (senior researcher at University College Dublin), Tarik Taleb (senior researcher at NEC Europe Laboratories), Ahmed Amokrane (Master student at University of Rennes 1) and Yassine Hadjadj-Aoul (Associate Professor at University of Rennes 1).

Bibliography

[1] A. Nafaa and A. Ksentini, "On Sustained Cross-layer QoS Guarantees in Operated IEEE 802.11 Wireless LANs ", in IEEE Transactions on Parallel and Distributed Systems (TPDS), August 2008.

[2] A. Ksentini and A. Nafaa, "Delay-based Admission Control to sustain QoS in a Managed IEEE 802.11 Wireless LANs ", Chapter in Quality of Service Architectures for Wireless Networks: Performance Metrics and Management, IGI Editors, January 2010.

[3] A. Ksentini, M. Ibrahim, "Modeling and performance analysis of an improved DCF-based mechanism under noisy channel", in Proc. of IEEE Broadnet 2007, Raleigh, NC, USA.

[4] A. Ksentini,"Enhancing VoWLAN Service Through Adaptive Voice Coder ", in Proc. IEEE ISCC 2009, The 14th Symposium on Computers and Communications, Sousse, Tunisie.

[5] A. Ksentini, Y. Hadjadj-Aoul, T. Taleb, "Cellular-based Machine-to-Machine (M2M): Overload Control", In IEEE Network magazine, November 2012.

[6] A. Amokrane, A. Ksentini, Y. Hadjadj-Aoul, T. Taleb, "Congestion Control for Machine Type Communications", in Proc. of IEEE ICC 2012 Adhoc and Sensor Networks Symposium, The IEEE Conference on Communication 2012. Ottawa, Canada.

[7] T. Taleb and A. Ksentini, "Impact of Emerging Social Media Applications on Mobile Networks", in Proc. of IEEE ICC 2013, Wireless Network Symposium, Budapest, Hungary.

[8] A. Nafaa, A. Ksentini, A. Mehaoua, "SCW: Sliding Contention Window For Efficient Services Differentiation Over IEEE 802.11 Wireless Ad-Hoc Networks", in Proc of. IEEE WCNC 2005, The IEEE Wireless Communication and Networking Conference. New Orlean, USA 2005.

[9] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," in IEEE J-SAC, vol. 18, pp. 535-547, March 2000.

[10] E. Ziouva and T. Antonakopoulos, "CSMA/CA Performance under High Traffic Conditions: Throughput and Delay Analysis" Elsevier's Computer Communications Journal, Vol. 25, No. 3, 2002, pp. 313-321.

[11] D.Pong and T. Moors, "Call admission control for IEEE 802.11 contention access mechanism," in Proc. IEEE Globecom, vol. 1, San Francisco, Dec. 2003, pp. 174-178.

[12] X. Chen, H. Zhai, X. Tian, and Y. Fang, "Supporting QoS in IEEE 802.11e wireless LANs," in IEEE Transactions on Wireless Communications, vol. 5, no. 8, pp. 2217-2227, Aug. 2006.

[13] H. Zhai, X. Chen, and Y. Fang, "How Well Can the IEEE 802.11 Wireless LAN Support Quality of Service", in IEEE Transaction on Wireless Communications, 2004.

[14] H. Zhai, X. Chen, and Y. Fang, "A Call Admission and Rate Control Scheme for Multimedia Support over IEEE 802.11 Wireless LANs," ACM Wireless Networks.

[15] Y. Xiao and H. Li, "Evaluation of Distributed Admission Control for the IEEE 802.11e EDCA", IEEE Communications Magazine, vol. 42, no. 9, 2004, pp. S20–S24.

[16] Y. Xiao, et al., "Protection and Guarantee for Voice and Video Traffic in IEEE 802.11e Wireless LANs,", in proc. of IEEE INFOCOM 2004.

[17] M. Barry, A. T. Campbell, and A. Veres, "Distributed Control Algorithms for Service Differentiation in Wireless Packet Networks," in proc. IEEE INFOCOM, vol. 1, Anchorage, Alasaka, 2001, pp. 582–90.

[18] A. Veres, et al., "Supporting service differentiation in wireless packet networks using distributed control," in IEEE J-SAC, vol. 19, no. 10, October 2001, pp. 2081-2093.

[19] Z.J. Haas and J. Deng, "On Optimizing the Backoff Interval for Random Access Schemes," in IEEE Transactions on Communications, vol. 51, no. 12, December 2003, pp. 2081-2090.

[20] G. Bianchi, et al., "Performance evaluation and enhancement of the CSMA/CA MAC protocol for 802.11 wireless LANs," in proc. of IEEE PIMRC'96, Taipei, Taiwan, Oct. 1996, pp. 392-396.

[21] H. Garcia-Molina. "Elections in a distributed computing system", IEEE Transaction on Computer, vol. 31, no. 1, Jan. 1982.

[22] 3GPP TS 23.682, "Architecture enhancements to facilitate communications with packet data networks and applications" version 11.0.0, 2012-03-12.

[23] S-Y. Lien and K-C. Chen, "Massive Access Management for QoS Guarantees in 3GPP Machine-to-Machine Communications", IEEE Communications Letter, Vol. 15, No. 3, March 2011.

[24] S-Y. Lien, K-C. Chen and Y. Lin,"Toward Ubiquitous Massive Accesses in 3GPP Machine-to-Machine Communications", IEEE Communication Magazine, Vol. 49, No. 4, April 2011.

[25] A-H. Tsai, L-C. Wang, J-H. Huang and T-M. Lin, "Overload Control for Machine Type Communications with Femtocells", in proc. Of IEEE VTC Fall 2012, Quebec, Canada 2012.

[26] CMCC TSG R2-113197, "Performance comparison of access class barring and MTC specific backoff schemes for MTC", 3GPP meeting 23-27 August 2010.

[27] A. Visioli, "Practical PID control". Springer-Verlag Editor, ISBN 1-84628-585-2, 2006.

[28] M. Chen, X. Fan, M-N. Murthi, T-D. Wickramarathna, and K. Premaratne, "Normalized queueing delay: congestion control jointly utilizing delay and marking.", in IEEE/ACM Transaction on Networking. 17, 2, pp. 18-631, April 2009.

[29] L. Romdhani, et al., "Adaptive EDCA: Enhanced Service Dif-ferentiation for IEEE 802.11 802.11 Wireless Ad-Hoc Networks," in Proc. of IEEE WCNC'03, New Orleans, USA, March 2003, vol. 2, pp. 1373-1378.

Chapter 3

Human-centric contributions

Contents

3.1 Introduction

Human-centric solutions have gained attention only the last few years. This is mainly due to the fact that most of QoE monitoring tools cannot be used in real-time. Thanks to Pseudo Subjective Quality Assessment (PSQA) tool, we were able to propose mechanisms that consider user QoE when controlling network functions. PSQA is a real-time and a parametric QoE monitoring tool developed at the INRIA Dionysos team.
It is worth noting that works in this category of contributions were not limited to the network control mechanisms, but also in developing new PSQA modules for emerging audio and video codecs.
Human-centric approaches were used to address the following issues:

- PSQA module for monitoring QoE
  - Emerging VoIP codecs (Silk, Speex and iLBC) ([1]):
    We built a new PSQA module for estimating user QoE in real-time for three emerging codecs: iLBC, Speex and Silk (used by Skype). To develop the new module, we replaced the subjective procedure, which is usually used to evaluate the distorted sequences, by an objective technique, which is less time consuming. This technique allowed us to reduce considerably the time needed for the learning procedure.

  - Hierarchical SVC video coding ([2][3]) (Details in Section 3.2)

- Energy conservation for VoWLAN ([4][5]):
  We addressed the problem of energy conservation in VoWLAN application by proposing a cross-layer solution. According to user QoE feedbacks (obtained at run-time), we proposed to derive the sleep periods that maximize power conservation while maintaining users' QoE above a certain threshold. Since the sleep period has a direct impact on user QoE, we modeled the transmission chain by using control theory. In fact, we used a PID controller in order to derive the sleep period that maintains user QoE around a fixed reference value.

- Multicast optimization in WLAN ([6][7]) (Details in Section 3.4)

- Network selection in the context of WLAN ([8]):
  We tackled the problem of network selection and load balancing in the presence of several WLAN AP, by proposing a QoE-oriented mechanism. Based on the concept of 802.11k [48], instead of giving radio measurement information, we append QoE information into Beacon and Probe Request frames. The QoE information (average QoE) is gathered from the stations connected to the AP. Then, users select the network that provides the highest score or they may not connect to any access point if they consider that the current scores are too low for the requirement of their applications. As a consequence, they will connect to the network where they will be best connected and avoid high-loaded networks automatically due to the lower QoE in those networks. Therefore, the scheme is profitable for preventing access networks from over- or under-utilization.

- Network selection in the context of Wireless Heterogeneous Networks ([9][10]):
  We considered the network selection problem in case of heterogeneous wireless access as a Multiple Attribute Decision Making (MADM) [50] problem. To solve it, we used the Preference by Similarity to Ideal Solution (TOPSIS) technique. The originality of our approach lies in the inclusion of the measured QoE value as an attribute for the selection of the access network.
  Two scenarios were considered. The first one represents the case where the access networks (3G/4G and WiFi) belong to different operators. Hence, the choice of the network, and the execution of the TOPSIS algorithm are done by the station. To facilitate the TOPSIS execution, potential access points communicate the mean MOS value perceived by their users to the concerned station.
  The second scenario denotes the case where the access networks belong to the same operator. Unlike the first scenario, the choice of the network is performed by a central entity at the operator core network. Here, the access points (WiFi and 3G) report periodically the value of the QoE perceived by users to the central entity, which in turn executes the TOPSIS algorithm. The selected network decision is then sent to the station.

- Uplink scheduling in 3G cellular networks ([11]):
  Most of the proposed scheduling mechanisms for cellular networks (mainly 3G) are taking into account only the signal quality and fairness as criteria, leaving user experience. To overcome this gap, our contribution consists of a novel QoE-oriented scheduler. It takes into account the quality of experience when selecting the stations having the opportunity to transmit on the uplink channel (scheduling process). The main idea is to give priority to video streaming users, which have more constraints

on the quality. For this, a coefficient is assigned to each user. This coefficient is then multiplied with a priority index as in traditional scheduling mechanisms. Based on the QoE measured at the mobile station, the QoE-oriented scheduler differentiates between the way of computing the coefficient of video users and the best-effort users as follows: if the QoE perceived by a video user is below a specific threshold, the scheduler increases the coefficient of these users and decreases those of best-effort users. Accordingly, the video users will have a higher probability to transmit in the next uplink frame.

- In-network adaptation of SVC streams
  - Case of DVB-T2 ([12][13]) (Details in Section 3.3)
  - Case of multipath communication in Video Delivery Networks (VDN) ([14][15]): In this contribution we proposed an approach that couples the SNR scalable video coding (SVC) extension of H264/AVC, with the path diversity provided by VDN. Our method adapts to the heterogeneity of end-users using the scalable video coding. Moreover, it adapts to network bandwidth fluctuation by observing the changes of the available bandwidth over the multiple overlay paths, and updating the streaming strategy accordingly.

In this chapter, we will particularly focus on three representative contributions belonging to the human-centric solutions: (i) QoE metric for SVC; (ii) QoE-based in-network adaptation of SVC in DVB-T2;(iii) QoE-based Multicast adaptation in WLAN. The first contribution shows an example of developing a new QoE monitoring module based on PSQA. This module is dedicated to the emerging SVC codec. The second contribution uses this QoE monitoring tool to adaptively select the number of SVC layers to decode and to display to the final user in DVB-T2 (unidirectional) networks. The third contribution shows the case of using the QoE as a metric to control the data rate of multicast communications in WLAN. Generally speaking, these three contributions show how we can use QoE for controlling network functions with or without using user feedback.

