Università degli Studi di Cagliari

PHD DEGREE

Electronic and Computer Engineering (DRIEI)

Cycle XXXI

TITLE OF THE PHD THESIS

QoE-aware Multimedia Service Management and Monitoring through OTT and ISP Collaboration

Scientific Disciplinary Sector(s)

ING-INF/03

PhD Student: Arslan Ahmad

Coordinator of the PhD Programme: Prof. Fabio Roli

Supervisor: Prof. Luigi Atzori

Final exam. Academic Year 2017 – 2018

Thesis defence: January-February 2019 Session

QoE-aware Multimedia Service Management and Monitoring through OTT and ISP Collaboration

Arslan Ahmad

Supervisor: Prof. Luigi Atzori

Department of Electrical and Electronic Engineering
University of Cagliari, Italy

This dissertation is submitted for the degree of
Doctor of Philosophy

January 2019

I would like to dedicate this thesis to my family and friends who have been motivating me to pursue and accomplish my goals as well as ambition throughout the life. I would like to dedicate this work my doctorate and master degree supervisors and mentors Prof. Luigi Atzori and Dr. Atif Bin Mansoor whom timely guidances have helped me throughout my life in order to accomplish this all.

Acknowledgements

First, I would like to prompt my inner, profound and sincere gratitude to my mentor and supervisor, Prof. Luigi Atzori. I will always do reverence and reminisce to his fatherly behavior, sympathetic attitude, perceptive pursuit, rational supervision, cheering perception and the most that I like his scholarly criticism. Actually, I had zilch knowledge about Quality of Experience and multimedia communication; the effort of Prof. Luigi Atzori changed me from crud coil to diamond. I'll never forget sampling tours under his organization. During this bond of more than three years under his supervision, I have improved my technical knowledge and understanding in the field of QoE Monitoring and Management.

Secondly, I would like to acknowledge the European Union Horizon 2020 Marie Curie initiative for researchers training and funding which indeed provides a significant contribution in my profile to emerge as a young researcher. All the research done during my doctorate is funded by QoE-NET MSCA ITN project. The young researchers' training under QoE-NET project not only assist me a lot in the technical domain but also in my personal life.

Lastly, I am grateful to my family and friends who motivated me throughout my life.

Abstract

Internet traffic is emerging as multimedia traffic nowadays because of the rising popularity of the Over-The-Top services (OTTS) which includes YouTube, Netflix, Facebook etc.. The drastically rising trend in the multimedia consumption over Internet by the end users combined with the competitive market and quality sensitive users require Internet Service Providers (ISPs) and OTTs to incorporate user perception in the multimedia service delivery chain. Necessarily, this will not only require the novel techniques for the network/application service monitoring and management but a collaborative network and application monitoring and management. Quality of Experience (QoE) is gaining researchers interest in multimedia service management and monitoring because of multidimensional aspects and being more user-centric. Thus, in order to incorporate user's perception in the multimedia monitoring and management, the inclusion of the QoE in multimedia service monitoring and management remains significant. As the OTT and ISP have different roles in multimedia service delivery, the work in this thesis focuses on collaborative QoE monitoring and management of multimedia services. The work in the thesis is composed of two parts: I) QoE Monitoring— state-of-the-art monitoring techniques, future challenges and opportunities regarding QoE monitoring are highlighted. Moreover, user terminal based multilayered QoE monitoring solutions are also proposed; II) QoE Management—collaborative QoE management strategies and reference architecture to drive these collaborative management solutions are proposed and evaluated. Moreover, the implementation of the Software-Defined Networking (SDN) based collaborative QoE management platform built during the doctoral course is also discussed. The results of the collaborative QoE monitoring and management highlights the effectiveness of the proposed methodologies.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Today Internet traffic is emerging as multimedia traffic due to popular Over-The-Top services (OTTs) such as YouTube, Netflix, and Facebook. Among all the multimedia services, the video streaming is the most popular with the predicted video traffic of 82% of all consumer Internet traffic by 2021 from 73% in 2016 [2]. Such a huge amount of traffic in combination with the the most of the time adopted best effort approach for multimedia service delivery of the OTT services requires on the part of the Internet Service Providers (ISPs) to implement novel network management techniques. By improving the network management techniques, the network congestion can be avoided (which may bring higher user churn rate) and better service quality can be delivered [3]. Indeed, the provision of the better quality to the end user is the ultimate goal of both ISP and OTTs in order to avoid the user churn and increment market share as well as reputation.