3.2 QoE metric

3.2.1 Research context and related work

Quality of Experience is defined in [16] as "the overall acceptability of an application or service, as perceived subjectively by the end user". Quality of Service is defined in [17] as "the collective effect of performance which determines the degree of satisfaction of a user of the service". In telecommunications, QoS is usually a measure of performance of the network itself. QoE instead focuses on the overall experience of the user. It depends on the global system behavior, going from the source of the services until the user, including the content itself and the network performance. There are several factors that can influence QoE for video applications. Characteristics, such as frame rate of a video stream, may impact the fluidity of the video: a lower frame rate means "choppiness" that can degrade the perceived quality. Spatial video resolution is another significant factor: depending on the limitations of the end device, users may prefer the highest available resolution. Another important factor related to the video quality is the Quantization Parameter (QP). In fact, this parameter relates to the compression of a video stream. Thus, during compression, some amount of information is thrown away and this will introduce certain distortion in the video that may, in turn, have an impact on QoE.
In addition to the preceding remarks, the type of video content, itself, may have significant importance. For example, a video of a news reader might have low frame rate requirements, but higher quantization requirements. On the other hand, a fast moving video, such as that of Formula One racing coverage, requires higher frame rates to ensure good QoE. The network used to provide the service can significantly impact the video quality. For example, packet losses can strongly degrade the video's perceived quality. Delays and jitter in the network may incur, first, a long initial delay before a video can start to play, and then, play-out disruptions and eventual data losses because of the video packets that miss the play-out deadline. In addition, other parameters, such as network bandwidth, impose limitations on the video characteristics because some quality of the video will get downgraded, either by lowering the frame rate or by using more compression, to accommodate the video with the available bandwidth.

Depending on the method used to evaluate user QoE, QoE measurement tools can be

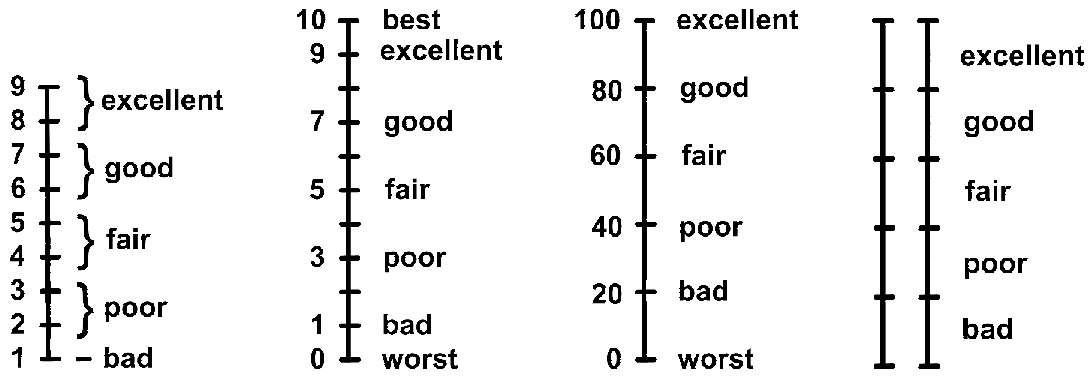

Figure 3.1: Examples of ITU standard scales for subjective test methods.

classified into two categories. The first class is based on subjective evaluation tests while the second class is based on objective evaluation (signal processing algorithm). Subjective

evaluation tests are based on personal evaluation of users, where a panel of selected persons rate video sequences. The output of these tests is a Mean Opinion Score (MOS), where different scales can be used as specified by the ITU-R, as shown in Figure 3.1. Indeed, the ITU-R recommendation BT.500-10 formalizes the subjective test procedure by introducing several experimental conditions, such as viewing distance and viewing conditions. Although subjective tests are highly accurate in estimating users' QoE, their preparation and execution are costly and time consuming. Furthermore, they cannot be used in real-time or automatically.

Objective quality evaluations, on the other hand, are algorithms and formulas (i.e., generally signal processing algorithms) that measure, in a certain way, the quality of a video stream. Objective video quality metrics range from very simple metrics to very complex ones [18] (e.g., metrics based on human vision systems (HVS)). Objective quality metrics can be further classified into three different categories, namely Full-Reference (FR), Reduced-Reference (RR) and No-Reference (NR), based on the amount of information available for comparison with the original content. Full- and reduced reference mechanisms are mainly used to evaluate video quality in non real-time scenarios where both the original video reference (or reduced data set) and the distorted video are available.

The PSNR metric is one of the most widely used full-reference metrics for objective video quality assessment, thanks to its simplicity and its low computational requirements. It calculates the ratio between the maximum value of a signal and the background noise. PSNR is used because of its physical significance and simplicity but the performance of this metric is quite poor, as it usually does not correlate well with subjective scores. Also, it is difficult to reliably derive MOS from this metric, despite the existence of heuristic mappings of PSNR to MOS. The Structural Similarity (SSIM) approach [19] provides an alternative and complementary way to tackle the problem of video quality assessment. It is based on a top-down assumption that HVS is highly adapted for extracting structural information from the scene, and hence a measure of structural similarity should be a good approximation of perceived image quality. Nevertheless, the SSIM index achieves the best performance when applied at an appropriate scale (i.e., viewer distance/screen height). Calibrating the parameters, such as viewing distance and picture resolution, represents an important challenge for this approach. Video Quality Metric (VQM) [20] is a standardized method for objectively measuring video quality by making a comparison between the original and the distorted video sequences based only on a set of features extracted independently from each video. The algorithm used by VQM measures the perceptual effects of several video impairments, such as blurring, jerky/unnatural motion, global noise, block distortion, and color distortion. These measurements are combined into a single metric that gives a prediction of the overall quality. For more details on FR methods reader can refer to [19].

Despite their efficiency in evaluating the video quality, FR methods are only applicable when the original video sequence is available. This constitutes a limitation when there is a need to evaluate QoE in real-time at the decoder side or at an intermediate network node. To address such situation, NR methods have been proposed. The main goal of NR methods is to create an estimator based on the proposed features that would predict the MOS of human observers, without using the original image or sequence data. Furthermore, since the model does not require any comparison of signals, the calculations can be performed in near real-time. Previously introduced NR methods do not estimate the overall user quality but estimate the degree of blockiness [21], which is the most prominent artifact of block-DCT based compression methods such as H.26x, MPEG and their derivatives. To increase the estimation accuracy, work in [22] uses the bit-stream information

which depends on the compression algorithm. Such a method suffers from the fact that it cannot differentiate video quality degradation from features of the video itself and those introduced by the network (e.g., loss and delay). Other NR methods incorporate network information to enhance user quality prediction. The ITU recommendation ITU G.1070 [23] (also known as "opinion model") is a NR model which uses the bit rate and frame rate of the compressed video along with the expected packet loss rate of the channel to predict video quality. Work in [24] showed that it is possible to enhance this model and make it more precise by replacing, for example, packet loss rate with packet loss event rate. A further extension of this mechanism is proposed in [25].

On other hand, PSQA is a quality assessment tool that is a hybrid between subjective

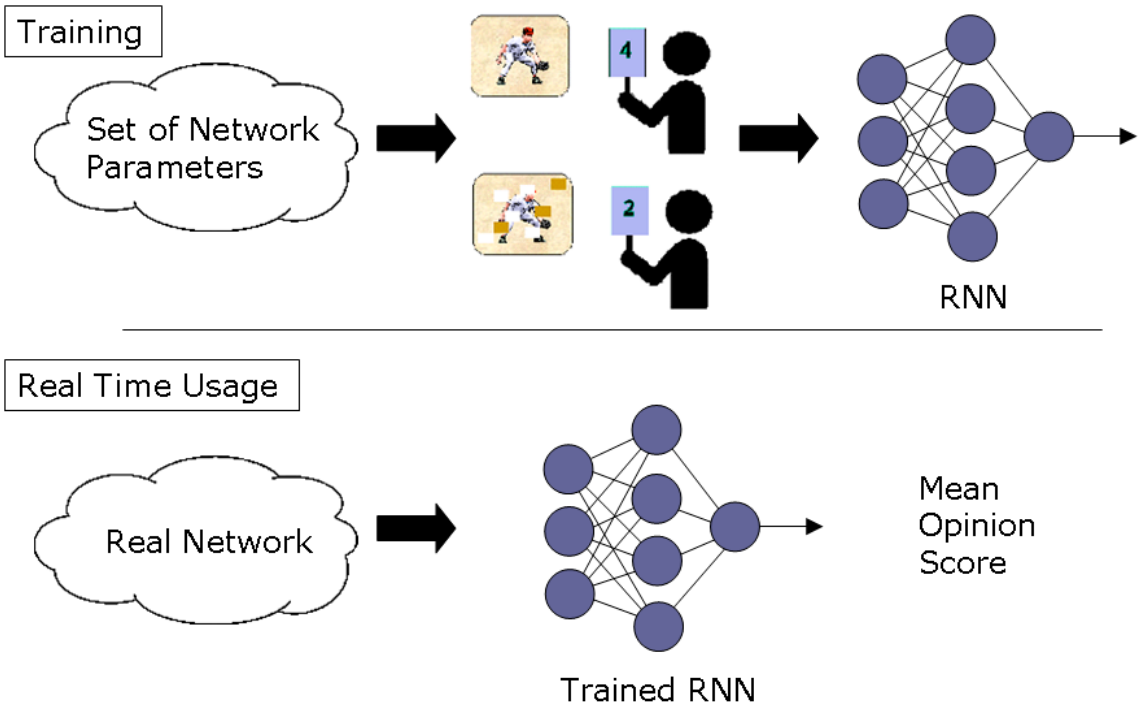

Figure 3.2: PSQA methodology.

and objective evaluation techniques. The idea is to do subjective tests for several distorted videos and use the results of this evaluation to teach a learning function $f()$ based on Random Neural Network (RNN) the relation between the parameters that cause the distortion and the perceived quality. The PSQA methodology is shown in Figure 3.2 and is explained in the following text.

The procedure consists in first identifying the parameters that have an impact on QoE in the given context. These parameters may vary from one context to another and some examples of these parameters are: type of codec, packet loss rate, delay, jitter, etc. A video database is then generated by simulating the identified parameters after choosing a set of representative values for each of them, together with an interval for the parameter, according to the conditions under which we expect the system to work. Then, a uniformly sampled subset Ω of this video database is subjectively evaluated by a panel of humans. After statistical processing of the answers (designed for detecting and eliminating bad observers whose answers are not statistically coherent with the majority), each video sequence in Ω receives a QoE value (often, this is a Mean Opinion Score, or MOS). It results

in Ω configurations of the parameters and a corresponding QoE score or MOS. Then some of the configurations are used for training the RNN and the remaining ones are used for validation that, in turn, are not shown to the RNN during training.
In order to validate the trained RNN, a comparison is done between the value given by the trained RNN at the point corresponding to each configuration in the validation set and to its actual MOS value; if they are close enough (having low mean square error), the training is validated. If the validation fails, a review of the chosen architecture and its configurations is needed.
Note that a comparison between PSQA and other QoE metric tools is presented in [27].