Quality of Experience (QoE) is the next generation evolution of the term Quality of Service (QoS). However, the latter is only technology-centric while the former is user-centric. Recent research reveals QoE as a multidimensional construct based on several influencing factors such as human, economic, network, application, and context [4]. Being user-centric, QoE is defined as *"the degree of delight or annoyance of the user of an application or service. It results from the fulfilment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state"* [4], [5]. Thus, the research paradigm in the area of multimedia service management and monitoring is shifting from the system centred QoS based Key Performance Indicators (KPIs) to user-centred QoE based Key Quality Indicators (KQIs) [6]. Therefore, the novel QoE oriented techniques for the multimedia service management/monitoring are required where the collaborative management of network and application is needed. QoE management and monitoring require the inclusion of the multidisciplinary QoE-aware concepts for the multimedia service management and monitoring.

QoE monitoring of the multimedia service may be considered the first step towards the QoE-aware Internet service delivery. QoE-aware multimedia service management requires the continuous QoE monitoring of the KQIs and KPIs in order to have a better view of the real quality perceived by the users as well as effective QoE management of the network and application. Considering the multidimensional aspect of the QoE, QoE requires the continuous monitoring of the KQIs representing different critical influencing factors such as user profile, context, system, application and network.

Lately, the research in the field of QoE has been conducted separately with the different prospects of the OTTs and the ISPs which gave birth to two different areas in the field of QoE: application-aware networks and network-aware applications. The network-aware applications [7] aim to adapt the delivery of multimedia contents on the best effort over the network by inducing the change in the application parameters, whereas application-aware networks [8] focus on effective management of network according to application requirements. However, the drawback of the above-mentioned research stream is that OTTs have no control over the network for enhancing users' QoE, whereas ISPs have neither access to application QoE models, nor can they do deep packet inspection due to content encryptin and privacy issues [9]. Hence, neither the ISP nor the OTT cannot deliver the best QoE to their valued customers, which results in user churn as well as the decrease in market shares.

Therefore, the work in this thesis is composed of the collaborative QoE management and monitoring approaches and the evaluations of those approaches. The thesis is composed of two parts: I) QoE Monitoring and; II) QoE Management. This chapter is structured as follows: First, the motivation on the topic is provided in Section 1.1. Section 1.2 discusses the problem definition in details. Section 1.3 provides the list of publications while Section 1.4 provides the details of thesis structure.

1.1 Motivation: Why is OTT-ISP collaboration needed?

It is well known that the QoE for any service over the Internet not only depends on the network parameters but also on the application parameters [10]. Accordingly, as the OTTs applications are being delivered over the ISPs best-effort Internet without considering the resource requirement of the application leading to degraded quality perceived by the users which on longer time scale may bring significant user churn rate. However, generally, the ISP is the entity which suffers more from the user churn because the average user thinks that the poor QoE perceived is mainly due to low network resources and then changes ISP. Additionally, the user may decide to move from one operator to another. For this reason, OTTs are usually not willing to collaborate with ISPs as well as because they do not want to share any precious information about their application and users. Nonetheless, the OTTs may accept to collaborate if this collaboration allows them to increase their revenue, i.e., the network services provided by the ISP allow the OTT's users to perceive a better QoE so that the number of users of that OTT provided through that ISP increases together with OTT and ISP's revenue.

Therefore, the collaboration between ISPs and OTTs must require a common ground of motivation that we identify as the revenue. Especially during the last years, users are more quality demanding and fulfilment of the quality expectations may lead to the reduction in user churn which in turn increases the number of the customers, resulting in higher revenue for both the service providers. Hence, we propose a collaboration approach driven by the maximization of the revenue based on different factors such as the user churn (modelled as a function of the QoE), pricing and marketing actions. In the following subsections, we further discuss the technical issues that the collaboration is addressing.