3.2.2 Contribution

One of the multimedia market trends is audiovisual service (TV or VoD) anywhere, at any time. To support such service, a Video Service Provider has to manage, store, and distribute content towards multiple kinds and scales of terminals, and over different and transient access technologies to reach the end user. To solve such issues, video scalability seems to be the most relevant solution. We recall that with SVC, it is possible to constitute a set of layer combinations to create the video streams. But to evaluate such combinations, and monitor the performance of the SVC encoding scheme, in term of user experience, there is a need to estimate automatically the QoE of SVC video streams.
To Build PSQA for SVC, we need to clearly identify the parameters impacting the perceived quality of the video streams (as SVC is composed by several layers). There are different parameters that affect the quality of H.264/SVC videos. They include parameters related to the content itself (such as brightness, contrast, sharpness, color, motion), the encoding parameters (such as QP) and other parameters dependent on the transport network (such as delays, loss rate, bandwidth). To efficiently use PSQA, it is important to consider parameters that can be obtained in real-time and with low complexity. The first parameter affecting the quality of the SVC streams is the frequency of IDR (Instantaneous Decoder Refresh) at the encoding side. Unlike coding mechanisms such as MPEG, SVC uses a different structure of the Group Of Picture (GOP). In SVC, the GOP structure consists of one key frame (IDR and P) and the remaining are B frames [28]. In addition, SVC adopts a hierarchical coding structure for B frames, as illustrated in Figure 3.3, to facilitate the temporal scalability implementation.

IDR frames (I frame) are special frames due to the fact that they are encoded without reference to another frame. From hereunder, we indifferently denote them as IDR or I frames. The IDR frames are periodically sent in order to refresh the decoder buffer and create a new point of reference. In fact, an increase in IDR frequency (IDR period) means an increase in the number of IDR frames that, in turn, decreases the number of P frames and B frames. High numbers of IDR frames are beneficial to reduce error propagation during the refresh period. In Figure 3.4 we illustrate the dependence of affected frames on the location of key picture within the intra refresh period. The affected frames vary according to the lost frame type (I, P or B) as well as to the position of the lost frame in the intra-refresh-period (or IDR period). Unlike MPEG, loosing a B frame has impact on other B frames, of course depending on its position in the hierarchy. For instance, in Figure 3.3, loosing frame B4 impacts all other B frames in GOP. However, loosing frame B1 has no impact. Loosing a P frame has an impact on all frames (B frames) that use this frame as a reference. The worst case occurs when an IDR frame is lost. Indeed, loosing such frame causes distortion not only on the current GOP, but also to precedent GOPs. Loosing IDR frames affects both P and B frames.

As noticed in the precedent versions of PSQA, the packet loss rate is an important pa-

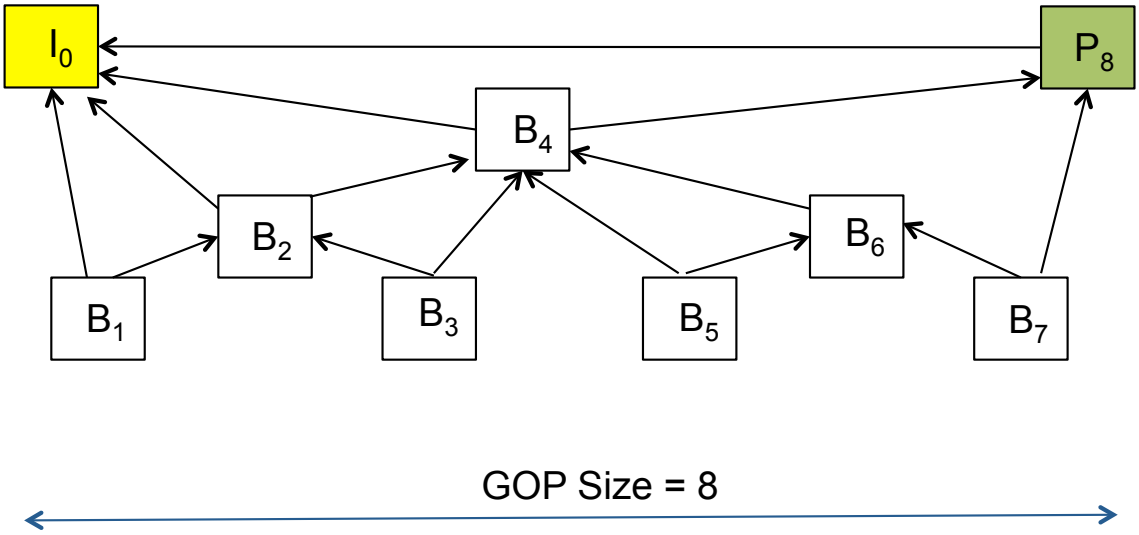

Figure 3.3: GOP structure in SVC (single layer case).

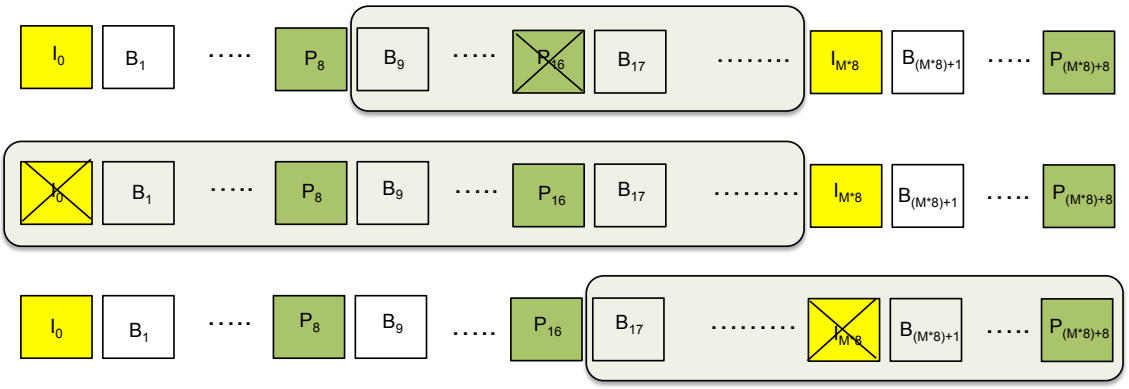

Figure 3.4: Frames range affected by the loss of key frames (single layer case).

rameter that affects the video quality. This is more evident in case of streaming SVC over UDP-based networks or over DVB, where the transport layer is not reliable. For SVC, we must consider the packet loss rate for each layer composing the SVC stream. The NALU (Network Abstraction Layer Unit) is the transport unit of video packet used in H.264 (SVC and AVC). In case of SVC, NALU can only carry information of one layer. The loss of a NALU affects only a single layer. However, it is worth noting that losing a NALU belonging to the base layer has more impact on the video quality, than the loss of a NALU belonging to other enhancing layers (particularly when using spatial or quality scalabilities). Besides affecting the other frames belonging to the base layer (Figure 3.4), a loss of base layer NALU impacts the other layers as all the other layers in SVC use the base layer as reference and any error in this layer propagates to other layers (see Figure 3.5 for intra layer prediction in case of SNR (CGS) scalability). Hence, this situation seriously downgrades video quality.

Usually, a NALU packet consists of one header (as AVC header), and a specific header extension (Figure 3.6). This extension has particular fields D, Q and T, which are used to identify the spatial quality and temporal layers, respectively. We denote by $P =$

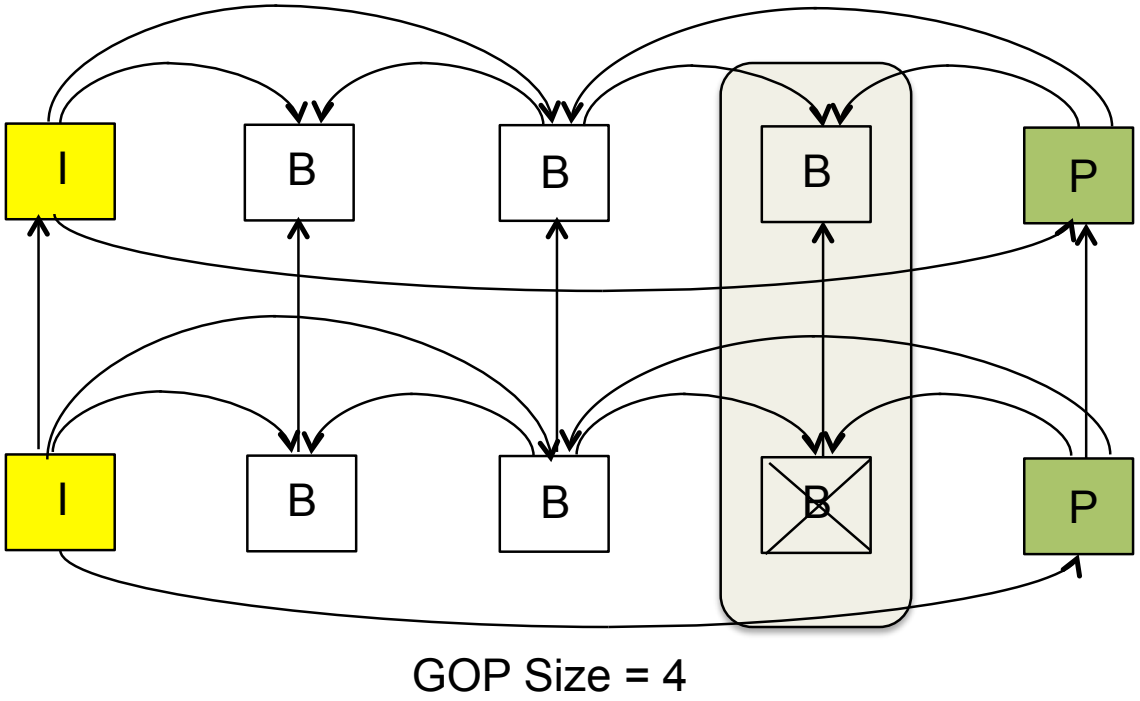

Figure 3.5: Range of frames affected by the loss of a B frame (base layer and one enhanced layer (CGS) case).

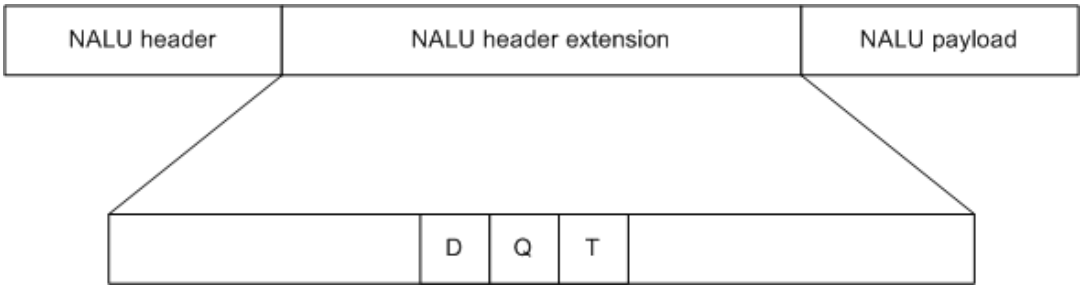

Figure 3.6: A typical NALU format.

$\{L_{BL}, L_1, L_2, ..., L_N, f_{IDR}\}$ the set of affecting parameters, where f_{IDR} represents the IDR frequency, L_{BL} and L_N denote the NALU loss rate of the base layer and layer N, respectively. It is important to note that there is another parameter that affects the quality of H.264/SVC videos, which is the error concealment procedure used by the decoder. Indeed, such a mechanism can help the decoder to replace information lost due to NALU losses. In this SVC module of PSQA, the error concealment is implicitly taken into consideration as the distorted videos are evaluated at the subjective phase and are obtained after the decoder has applied the error concealment procedure. Therefore, in this version of PSQA, we are relying on the error concealment provided by the SVC decoder.