1.1.1 Application-aware Traffic Engineering vs Encryption

The delivery of users' perceived quality is a big issue nowadays considering that different multimedia applications have unique requirements [11]. The proposed quality management approaches at the hands of the ISPs, such as DiffServ [12] and IntServ [13], have their limitation over best-effort Internet [14]. Though some past works highlighted to be application-oriented, such as in [15, 16], today, the OTT services are being encrypted with the concern of the user privacy issues. This is the case for example of YouTube that has been turned from HTTP to HTTPS, where the videos are now being transmitted in the encrypted sessions [17]. The traffic encryption is leading to a major challenge for application-aware QoE-based Internet service delivery as ISPs may not be able to either perform the Deep Packet Inspection (DPI) and packet marking in order to apply the core traffic engineering concepts such as packet prioritization, traffic shaping, admission control etc. for the multimedia traffic management [17]. Moreover, shortest path routing concept [18] cannot be applied to delay sensitive traffic. Furthermore, as the network resource requirements vary in accordance with the QoE model of the application, only an OTT may know the best application-aware QoE model according to the users' context of use.

1.1.2 Different Roles in QoE based Optimization and Control

As also highlighted in [19, 20], the optimization of the application rate simultaneously with the packet prioritization and error concealment techniques not only can save the network resources but can also increase the QoE of the multimedia streaming services significantly.

However, the optimization of the application parameters is in the hands of OTT-only while the ISP has control over the network resource usage. Therefore, the collaboration is required, which may not only results in saving the number of resources but will also lower down the user churn.

1.2 Problem Definition

The objective of QoE management is the optimization of end-user QoE while making efficient (current and future) use of network resources and maintaining a satisfied customer base (provider perspective) [9]. QoE management relies on accurate monitoring and prediction of the QoE perceived by the users, so that QoE-aware optimization and control techniques can be applied to the available resources to maximize the user's perceived QoE. In this section, the different perspectives of OTT and ISP are discussed, which are mainly influenced by their roles.

The ISP is the owner of the *last-mile* network through which the end-users are allowed to connect to the Internet and then subscribe to OTT services. Nowadays, users are more quality demanding so that the ISP has the difficult role in managing the enormous amount of traffic generated by OTT services (especially multimedia content) to provide adequate QoE to the users. Application level throughput and download time may be considered the most important network parameters for the delivery of current multimedia applications. The former is the actual transmission rate of the information when the end user utilizes an application or accesses a web content: it mainly depends on latency and packet losses, which typically increase with the distance between the end user and the server. Caching platforms, such as transparent caching and CDN, can improve the throughput by placing contents near to the end users. The download time represents the actual time of information download (web pages, contents), which mainly depends on the protocol efficiency (rules for information exchange between the end user and server) and it is also related to the previous parameter. To decrease the download time, techniques such as protocol optimization and web acceleration platforms may be used. Though, besides best-effort services, to improve the QoE the ISP typically implements QoS-based traffic management techniques, such as the DiffServ model, which is mainly based on packet prioritization and flow management [21]. Although these actions are effective for congested network paths, they are not enough to guarantee higher throughput as well as lower delay/packet loss/download time. To this, the transfer of the multimedia servers closer to the users seems to be one of the best solutions.

The ISP has a network-centered view of the quality provided and predicts the QoE on the basis of QoS-to-QoE models, which are a function of monitored Key Parameter Indicators

(KPIs), i.e.,

$$QoE^{ISP} = f_1(KPI_1, KPI_2, \cdots, KPI_a, \cdots, KPI_A), \quad (1.1)$$

which means that the QoE predicted by the ISP is a function over set of the KPIs related to the network performance. Examples of QoE-based network management approaches are provided in [8, 17]. Though, the QoE predicted with Eq. (1.1) does not represent the actual quality perceived by the users because the QoE depends on QoS factors as well as on application, human and context factors, which are not considered in QoS-to-QoE models [22]. Indeed, the ISP does not have direct access to users' devices, nor application information can be acquired by inspecting network traffic due to the data encryption applied by the OTTs. Even if the Deep Packet Inspection (DPI) worked for some of the OTT services, the human and context influence factors would be still missing, and more complex context-aware QoE models for specific applications cannot be integrated into the service delivery [23]. Therefore, the provision of high QoE for each type of service requires the ISP to adopt novel QoE-aware traffic management approaches (e.g., CDN, transparent Internet caching, Web acceleration platform) that must be aware of the type of traffic to be managed as well as of user- and context-related information.