In order to train the learning function $f()$, which monitors user QoE, we used five different video sequences (m) (Table 3.1) for constituting the set of the media sample (δ_m). We encoded the different videos by using the JSVM encoder [29]. For the decoder side, we

Table 3.1: Videos used for training

Video	NAL number
CITY	300
CREW	300
HARBOUR	300
ICE	240
SOCCER	300

Table 3.2: Video parameters

Fixed parameter	Value
Resolution	704x576 (4CIF)
Frame Rate	30
Layer(QP)	BL(34) - EL1(28) - EL2(22)

used the openSVC soft [30]. The resolution is 4CIF (704 x 576), the frame rate is set to 30 frames/s, and the values of Quantization Parameter (QP) are set to 34, 28, 22, respectively for layer 0 (Base Layer), 1 and 2 as shown in Table 3.2. The video scalability is based on the SNR (CGS) quality, by reducing the QP parameter for each enhanced layer by Delta QP (DQP). As our focus is on employing only quality-based scalability, we used only key frames (IDR and P frames) to constitute the GOP of each video sequence. In this case the GOP size is equal to 1, which means that no B frames are used and the GOPs are in the form of "ÌPPP...PIPPP...". Since the NALU losses (L_{BL}, L_1, L_2) have a serious impact on the final quality of a video, we proposed that each value of these parameters (NALU losses) is to be taken from the set noted $V = \{0, 0.3, 0.5, 1, 3, 5, 10\}$. The set V represents the loss rate (i.e., from 0% to 10%, quality is already very bad at 10% loss). Regarding the parameter f_{IDR}, three values were considered: 75, 150 and 300. Table 3.3 shows the set of values associated with each parameter.

A large number of videos were created using different combinations of the above parameters and their values given in Table 3.3. The obtained sequences correspond to the set of configurations $\{\Omega_1, \Omega_2, ..., \Omega_n\}$. For instance, $\Omega_1 = \{0, 0, 0, 75\}$ represents a configuration, where $L_{BL} = 0\%$, $L_1 = 0\%$, $L_2 = 0\%$, $f_{IDR} = 75$. Next steps consist of: (i) reducing the set of video sequences $\{\delta_{1S}, ..., \delta_{iS}\}$ to 500 by employing uniform sampling; (ii) using manual evaluation to select around 100 among 500 for the subjective test session, such that 25 videos corresponded to a MOS score between 1 and 2, 25 videos between 2 and 3, and so on. The MOS scale, shown in Table 3.4, was used for all quality evaluations.

Table 3.3: The values of parameters affecting QoE.

Parameter	Set of values
NALU loss rate for Base Layer (%) (L_{BL})	0, 0.3, 0.5, 1, 3, 5, 10
NALU loss rate for Layer 1 (%) (L_1)	0, 0.3, 0.5, 1, 3, 5, 10
NALU loss rate for Layer 2 (%) (L_2)	0, 0.3, 0.5, 1, 3, 5, 10
IDR Frequency	75, 150, 300

Table 3.4: The values of parameters affecting QoE.

MOS	Quality	Level of Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

These 100 video sequences were evaluated by 15 humans using the DSIS (Double Stimulus Impairment Scale) methodology [31]. Indeed, the reference video is first shown to the evaluators' panel, and then followed by the distorted sequences. This method is more useful for evaluating clearly visible impairments caused by transmission. The obtained scores were then pre-processed, as in [32], to check for inconsistent scorers. Then the data was used to train the learning function.

3.3 QoE-based in-network adaptation of SVC flows in DVB-T2

3.3.1 Research context and related work

DVB-T2 [33] is the new standard for digital video broadcasting that aims at replacing the DVB-T (i.e., the first generation of the terrestrial broadcasting standard) for broadcasting terrestrial television. Based on advances made in digital signal processing, and specifically in channel coding, DVB-T2 brings a new flexibility in services' broadcasting with an increased transfer capacity of 50%, when compared to DVB-T. Besides using the Coded Orthogonal Frequency Division Multiplexing (COFDM) in order to be more robust against multipath channels [34], the DVB-T2 physical layer data channel is divided into logical entities called Physical Layer Pipes (PLP). Each PLP carries one logical data stream that could be an audio-visual multimedia stream along with the associated signaling information, or hierarchical video streams which can address at the same time different qualities. The PLP architecture is designated to be flexible so that arbitrary adjustments of robustness and capacity can be easily done. Thus, using different PLPs enables broadcasting, on a single radio channel, multiple services, or groups of services, with different channel coding and modulation settings. Broadcasting several service components over the same channel has thus become possible, with differentiated levels of robustness, which was not possible with the previous DVB-T standard or other broadcasting technologies [35]. Using this new capability allows handling users' channel diversity, where users with good channel condition can decode all PLPs and access to high quality contents, while users with poor channel conditions (such as mobile terminals) can decode only robust PLP but at least can access the lowest service quality; i.e., in case of DVB-T all services are lost. Needless to say that associating SVC with DVB-T2 can easily address the problem of broadcasting added value services to high number of users, despite their radio environments and terminal capabilities, which represent an interesting solution for operators [36]. In fact, the hierarchical physical layer provided by DVB-T2 can be easily combined with hierarchical video coding features proposed by SVC. Thus, each SVC layer is broadcasted through different PLPs. The base layer is sent through the most robust PLP, usually PLP0. The enhanced layers are sent through other PLPs, which use less robust physical modulation, but allow using higher data rates. Thus, stations with good physical channel conditions can decode all the SVC layers and benefit from high video quality, while users with poor channel conditions can at least decode the base layer and benefit from acceptable quality.
Apart from the work presented in [37], most of related research works on SVC and DVB-T2 have focused on exploiting each technology separately. In [37], the authors discuss the deployment of SVC in DVB-T2, and particularly concentrate on providing optimal usage of DVB-T2 features from SVC's point of view. In addition, they provide modifications to the error protection mechanism, at the physical layer, in order to improve users' experience. Besides considering only physical layer enhancements, the authors employed the PSNR tool to calculate user perception, which is known for its lack of efficiency to reflect user's QoE.

3.3.2 Contribution

The aim of this contribution is to further concentrate on the enhancement achieved, at the user side, when associating SVC and DVB-T2. We concentrate on user's experience, in terms of QoE, as the main criteria for evaluating this association, and demonstrate how

QoE can help optimizing this association for providing context-aware video services to an end-user through the broadcast channel. Let us consider Figure 3.7, which presents the

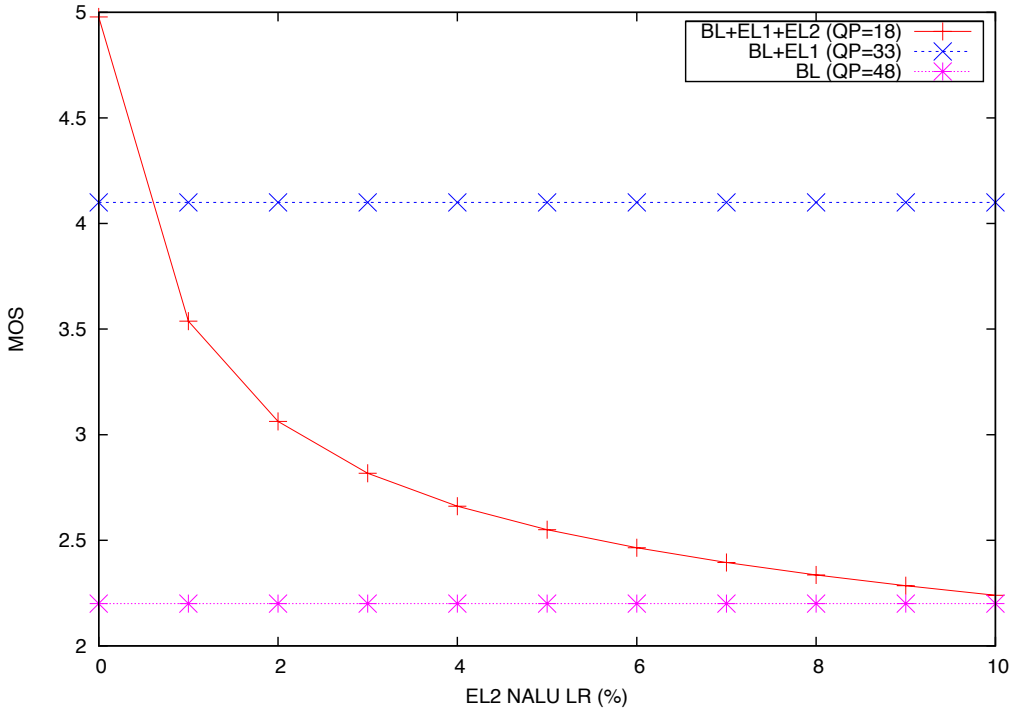

Figure 3.7: A typical NALU format.

PSQA scores computed for the following three cases: (i) the three layers are decoded and displayed to users, where the enhanced layer 2 experiences NALU loss; (ii) the two lower layers are decoded; (iii) only the base layer is decoded. Although NALUs belonging to the enhanced layer 2 have less importance than those of the base layer, loosing NALUs from this layer degrades the user's QoE. Therefore, if the enhanced layer 2 (highest layer) experiences losses, for the sake of maintaining good QoE, it is worthwhile withdrawing this layer and not displaying it to the end users. Indeed, decoding only the base layer and the enhanced layer 1 achieves better MOS than decoding all layers, particularly if the L_2's NALU Loss Rate exceeds 1%. It is clear that, in some cases, withdrawing one SVC layer can maintain higher MOS than the case where this layer is decoded. Accordingly, we proposed QoE-BASD, a new scheme that dynamically selects the number of SVC layers to be decoded and shown to the end-user aiming at maximizing the user's QoE. Based on the results of Figure 3.7, we proposed to use QoE as the main criteria for selecting or withdrawing a SVC layer at the receiver side. The proposed QoE-BASD algorithm assists the decoder to decide which layers to be displayed at the end user. QoE-BASD is compatible with DVB-T2 unidirectional communication principle, since it is executed at the receiver side without the need for any feedback or signaling messages.
Let us assume that n layers compose the broadcasted video stream. We consider representing the initial SVC stream with a Matrix noted Mat. Indeed, the aim of this matrix is to

Algorithm 1

1: **loop**
2: **for** $i = 1$ to n **do**
3: $\Delta[i] \leftarrow$ Compute_QoE(Mat_i) // Mat_i is a vector constituted by the line i of the matrix Mat.
4: **end for**
5: $\Delta[k] \leftarrow \max(\Delta)$
6: Decode and display the k layers
7: **end loop**

decompose the initial SVC stream into a combination of layers composing n streams. Each line of the matrix corresponds to a possible SVC stream composed of k layers ($n \geq k \geq 1$), whereby each column represents a layer. The element $Mat(i, j)$ is equal to one if the jth layer is present in the ith SVC stream. Otherwise, the element $Mat(i, j)$ is equal to null. Usually, the received SVC stream at the decoder side represents the last line of the matrix. Using QoE-BASD, the aim is to build the matrix Mat and selects the line of the matrix (SVC stream) that maximizes the user's QoE. In contrast, the classical approach systematically decodes all the layers composing the last line of the matrix.