On the other hand, the OTT has access to some types of user-related information having its application running on the users' devices, from which the OTT may even monitor the most important network KPIs (e.g., throughput, delay and packet loss) and application KQIs (e.g., stalling events and playout buffer status) as well as understanding about the context in which applications are used, such as the user position (acquired from the GPS sensor), the type of device, the type of Internet connection. User profiles are also known by the OTT because each user to subscribe to the OTT service must have its own account, which contains information such as service's price and preferences and in some cases also user characteristics, which can be useful to implement a better prediction of the user's perceived QoE. Although all this information is in the hands of the OTT, the OTT may be limited in providing a good QoE to the users because of issues which may occur at the client device, at the ISP's network or in the CDN. However, even if the OTT has no control on the ISP's network, different optimization techniques may be implemented by the OTT to optimize user's QoE, such as adaptation of application parameters, CDN selection strategies, and resource management at the cloud/server side. Then, we express the QoE predicted by the OTT as a function of B Key Quality Indicators (KQIs), which in turn for the most part depend on the network QoS and on CDN strategies.

$$QoE^{OTT} = f_2(KQI_1, KQI_2, \cdots, KQI_b, \cdots, KQI_B / QoS, CDN), \quad (1.2)$$

Examples of QoE-based application management approaches are rate adaption algorithms for HTTP Adaptive Streaming (HAS) [24] and CDN technologies for content-centric networking [25].

From Eqs. (1.1) and (1.2), it is clear that not only the OTT and ISP have different perspectives in terms of QoE measurements, but they also have different roles in the management and optimization of the delivered quality. However, optimal end-to-end QoE-based service delivery can only be possible if OTT and ISP collaborate and agree upon a common QoE-based management approach and sharing of the QoE-oriented information. Indeed, by acting independently, it is difficult for ISPs and OTTs to deliver adequate QoE to their customers. Therefore, a collaboration between the two providers seems to be the natural solution to provide a better QoE to their customers. Since the OTT is more QoE-oriented, its role in the collaboration may be the monitoring of the QoE perceived by the users. By sharing this information with the ISP, this may lead to QoE-aware network management. Also, by sharing the information regarding the type of multimedia services delivered over the network, and their respective QoS requirements, the ISP does not need to employ DPI techniques but can directly provide specific network resources as a function of the service to be delivered. Then, by following a collaborative service management paradigm, both the entities may overtake their limitations in QoE management: the ISP can manage its network in a QoE fashion and the OTT can rely on the ISP if better network resources are needed for service delivery.

1.3 List of Publications

During the doctoral course, eight research articles are published which includes three journal articles and five conference research papers. Section 1.3.1 and Section 1.3.2 list the journal publications and conference proceedings during the doctoral program respectively.

1.3.1 Journal Publications

1. Alessandro Floris, Arslan Ahmad, Luigi Atzori: QoE-aware OTT-ISP Collaboration in Service Management: Architecture and Approaches. ACM Transactions on Multimedia Computing Communications and Applications 01/2018; 14(2s)., DOI:10.1145/3183517

2. Werner Robitza, Arslan Ahmad, Peter A. Kara, Luigi Atzori, Maria G. Martini, Alexander Raake, Lingfen Sun: Challenges of future multimedia QoE monitoring for internet service providers. Multimedia Tools and Applications 06/2017;, DOI:10.1007/s11042-017-4870-z

3. Arslan Ahmad, Alessandro Floris, Luigi Atzori: QoE-centric Service Delivery: A Collaborative Approach among OTTs and ISPs. Computer Networks 09/2016; 110(9 December 2016):168-179., DOI:10.1016/j.comnet.2016.09.022

1.3.2 Conference Proceedings

1. Arslan Ahmad, Alessandro Floris, Luigi Atzori: Timber: An SDN based emulation platform for QoE Management Experimental Research. 10th International Conference on Quality of Multimedia Experience (QoMEX 2018), Sardinia; 05/2018, DOI: 10.1109/QoMEX.2018.8463387

2. Arslan Ahmad, Alessandro Floris, Luigi Atzori: Towards QoE Monitoring at User Terminal: A Monitoring Approach based on Quality Degradation. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting 2017, Cagliari; 06/2017, DOI:10.1109/BMSB.2017.7986131

3. Arslan Ahmad, Luigi Atzori, Maria G Martini: Qualia: A Multilayer Solution for QoE Passive Monitoring at the User Terminal. IEEE ICC 2017 Communications Software, Services, and Multimedia Applications Symposium (ICC'17 CSSMA). IEEE, Paris, France (2017); 05/2017, DOI:10.1109/ICC.2017.7997262

4. Arslan Ahmad, Alessandro Floris, Luigi Atzori: OTT-ISP joint service management: A Customer Lifetime Value based approach. IFIP/IEEE International Symposium on Integrated Network Management. IEEE, Lisbon; 05/2017, DOI:10.23919/INM.2017.7987431

5. Arslan Ahmad, Alessandro Floris, Luigi Atzori: QoE-aware Service Delivery: A Joint-Venture Approach for Content and Network Providers. Eighth International Workshop on Quality of Multimedia Experience (QoMEX) 2016, Lisbon; 06/2016, DOI:10.1109/QoMEX.2016.7498972

1.4 Thesis Structure

This thesis is composed of two parts: I) QoE Monitoring and; II) QoE Management. The chapters in the QoE Monitoring part provide the details of the state-of-the-art works and proposed QoE monitoring solutions while QoE Management part discusses the QoE management strategies and platform proposed in this thesis. Each chapter in this thesis represents the work done in the published articles where each chapter contains the abstract, introduction, relevant state-of-the-art work, proposed methodology, research finding. Moreover, the chapters are organized in a way that reader can read the chapter of his/her interest while skipping the rest. Chapter 9 provides the concluding remarks and the future works. The rest of this section provides the details of the thesis parts in details.

1.4.1 Part I– QoE Monitoring

This part of the thesis consists of three chapters from Chapter 2 to Chapter 4.

Chapter 2 provides a tutorial discussion on the QoE models and monitoring solutions. Moreover, the key future challenges and opportunities regarding QoE monitoring with the ISP's perspectives are investigated. Chapter 2 represents the work published in [6].

Chapter 3 focuses on the proposed multilayered QoE monitoring solution where the installation of the probe at the user terminal is proposed. The proposed solution was implemented for the Android devices by developing the Android application for the passive QoE monitoring purposes with the ISP's perspectives. The research is conducted to see the impact of the multilayered user end passive probes operating frequency on the user terminal in terms of CPU, RAM and battery utilization. Chapter 3 represents the contribution of the work presented in [26].

Chapter 4 discusses the proposed QoE monitoring solution through the user end the passive probe. In this proposed solution, the activation of the probe is proposed on the quality degradation event. The effectiveness of the proposed approach is investigated as compared to conventional passive monitoring solutions whose probes send monitored data to the collection centers on regular intervals basis. Chapter 4 presents the research contribution of the work in [27].

1.4.2 Part II– QoE Management

QoE management part of the thesis consists of four chapters from Chapter 5 to Chapter 8.

Chapter 5 represents the collaborative QoE management schemes between OTTs and ISPs with a reference framework is devised. Moreover, collaboration driven by the maximization of the revenue is proposed where different factors, such as the user churn (which is modeled as affected by the QoE using the Sigmoid function), pricing and marketing actions are considered. Chapter 5 represents the research contribution of the work in [3].