$$Mat_{4,4} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \tag{3.1}$$

For simplicity, we consider the above matrix $Mat_{4,4}$, which represents the case of an initial SVC stream composed by 4 layers. For instance, the first line shows the case of a SVC stream (noted 1) composed only by the base layer, while the $4th$ line represents a SVC stream (noted 4) constituted by the base layer and 3 enhanced layers. The QoE-BASD algorithm is illustrated in Algorithm 1. First, for each line of the matrix Mat, QoE-BASD computes MOS by considering the parameters $(f_{IDR}, L_{BL}, L_1, L_2)$ for line i. We denote by $\Delta = (\Delta_1, \Delta_2, \Delta_3, .., \Delta_i)$ the vector containing the i MOS values obtained for each line. After that, the maximum value of MOS (noted Δ_k) of this vector is selected. Accordingly, the kth layers composing the SVC stream are decoded and displayed to the user, while the other layers (layers higher than k) are withdrawn. The execution periodicity of QoE-BASD (the loop) is repeated every t seconds, which corresponds to the duration of a GOP. To build the vector Δ, we used the PSQA SVC module. As stated before, PSQA needs to be able to measure the parameters $(f_{IDR}, L_{BL}, L_1, L_2)$ affecting the SVC quality automatically, above all in real-time. Usually, f_{IDR} is a static value; it can be obtained when encoding the video stream. Another way of obtaining f_{IDR} is by tracking the appearance of the IDR frames in the video stream. An IDR frame, in turn, can be identified by parsing the NALU headers that contain the information about whether its payload corresponds to an IDR frame or not. As mentioned before, higher values of the IDR frequency means higher resilience to the error propagation, but for the price of larger video sizes. On the other hand, the NALU loss rate $(L_{BL}, L_1, L_2, ..., L_N)$ for each SVC layer is obtained by relying on the Real Time Protocol layer. Combined with RTP simple packetisation mechanism (Single NAL Unit) [38], we propose using a multi-session RTP connection for each layer. In other words, for each layer an independent RTP session is established, where one NALU is conveyed in one RTP packet. Thus, we can obtain the NALU loss rate for each SVC layer at the RTP level, either at the decoder side, or by enabling the RTCP protocol for a remote monitoring if there is a return channel available from the receiver

to the server. IP traffic is supported in DVB-T2 by either using MPEG-TS encapsulation or through the Generic Stream Encapsulation (GSE) [39] that provides an appropriate IP encapsulation over PLPs.
It is important to note that the complexity of the QoE-BASD algorithm mainly depends on the number of layers n composing the SVC video stream. Since most of the existing implementations of SVC use three layers (i.e., one base layer and two enhancement layers), QoE-BASD can be easily implemented in most DVB-T2 products; even in those with low CPU capacities.

3.4 QoE-based Multicast optimization in WLAN

3.4.1 Research context and related work

Various applications support multicast; for example, conference meeting, mobile commerce (mobile auctions), military command and control, distance education, entertainment services, and intelligent transport systems. In WLAN, multicast traffic has been set to the lowest transmission rate (basic rate) in order to reach all mobile nodes especially the further ones, because they suffer from important signal fading and interference. The lower rates disadvantage transmission in terms of channel occupancy since they take longer time than the higher rates to send the same amount of information. This performance anomaly has been presented in [40], where it is mentioned that a slow host may considerably limit the throughput of other hosts roughly to the level of lower rate. Another constraint in multicast transmission is the lack of acknowledgment and retransmission due to huge amount of traffic overhead generated by these packets. This is severe when transmission mode is multicast because the number of acknowledgments/retransmissions will be multiplied by the number of recipients in the multicast group; this could cause collision due to feedback implosion.
Rate-adaptation in WLAN has gained high interest, not only in the multicast context but also for unicast communications. The first and widely used rate adaptation protocol in commercial products is Auto Rate Fallback (ARF) proposed in [41]. In ARF, when SNR decreases, an access point tries to recover by decreasing the bandwidth. In fact, the access point switches to a higher rate when a certain number (ten) of packets has been successfully received; it switches back to the lower rate when a failure occurs right after rate increase. If a failure occurs when the number of consecutive successful transmissions is less than ten, the access point switches to a lower rate only after two consecutive failures. Regardless of its wide implementation in commercial products, the protocol has a drawback resulting from the static-threshold approach, which does not adapt well to varying condition in wireless networks. To solve disadvantages from the static-threshold approach, Adaptive ARF (AARF) has been introduced in [42]. The authors also use threshold-based mechanism as in ARF but instead of setting it to a fixed number, the threshold follows binary exponential backoff continuously at runtime to better reflect to the channel conditions. In other words, they multiply by two the number of consecutive successful transmission required to switch to a higher rate. The mechanism increases the period between successive failed attempts to use a higher rate results in fewer failures and retransmissions, thus the overall throughput is improved. Despite that AARF is an efficient mechanism, it cannot be used in multicast transmission since the implementation of this protocol requires acknowledgment and retransmission, which are disabled in multicast. Another popular protocol is Receiver-Based Auto Rate (RBAR) presented in [43]. It enables the RTS/CTS mechanism in order to get/send feedback from receiver. In fact, RTS is sent out before each transmission by the sender and it is received by the receiver which computes the SNR of the frame. After consulting a table mapping of SNR and rate, the receiver sends back the transmission rate for the sender to use in the next transmissions in CTS. RTS and CTS headers have been modified for this purpose. This mechanism is based on SNR (computed with a priory channel model), which is a physical parameter that does not always correlate well with human perception. Moreover, RTS/CTS mechanism is unusable in multicast transmission. Based on similar idea of using RTS/CTS in RBAR, Rate Adaptive Multicast protocol (RAM) has been proposed in [44] for channel estimation and rate selection. In this protocol, multicast receivers use the RTS packet to measure channel condition and send back the transmission rate to be used by the sender through CTS packet. In case that a member

does not receive the data frame correctly, it will send a NACK (Not Acknowledge). For enhancing the throughput, the authors added a frame sequence field to RTS. This field is used by the member to check whether multicast data frame is a new frame or retransmission. If a frame is a retransmission of a previously successfully received frame, a member will not participate in this multicast transmission. This reduces the number of retransmissions. It can be noticed that the protocol makes use of RTS/CTS, NACK and retransmission, which are disabled in multicast. In addition, there are many modifications to existing frames. To overcome the feedback implosion problem, the Leader-based Rate Adaptive Multicasting for WLANs (LM-ARF) protocol that deploys a leader-based feedback approach and adapts data rate according to ARF is introduced in [45]. One of the receiving stations, which is the leader, is responsible for sending ACK on behalf of the participating multicast stations. If any multicast receiver, which is not the leader, fails to receive a multicast frame, it will send a negative acknowledgment (NAK) to request retransmission. The AP adjusts the contention window size in a manner same as for unicast transmission, thus keeping fairness between unicast flows. A new frame type called CTS-to-Self has been added in order to guarantee the channel access and announcing the transmission of a multicast frame. This mechanism covers several aspects such as fairness, reliability, and performance; however, since it uses ARF, it also inherits the static-threshold approach and drawback of ARF as well.
To avoid using RTS/CTS, Auto Rate Selection for Multicast (ARSM) protocol is proposed in [46], which uses a multicast channel probe operation (MCPO), where a multicast probe frame sent out by AP before sending multicast traffic. In this protocol, the user having the lowest SNR will be the one in charge of replying to the AP by a multicast response. The AP then selects the multicast data rate based on feedback in three different ways: explicit, implicit, and no feedback. For avoiding collision, multicast users select backoff timer according to their SNR value.
Taking into account user perception, SNR-based Auto Rate for Multicast (SARM) is presented in [47]. It adapts transmission rate according to SNR of the node experiencing the worst channel condition. SNR references are obtained from a table listing required SNR for PSNR (peak signal to noise ration) to be higher than 30 (representing good quality) for each transmission rate. By changing multicast transmission rate on the basis of SNR values reported by mobile nodes, the wireless channel is used more efficiently. To overcome the lack of feedback mechanism in multicast, the authors propose a channel probing mechanism to inform the access point of the channel quality at mobile nodes. To avoid collision when nodes transmit feedback to the access point, the author also proposed a backoff timer for each mobile node based on the received SNR. But using PSNR as the metric for changing rate does not always reflect accurately the user experience.
For better comprehension, we summarized the described schemes in Table 3.5.

3.4.2 Contributions

Similar to the SARM mechanism, we proposed two human-centric mechanisms for controlling the data rate of multicast communications in WLAN. Unlike SARM, we used PSQA which is more efficient to reflect user satisfaction than PSNR. The first contribution uses a static threshold for deploying the multicast communication data rate, whereas the second one employs a dynamic threshold depending on the network conditions.

Table 3.5: Summary of rate adaptation protocols

	Protocol	Threshold	Metric	Feedback
Unicast	ARF	static	tx failure	ACK
	AARF	dynamic	tx failure	ACK
	RBAR	static	SNR	RTS/CTS
Multicast	RAM	static	SNR	RTS/CTS, NACK
	ARSM	static	SNR	Channel probing
	LM-ARF	static	tx failure	Leader-based
	SARM	static	SNR, PSNR	Channel probing

Static Approach

We used QoE as indicator for switching from one transmission rate to another because it is more relevant to adapt the transmission rate taking into account the quality perceived at the end-user rather than other signal parameters. Also, as we explained earlier, the physical modulation plays an important role in such an environment and hence adapting modulation would help facing the bad condition.
Assuming PSQA running on every multicast client, Figure 3.8 illustrates the behavior of an access point in our scheme during a multicast session. At the beginning, the access point transmits multicast traffic at its highest rate. For communications between access point and mobile nodes, the IEEE 802.11k standard is used, which specifies many measurement requests and reports that are useful for our schemes. It can be noticed that with IEEE 802.11k measurements, our control traffic is less significant in terms of overhead as control traffic is sent much less frequently than other packet-level schemes. For example, control traffic can be sent every second in our scheme comparing to every single packet in the other packet-level schemes. An access point in our schemes initiates requests for the actual QoE to users at different timestamps at the beginning of monitoring period in the order of session joining. This is to avoid collision explosion of reports sent back by all users. With PSQA running on every station, users compute their QoE and return it to the access point. Thanks to this information, when condition changes, the access point adapts its transmission rate accordingly. The AP monitors its attached clients every monitoring interval (m_i). Note that our scheme uses time scale in order of a second, which is more reasonable than using a per-packet scale when dealing with human perception. When the timer expires, the AP begins by sending requests to multicast members in order of membership precedence, which avoids collision of reports sent back from members. When a report is received, AP updates the minimum MOS (min) of the group accordingly. Once the last report has been received, it compares min with the lower bound (lb). This lb is computed by adding a margin (mg) to a reference score (rf), which is an acceptable score for the application. If min is less than lb, then AP switches immediately to one-step lower rate until minimum rate. If min is higher than lb, then AP increases the counter (representing the duration that AP has been waiting). If the counter reaches a threshold (th), then AP switches to one-step upper rate until maximum rate.

It can be noticed that when condition degrades, the AP decreases the transmission rate immediately. This is to adapt instantly to bad conditions because it is essential to recover from the bad situation rapidly. When network condition becomes better (i.e. min is higher than lb) for a certain number of times, the AP switches to higher rate. This

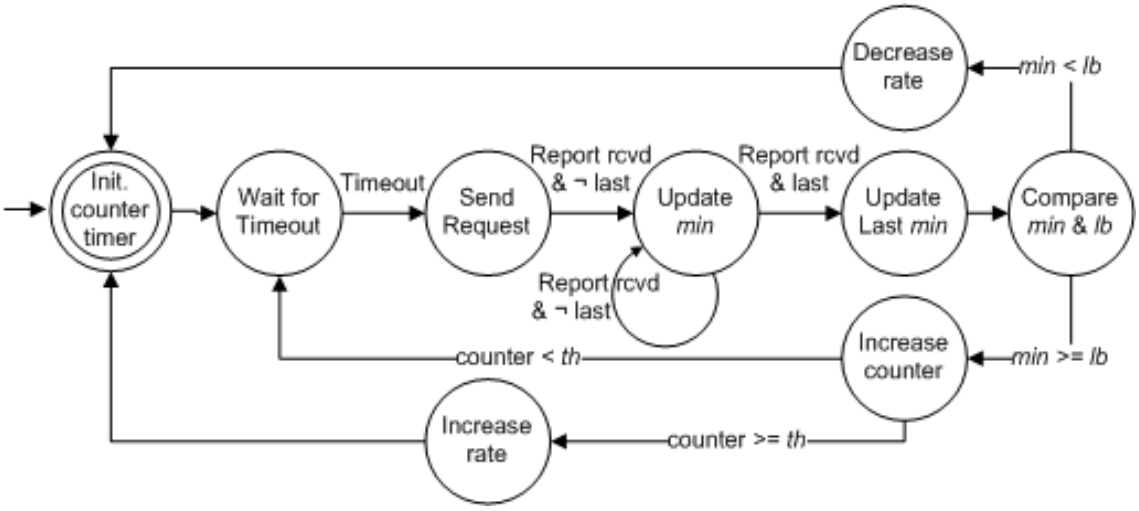

Figure 3.8: Access point behavior during multicast session

waiting threshold is used to avoid a ping-pong effect; before sending at higher rate (higher risk of BER), we make sure that this condition remains quite stable.

Dynamic Approach

We have noticed from previous works that all proposed schemes use a static-threshold rate adaptation for multicast, and none of them has considered dynamic threshold adaptation; the parameters are number of transmission failure or SNR as shown in Table 3.5. The problematic issue associated with static approach is the adaptability to the network condition fluctuation, which is common in wireless environment.

In order to overcome different limitations and to adapt to the environment and user perception dynamically, we propose QoE-based Dynamic Rate Adaptive Multicast (Q-DRAM). It is a novel mechanism that dynamically adjusts the transmission rate according to end-user perception terms of QoE. Although it is similar to the static approach because it also uses QoE as metric, this mechanism adds dynamic adaptability since it can adjust to network condition dynamically.

The most important decision to make in rate adaptation is mainly on how long to wait (backoff) before changing rate.

- For switching down, the decision is quite simple because we do not want to stay in bad situation so we switch rapidly to a lower rate. From the literature, there are two causes for switching down. The first one is failure due to varying network condition; this is naturally happened when network condition changes due to mobility, interference, etc. In this case, the sender should wait for two consecutive failures before switching down in order to avoid constantly changing rate up and down (ping-pong effect). The second cause is due to the action that the sender just took to switch to a higher rate; in this case of failure here, the sender switches immediately to the previous rate because the new rate appears to be too high. For D-QRAM, we decided to switch down immediately after both cases. Note that failure in this scheme occurs when QoE is less than a desired threshold.

- For switching up, we use dynamic-threshold strategy called binary *exponential backoff (BEB)* similar to AARF. This strategy allows us to adapt to varying network

condition. With BEB, we increase the backoff exponentially when failure occurs or repeats after the successful attempt of rate increase. It means that if the QoE is less than desired (*fail*) right after switching up (*just_up*), we switch down immediately and before trying to switch up again we wait longer by setting the backoff timer to be twice of the previous value. For the other case of failure (varying condition), we do not update the backoff stage. Figure 3.9 illustrates how BEB works in our scheme. At the beginning, the backoff timer is set to minimum value (*thMIN*). During multicast session, it will be reset to *thMIN* again after a successful attempt of rate increase. The backoff timer cannot exceed *thMAX*. Therefore, the backoff timers corresponding to each stage in Q-DRAM are {0:1; 1:2, 2:4}.

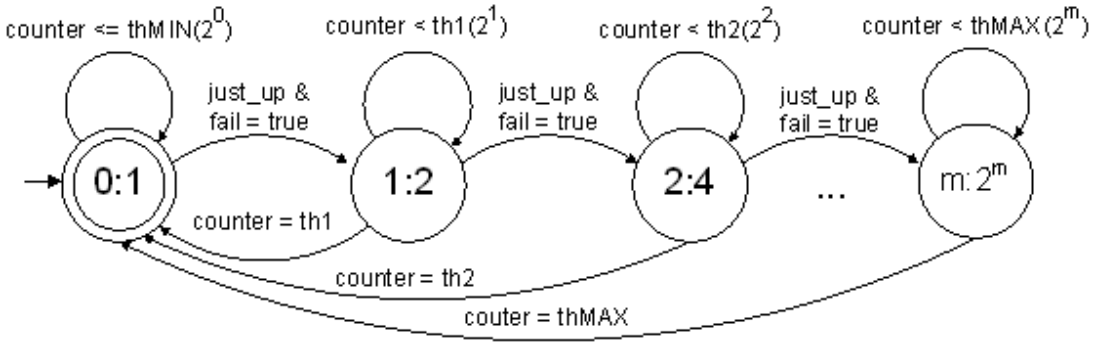

Figure 3.9: Binary exponential backoff in Q-DRAM

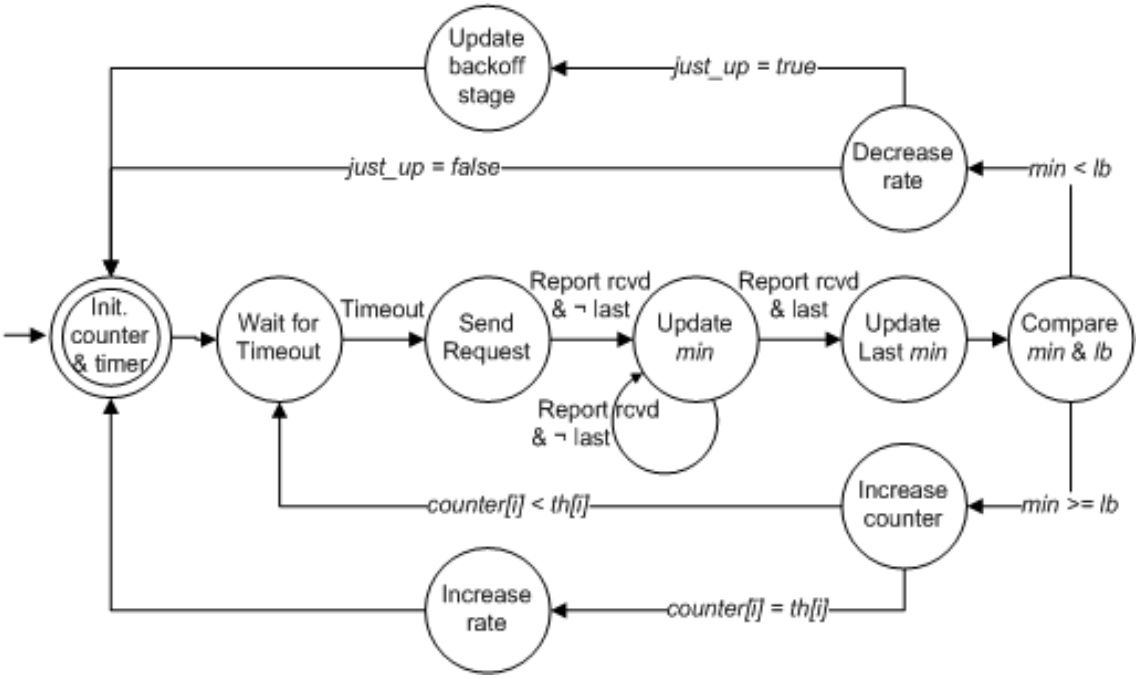

Figure 3.10: Access point behavior during multicast session

We describe in Fig. 3.10, the behavior of an AP during a multicast session. We assume PSQA running on every multicast node. At the beginning, the access point transmits multicast traffic at its highest rate. The AP monitors its ongoing clients every monitoring

interval (m_i) in unit of second. When the timer expires, AP begins by sending requests to multicast members in the order of membership precedence. This helps to avoid collision of reports sent back from members. When a report is received, AP updates the minimum MOS (min) of the group accordingly. Once the last report has been received, it compares min with the desired QoE called as lower bound (lb), computed as in our static scheme. If min is less than lb, then AP switches immediately to one-step lower rate until minimum rate; in case of unsatisfied QoE just after rate increase, the backoff stage is updated. If min is higher than lb, then AP increases the counter (representing the duration that AP has been waiting); if the counter reaches a threshold (th_i) where i is backoff stage, then AP switches to one-step upper rate and the backoff stage is reset.

3.5 Summary of results

To conclude, we can summarize our representative research works relaying on human-oriented contributions as follows. We built a new automatic QoE measurement tool for SVC. After describing the parameters that affect the quality of SVC streams and their relationship with QoE, we explained how to capture this non-linear relationship with PSQA tool. The results showed that our QoE estimation is highly accurate and the obtained scores are close to those given by real users. This is reflected by the overall mean square error of about 0.1777 on the MOS scale from 0 to 5. In addition, we used the Pearson Correlation Coefficient (PCC) as recommended by the Video Expert Group (VQEG) [49] to validate the objective models for video quality assessment. The mean PCC for the tested video is equal to 0.9406, which represents a good result as the highest accuracy obtained with a PCC equals to one. To the best of our knowledge, this automatic QoE tool measurement is the only one that considered packet loss as a parameter to evaluate user perception for SVC encoded videos.
We evaluated the concept of associating DVB-T2 with SVC in order to broadcast high added-value video services to end-user. Based on the above mentioned SVC QoE tool we conducted measurements, which showed that decoding all SVC layers is not always the best way to enhance QoE, especially when SVC enhanced layers experience packet losses. To improve user QoE, we demonstrated that, in some cases, it is better to withdraw an enhanced layer that experiences packet losses rather than decoding its content. We considered this observation when designing our proposed QoE-BASD mechanism that dynamically selects the SVC layers to be decoded and displayed to end-users in order to increase QoE. We modeled QoE-BASD using a Time Discrete Markov chain model, and validated the analytical results using computer simulations. The obtained results clearly demonstrated the enhancements achieved by QoE-BASD as it improves user QoE under different channel conditions.
Another contribution that showed the efficiency of using QoE for controlling network resources is the one that addressed the case of multicast communications in WLAN. We have seen that using the default basic rate for multicast transmission (conservative approach), has disadvantages not only in network utilization but also in quality perception at the users. Deploying the maximum rate achieves better performance if channel quality is good, however when the channel condition degrades this high rate leads to poor network performance. Unlike the existing schemes that used SNR as the main metric for changing rate, we have proposed a novel mechanism using QoE to trigger rate adaptation in wireless multicast. Our scheme is dynamic, in a sense that it can adapt to varying wireless environment. By using simulation, we compared Q-DRAM approach to 1Mbps fixed-rate multicast (conservative approach), 11Mbps fixed-rate (throughput maximizing

approach), and SARM-like mechanism (user-perception approach). Thanks to PSQA, Q-DRAM achieved the maximum MOS scores despite the considered network configurations (in term of loads and user mobility). In addition, Q-DRAM successfully achieved a trade-off between maximizing the network resources utilization (it uses the maximum rate since it is possible), and increasing user QoE.
QoE metric contributions were conducted as part of the PhD program of Wael Cherif, and in collaboration with Kamal Deep Singh (Post-doc at Dionysos team). Human-centric contributions were done as part of the PhD program of Kandaraj Piamrat and Majd Ghareeb, and in collaboration with Jean-Marie Bonnin (Professor at Télécom Bretagne) and César Viho (Professor at the University of Rennes 1).

Bibliography

[1] W. Cherif, A. Ksentini, D. Negru, M. Sidibé, "A_PSQA: PESQ-like non-intrusive tool for QoE prediction in VoIP services", in proc. of IEEE ICC 2012 CSSMA Symposium, The IEEE Conference on Communication 2012, Ottawa, Canada.