Chapter 6 discusses the proposed collaborative QoE management solution by considering Customer Lifetime Value (CLV) in the multimedia service delivery chain. A reference architecture to drive the collaborative QoE management of the multimedia services is proposed in combination with CLV based collaborative service management strategy. Chapter 6 discusses the work in [28].

Chapter 7 provides the details of the Software-Defined Networks (SDN) based QoE management platform called *Timber* developed during the doctoral degree. *Timber* is developed on the top of Mininet SDN emulator and Ryu SDN controller, which provides the major functionalities of the traffic engineering abstractions in SDN environment. Moreover, the platform provides an actual complete video streaming application including the implementation of the server side and client side probes for QoE measurements which have functionalities to store the quality measurements into the cloud database accessible to the SDN controller application. In this chapter, we first discuss the general architecture, implementation details and major functionalities of the platform. Chapter 7 provides details of the work published in [29].

Chapter 8 provides the comparative analysis of the collaborative QoE management strategies. Firstly, a reference architecture for possible collaboration and information exchange among OTT and ISP is proposed. Secondly, three different approaches, namely: joint-venture, customer lifetime value-based, and QoE-fairness-based are proposed and compared concerning QoE provided to the users, profit generated for the providers and QoE fairness. Chapter 8 represents the work published in [30].

Chapter 9 provides the thesis conclusion and discussion on the future work to extend the research in the domain of Collaborative QoE management and monitoring for multimedia services.

Part I

QoE Monitoring

Chapter 2

QoE Monitoring Challenges for Inter Service Providers

Abstract

The ever-increasing network traffic and user expectations at reduced cost make the delivery of high Quality of Experience (QoE) for multimedia services more vital than ever in the eyes of Internet Service Providers (ISPs). Real-time quality monitoring, with a focus on the user, has become essential as the first step in cost-effective provisioning of high quality services. With the recent changes in the perception of user privacy, the rising level of application-layer encryption and the introduction and deployment of virtualized networks, QoE monitoring solutions need to be adapted to the fast changing Internet landscape. In this contribution, we provide an overview of state-of-the-art quality monitoring models and probing technologies, and highlight the major challenges ISPs have to face when they want to ensure high service quality for their customers.

2.1 Introduction

In our world of rapidly evolving multimedia services, requirements on networks are growing every day. Services become more and more real-time and require higher bandwidth (e.g., Ultra-HD (UHD) video, telepresence, Virtual Reality (VR), gaming and the Internet of Things (IoT)), putting the network performance of Internet Service Providers (ISPs) under critical highlight. This holds true especially for ISPs delivering third party content from Over-the-Top (OTT) providers (e.g., *YouTube* or *Netflix*). In this ecosystem, all stakeholders aim to provide great experience for their customers — defined as Quality of Experience (QoE).

QoE is the “degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user’s personality and current state.” [31] While this definition is commonly accepted and used in the research domain, traditionally, Quality of Service (QoS) approaches and their associated Key Performance Indicators (KPIs) (e.g., delay or jitter) are the core of monitoring solutions, which are deployed by the industry. Research has shown that raw QoS measurements and the associated KPIs are not necessarily a reliable predictor of QoE [32, 33], which is why modern QoE models focusing on perceptual characteristics of the monitored systems can help at providing a user perspective, based on measurements from the network.

Quality monitoring is one of the first steps in a “quality engineering” process that ISPs implement to ensure the best experienced quality for their customers. The values obtained from measurements are used for getting insight on the current state of the network, but also for managing traffic and providing input for economic decisions. It must therefore be ensured that the acquired data is valid in the sense of providing the ISPs with a representative view of customer-experienced quality. At the same time — in addition to quality — ISPs monitor other parameters in the network, for example related to security and billing. In this chapter however, we want to exclusively focus on the quality monitoring step as an input to other important business processes. Figure 2.1 shows the most important network-related monitoring and management processes for an ISP and highlights the concepts under study in this chapter, namely QoS and QoE monitoring and the respective interfaces to ISP traffic management.