[2] W. Cherif, A. Ksentini, D. Negru, "No-reference Quality of Experience estimation of H264/SVC stream", in Proc. IEEE QoEMC 2012 Workshop helds in conjunction with IEEE Globecom 2012. Anaheim, USA.

[3] K. Singh, A. Ksentini and B. Marienval, "Quality of Experience measurement tool for SVC video coding", in Proc. of IEEE International Conference on Communications (ICC), Kyoto 2011.

[4] A. Ksentini, Y. Hadjadj-Aoul, "QoE-based energy conservation for VoIP applications in WLAN", ,Book Chapter in energy Scavenging and Optimization Techniques for Mobile Devices and Networks, published by CRC Press, Taylor and Francis Group, USA.

[5] A. Ksentini, Y. Hadjadj-Aoul, "QoE-based energy conservation for VoIP over WLAN", in Proc. of IEEE WCNC 2012, The IEEE Wireless Communications & Networking Conference (WCNC), Paris, France.

[6] K. Piemrat, A. Ksentini, C. Viho, J-M. Bonnin, "Rate Adaptation Mechanism for Multimedia Multicasting in Wireless Networks", in Proc. of BROADNETS 2009, Sixth International Conference on Broadband Communications, Networks, and Systems, Madrid, Spain.

[7] K. Piemrat, A. Ksentini, C. Viho, J-M. Bonnin, "Q-DRAM: QoE-based Dynamic Rate Adaptation Mechanism for Multicast in Wireless Networks", in Proc. of IEEE GLOBECOM 2009 Wireless Networks Symposium, The 52th IEEE Global Telecommunication Conference. Hawaii, USA.

[8] K. Piamrat, A. Ksentini, C. Viho, J-M. Bonnin, "QoE-based Network Selection for Multimedia Users in IEEE 802.11 Wireless Networks", , in Proc of IEEE LCN 2008, The 33rd Conference on Local Computer Network, Montreal, Canada.

[9] K. Pimrat, A. Ksentini, C. Viho, J-M. Bonnin, "QoE-aware Vertical Handover in Wireless Heterogeneous Networks", In Proc. of IEEE IWCMC, The International Wireless Conference on Mobile Communications, Istanbul, Turkey 2011.

[10] T. Taleb, A. Ksentini, F. Filali, "Wireless Connection Steering for Vehicles", , In Proc. of IEEE Globecom 2012 Adhoc and Sensor Networks Symposium, The 55th IEEE Global Telecommunication Conference 2012. Anaheim, USA.

[11] K. Piamrat, K. Singh, A. Ksentini, C. Viho, J-M Bonnin, "QoE-aware scheduling for video-streaming in High Speed Downlink Packet Access", in Proc. IEEE WCNC 2010, The IEEE Wireless Communications & Networking Conference (WCNC 2010), Sydney, Australia,

[12] A. Ksentini and Y. Hadjadj-Aoul, "On associating SVC and DVB-T2 for Mobile Television Broadcast", in Proc. IEEE Globecom 2011 Wireless Networks Symposium, The 54th IEEE Global Telecommunication Conference. Houston, USA.

[13] A. Ksentini, T. Taleb, "QoE-Oriented Adaptive SVC Decoding in DVB-T2", in IEEE Transactions on Broadcasting, Accepted in March 2013.

[14] M. Ghareeb, A. Ksentini, C. Viho, "An adaptive QoE-based multipath video streaming algorithm for Scalable Video Coding (SVC)", in Proc. of IEEE ISCC 2011, The 16th Symposium on Computers and Communications, Corfou, Greece 2011.

[15] M. Ghareeb, A. Ksentini, C. Viho, "A Multipath Video Streaming Approach For SNR Scalable Video Coding (SVC) In Overlay Networks", in Proc. IEEE CCNC special session on IPTV and CDN, The IEEE Consumer Conference on Networking and Communication 2011, Las Vegas, USA.

[16] ITU-T SG12, "Definition of Quality of Experience", COM12 – LS 62 – E, TD 109rev2 (PLEN/12), Geneva, Switzerland, Jan. 2007.

[17] ITU-T E.800, "Telephone Network and ISDN Quality of Service, Network Management and Traffic Engineering: Terms and Definitions Related to Quality of Service and Network Performance Including Dependability", Aug. 1994.

[18] B. Ciubotaru, G-M. Muntean and G. Ghinea, "Objective Assessment of Region of Interest-Aware Adaptive Multimedia Streaming Quality", in IEEE Trans. on Broadcasting, Vol. 55, No. 2, Jun. 2009.

[19] S. Chikkeur, V. Sundaram, M. Reisslein and L-J. Karam, "Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison", in IEEE Trans. on Broadcasting, Vol. 57, No. 2, Jun. 2011.

[20] M-H. Pinson and S. Wolf, 'A New Standardized Method for Objectively Measuring Video Quality", in IEEE Trans. on Broadcasting, Vol. 50, No. 3, Sep. 2004.

[21] S. Winkler, A. Sharma, and D. McNally, "Perceptual video quality and blockiness metrics for multimedia streaming applications", in Proc. Int'l Symp. on Wireless Personal Multimedia Communications, Aalborg, Denmark, Sep. 2001.

[22] A. Ichigay, M. Kurozumi, N. Hara, Y. Nishida and E. Nakasu, "A Method of Estimating Coding PSNR using Quantized DCT Coefficients", in IEEE Trans. on CSVT, Vol. 16, No. 1, Feb. 2006.

[23] ITU-T G. 1070, "Opinion model for video-telephony applications," Apr 2007.

[24] K. Yamaghachi and T. Hayashi, "Parametric Packet-Layer Model for monitoring Video Quality of IPTV services", In Proc. of IEEE ICC, Beijing, China, Jun. 2008.

[25] F. You, W. Zhang, and J. Xiao, "Packet Loss Pattern and Parametric Video Quality Model for IPTV," in Proc. of IEEE Int'l Conf. on Computer and Information Science, Shanghai, China, Jun. 2009.

[26] S. Mohamed and G. Rubino, "A study of Real-Time Packet Video Quality using Random Neural Networks", in IEEE Trans. on Circuits and Systems for Video Technology, Vol. 12, No.12, Dec. 2002.

[27] W. Cherif, A. Ksentini, D. Negru and M. Sidibe, "A_PSQA: Efficient Real-time Video Streaming QoE Tool in a future media Internet Context", in Proc. of IEEE Int'l Conf. on Multimedia and Expo (ICME), Barcelona, Spain, Jul. 2011.

[28] G. Auwera, P-T. David, and M. Reisslein, "Traffic and Quality Characterization of Single-Layer Video Streams Encoded with H.264/MPEG-4 Advanced Video Coding Standard and Scalable Video Coding Extension", in IEEE Trans. on Broadcasting, Vol. 54, No. 3, Sep. 2008.

[29] JSVM, URL: http://ip.hhi.de/imagecom_G1/savce/downloads/SVC-Reference-Software.htm.

[30] OpenSVC, URL: http://sourceforge.net/projects/opensvcdecoder.

[31] ITU-R BT.500-11, "Methodology for the subjective assessment of the quality of television pictures", Jun. 2002.

[32] S. Mohamed and G. Rubino, "A study of Real-Time Packet Video Quality using Random Neural Networks", in IEEE Trans. on Circuits and Systems for Video Technology, Vol. 12, No.12, Dec. 2002.

[33] "DVB-T2 – the HDTV generation of terrestrial DTV", DVB-TM-3997, Mar. 2008.

[34] M-J. Rahman, W. Yiyan, T. Bin, Y. Kechu and J.-Y. Chouinard, "A Time Slicing Adaptive OFDM System for Mobile Multimedia Communications", in IEEE Trans. on Broadcasting, Vol. 56, No. 2, Jun. 2010.

[35] J-Y. Chouinard, A. Semmar, W. Xianbin and W. Yiyan, "On the Channel and Signal Cross Correlation of Downlink and Uplink Mobile UHF DTV Channels With Antenna Diversity", in IEEE Trans. on Broadcasting, Vol. 56, No. 2, Jun. 2010.

[36] T. Taleb and K. Hashimoto, "MS2: A Novel Multi-Source Mobile-Streaming Architecture," in IEEE Trans. on Broadcasting, Vol. 57, No. 3, Sep. 2011.

[37] L. Kondrad, I. Bouzizi, V. K. M. Vadakital, M.M Hannuksela, M. Gabbouj, "Cross-Layer Optimized Transmission of H.264/SVC streams over DVB-T2 Broadcast System", in Proc. of IEEE Int'l Conf. Symp. on Broadband Multimedia Systems and Broadcasting (BMSB'09), Bilbao, Spain, May 2009.

[38] S. Wenger, Y. Wang and T. Schierl, "RTP Payload Format for SVC Video", IETF Internet Draft, draft-ietf-avt-rtp-svc-20.txt, Dec. 2009.

[39] ETSI TS 102 606 V.1.1.1, "Digital Video Broadcasting (DVB); Generic Stream Encapsulation".

[40] M. Heusse, F. Rousseau, G. Berger-Sabbatel, and A. Duda. "Performance anomaly of 802.11b", in proc. INFOCOM, Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies, 2:836–843 vol.2, March-3, April 2003.

[41] A. Kamerman and L. Monteban, "WaveLAN-II: A high-performance wireless LAN for the unlicensed band", Bell Labs technical journal, 2(3), 1997.

[42] M. Lacage, M-H. Manshaei, and T. Turletti, "IEEE 802.11 rate adaptation: a practical approach", in Proc. MSWiM the 7th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems, pages 126–134, New York, NY, USA, 2004.

[43] G. Holland, N. Vaidya, and P. Bahl, "A rate-adaptiveMAC protocol for multi-Hop wireless networks", in proc. MobiCom, the 7th annual international conference on Mobile computing and networking, pages 236–251, New York, NY, USA, 2001

[44] A. Basalamah, H. Sugimoto, and T. Sato, "A Rate AdaptiveMulticast Protocol for Providing MAC Layer Reliability in WLANs", in IEICE Transactions on Communications, E89-B(10):2733–2740, 2006.

[45] S. Choi, N. Choi, Y. Seok, T. Kwon, and Y. Choi, "Leader-Based Rate Adaptive Multicasting for Wireless LANs", in proc. IEEE Global Telecommunications Conference, 2007. GLOBECOM '07, pages 3656– 3660, Nov. 2007.

[46] J. Villalon, P. Cuenca, L. Orozco-Barbosa, Y. Seok, and T. Turletti, "ARSM: a cross-layer auto rate selection multicast mechanism for multi-rate wireless LANs", in IET Communications, 1(5):893–902, Oct. 2007

[47] Y. Park, Y. Seok, N. Choi, Y. Choi, and J.-M. Bonnin, "Rate-adaptive multimedia multicasting over IEEE 802.11 wireless LANs", in proc. of 3rd IEEE Consumer Communications and Networking Conference, 2006. CCNC 2006., 1:178–182, 8-10 Jan. 2006.

[48] IEEE Computer Society. IEEE 802.11k, Part 11: Wireless LANMedium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 1: Radio Resource Measurement of Wireless LANs. IEEE Std 802.11k-2008 (Amendment to IEEE Std 802.11-2007), pages c1–222, 12 June 2008.

[49] Video Quality Experts Group, "Validation of Objective Models of Video Quality Assessment, Phase II", Aug. 2003.

[50] F. Bari and V. Leung, "Multi-Attribute Network Selection by Iterative TOPSIS for Heterogeneous Wireless Access," in Proc. IEEE Consumer Communications and Networking Conference, Las Vegas, NV, USA, Jan. 2007.

Chapter 4

Conclusion

My works on WLAN were highly motivated by the ineffectiveness of the IEEE 802.11 standard and its extension 802.11e to guarantee QoS for emerging added-value applications like VoIP, video streaming, IPTV and network gaming. In this context, we proposed several contributions to improve the 802.11 QoS oriented procedures: intra and inter TC service differentiation, admission control, efficient support of multicast communication, etc. Our control-oriented approaches have the originality to be dynamic and adaptive to the wireless network dynamic (contention level, channel conditions) as well as to service application requirement (delays, bandwidth, QoE). Besides enhancing the user experience, our contributions were also beneficial for the network operator, where we were able to efficiently optimize the network resources for a better QoS support. Our contributions were based on both network-centric and human-centric information to control network mechanisms. By using human-centric information we were capable to consider user experience rather than just technical QoS information, which represents one of the novel contribution in this field of QoS support in WLAN.
Now, the 802.11-based WLANs are mature and gained significant support from industrials and academics to update and build new standards for increasing the capacity of WLAN to support added-value applications, which impose strict QoS guarantee. Indeed, the IEEE 802.11 group has launched recent activities through new working groups, such as:

- IEEE 802.11aa: this group is dedicated to efficiently support video streaming and multicast communication in WLAN,
- IEEE 802.11ad [1]: this group aims to define new MAC and physical protocols to support the very high throughput.

Accordingly, I will concentrate more on LTE-based cellular networks where there are many challenges to address, especially in the EPC part. Indeed, mobile traffic is increasing at a tremendous pace, exceeding far beyond the original capacities of mobile operator networks. This huge mobile traffic is associated with a wide plethora of emerging bandwidth-intensive mobile applications popular among an ever-growing community of mobile users. Besides the non usual traffics generated by MTC and social networks, current mobile networks are highly centralized, leading to high demand on central locations due to "backhauling" of all data traffic, to dramatic increases in bandwidth requirements and processing load resulting in undesirable bottlenecks, and last but not least, to long communication paths between users and servers. The effects are wasting core network resources, leading to undesirable delays, and ultimately resulting in poor QoE for the users.
Meanwhile, I will pursue working on the QoE monitoring techniques and their use to control network mechanisms, where I am particularly interested to work on the QoE prediction

techniques.
In the following I will detail some research perspective on the above-mentioned fields.

4.1 QoE prediction and its use for controlling network mechanisms

Most of our human-centric oriented contributions are based on the QoE information measured at a specific moment (the current user experience) with PSQA. But, what if we can predict the future QoE values?
In fact, QoE prediction will allow the network operator to have a better view in time of the evolution of QoE, and hence to better allocate the network resources. We begun exploring this principle of QoE prediction in [2], where we proposed a model of prediction based on the QoS log information of a network operator. Thanks to this model, we proposed an admission control between Femtocell and Macrocell, which aims to maximize user QoE. However, this work concentrates further on the benefits of using QoE prediction for enforcing admission control algorithm, than on the details of the prediction mechanism. Indeed, we used a static model which cannot adapt to the network dynamics. Therefore, there is a gap to fill in order to develop an accurate QoE prediction model based on PSQA.

4.2 Mobile Cloud

A straightforward solution to the centralized issue of the cellular networks may consist in having operators invest in speed or upgrading their core network nodes to comfortably accommodate traffic peak hours of the emerging bandwidth-intensive mobile applications. Whilst this is technically and technologically possible, it economically represents a significant challenge for operators, particularly due to the fact that the Average Revenue per User (ARPU) is not growing as quickly as traffic demand, particularly given the trend towards flat rate business models. There has been thus need for cost-effective solutions that can help operators accommodate such huge mobile network traffic while keeping additional investment in the mobile infrastructure to the minimum. In addition to application type-based traffic admission control techniques (e.g., throttling video traffic), an important solution consists in Selective IP Traffic Offload (SIPTO) as close to the Radio Access Network (RAN) as possible [3]. The key enabler of efficient SIPTO is to place data anchor and mobility gateways close to RAN, essentially leading to a relatively decentralized mobile network deployment.
On the other hand, cloud computing is gaining a great momentum. Its market is expanding at a high speed, thanks to the multiple features it supports (e.g., multi-tenancy support, pay as you go, elasticity and cost-efficient scalability) and the new business models it provides based on infrastructure sharing (Infrastructure-, Platform-, Software-as a Service – IaaS, PaaS, and SaaS). In the telecommunications area, cloud computing has been gaining lots of attention. Indeed, there are already many Telcos and carrier providers deploying cloud-based services; some deployments are only for internal use whereas others are being sold as a service. The fast growing business of clouding computing is calling for distributed regional Data Centers (DC) [4][5].
There are mainly two perspectives which motivate us to explore the concept of using cloud in the context of cellular networks: Follow Me Cloud (FMC) concept and EPC as a Service.

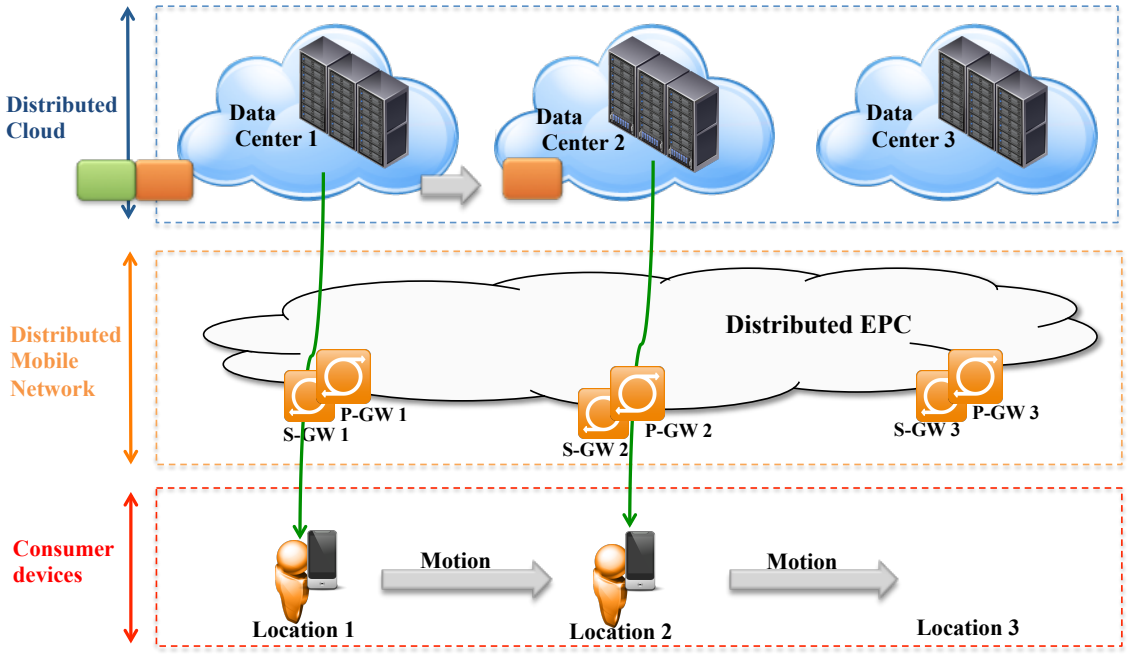

Figure 4.1: FMC concept.

4.2.1 Follow Me Cloud (FMC)

Actually, cloud providers are distributing their DCs due to growing business. As for mobile operators, they need to decentralize their networks to cope with the growing number of smart phones and associated bandwidth-intensive services. The expected outcome network architecture is as depicted in Figure 4.1. Indeed, the figure shows the case of a decentralized mobile network architecture whereby core network gateways such as PDN-GWs and S-GW, in the context of the EPS, are geographically distributed. Also shown is a distributed cloud network consisting of multiple regional DCs, which are geographically distributed and interconnected.

In such decentralized mobile networks, the main objective of any mobile operator behind SIPTO is to ensure an optimal mobile connectivity service; i.e., a User Equipment shall be always connected to the optimal data anchor and mobility gateways such as PDN-GWs and S-GWs. However, it is very likely to have a UE connected to an optimal data anchor gateway (as per its current location) but accessing a mobile service from a distant DC in a distant location (e.g., UE being in location 2, having its data anchored at P-GW2, but receiving service from DC 1). This intuitively results in an inefficient "mobile connectivity service" given the absence of an optimal end-to-end connectivity. The objective of the FMC concept is to enable a user to be always connected to the optimal data anchor and mobility gateway and to access its data and/or service from the optimal DC, i.e., geographically/topologically nearest (or any other metric such as load, processing speed) DC.

There are several interesting issues to study before that FMC concept will be a reality:

- IP Service continuity during UE mobility. It is important to maintain the service continuity, as the user is mobile, and can change the anchor PDN-GW and hence the IP address during its movement.

- Service migration algorithm, where it is important to know the criterion used to decide to migrate or not a service from one DC to another one.
- Use the concept of Software Defined Networking (SDN) in order to implement the FMC concept

4.2.2 EPC as a Service

One solution to reduce the congestion at the cellular core nodes is by virtualizing the EPC nodes like MME, PDN-GW and S-GW. In fact, virtualizing EPC nodes permits the network operator to create or destroy a virtual instance of the network on demand. For instance, it can increase the number of gateway nodes when the traffic grows.
This virtualization is now implemented with SDN tools, where it is possible to host the virtualized EPC throughout the distributed regional DC of a cloud operator. Thus, it will be easy to set up a network node on demand and in a distributed way, implementing thus the concept of "EPC as a service". Besides reducing operator cost (no need to invest in more infrastructure elements), this virtualization can help to reduce congestion at the EPC part.
We are interested in studying the concept of EPC as a service to build tools for network dimensioning. These tools could be based on simulation or analytical methods, where the aim is to establish relation between, for instance, the number of gateway to create and an offered amount of traffic.

4.3 Small Cell Network (SCN)

SCN is a new architectural model for cellular networks, where the principle is to use base stations (picocell, microcell, femtocell) with low radio coverage (about few meters). Associated to the classical macrocell, SCN increases the capacity of a network operator to handle more traffic and avoid congestion in the EPC (by offloading mobile traffic through the SCN). Thanks to their reduced cost, SCN constitute an interesting opportunity to extend the network operator coverage for regions where it is not profitable to invest in network infrastructure (due to economic or technical reasons).
Due to their flexibility and the low cost deployment, SCNs have a strong potential to be used for connecting areas like those in the emerging countries, where it is important to propose solutions that: (i) minimize the cost of deployment and maintenance; (ii) minimize the energy consumption or use other alternative energy sources.
Deploying SCN while considering the above constraints introduces several issues, such as routing between the SCNs (there is no backhaul network), forcing the non active BS to enter the sleep mode, sharing the infrastructure among the operators; which still are untreated by the research community.
Rather than using SCN for offloading traffic, we intend to use the SCN concept as a solution for increasing the operator coverage in areas where investing for costly network infrastructure is not profitable.

Bibliography

[1] IEEE p802.11ad/Draft9.0 2013, modification to both the 802.11 physical layers (PHY) and the 802.11 Medium Access Control Layer (MAC) to enable operation in the 60 Ghz frequency band (typically 57-66 Ghz) capable of very high throughput.

[2] T. Taleb and A. Ksentini,"QoS/QoE Predictions-based Admission Control for Femto Communications", in Proc. of IEEE ICC 2012 Wireless Networks Symposium, The IEEE Conference on Communication 2012. Ottawa, Canada.

[3] 3GPP TR 23.829, "Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO)".

[4] R. Miller, "AOL Gets Small with Outdoor Micro Data Centers," Data Center Knowledge, Jul. 2012.

[5] R. Miller, "Solar-Powered Micro Data Center at Rutgers," Data Center Knowledge, May 2012.