In our perspective — despite the advent of new “killer applications” such as cloud gaming and VR — there are still challenges that need to be addressed by ISPs with regard to the monitoring of “classic” services like video and speech. As researchers and practicioners in the domain of quality monitoring, we still observe a discrepancy between ongoing research and the practical implementation of monitoring solutions, where operators still often take a network-centric perspective, not fully taking into account user factors and QoE.

In this chapter, our first presented challenge is based on asking ourselves what ISPs should optimize a network for: should it still be just best-effort bandwidth, or can a new and more comprehensive focus on tailoring services to user QoE increase the satisfaction level for all users? This calls for new monitoring approaches to ensure that the required user-centric QoE data is accessible for later traffic management steps.

High-performance probing systems may allow ISPs to monitor thousands of connections simultaneously and provide diagnostic analyses as well as QoE estimations. However, peeking into all media streams transmitted over a network, capturing user-specific information

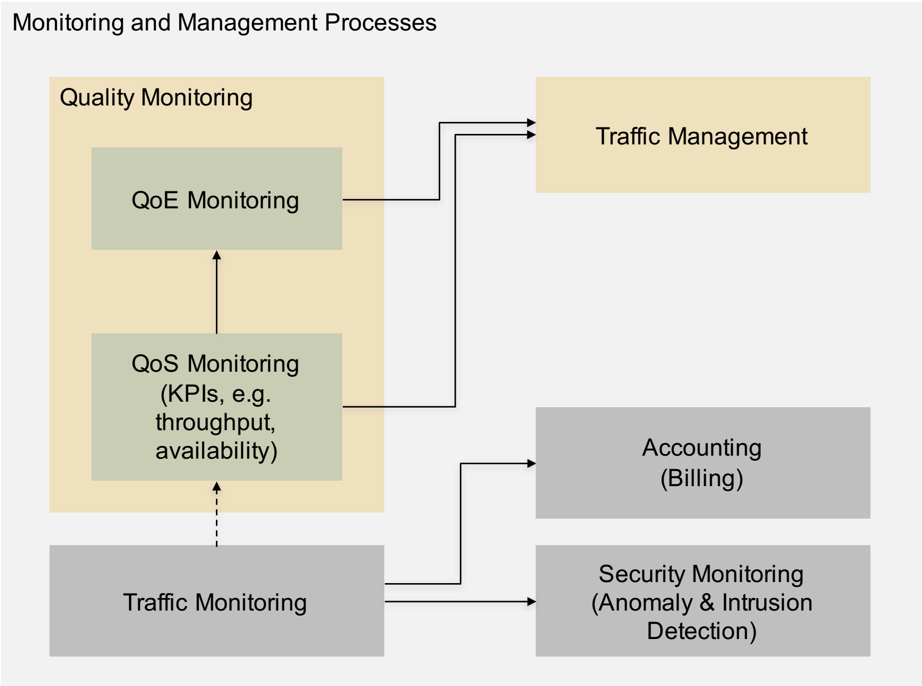

Fig. 2.1 Concepts related to ISP monitoring and management.

touches the critical subject of customer privacy. Here, OTTs have reacted by encrypting their streams; the increased amount of application-level encryption however prevents the ISPs from accessing the same data that their monitoring systems relied on for precise quality measurements. Instead, ISPs will see less information, from which they have to estimate QoE. This paradigm shift presents a big challenge for ISPs and calls for new measurement approaches and redesigns of monitoring systems.

Another set of challenges that we will discuss are rooted in fundamental devleopments in Internet architecture; they relate to how the Internet has evolved over the last years and will continue to change in the near future: the rise of Software-Defined Networks (SDNs) and Network Function Virtualization (NFV) will allow for more diverse and scalable monitoring solutions, but can come with drawbacks due to the increased complexity and abstraction of the system architectures.

Softwarization and virtualization technologies also bring other advantages. In such a scenario, network components are equipped with embedded software procedures; these allow for an easier introduction of new quality models, which is required in monitoring procedures to follow the fast evolution of the deployed applications. Another category of new applications is related to the Internet of Things (IoT). Here, physical (multimedia) sensing devices are being deployed for collecting information about our physical world. These sensors are often virtualized through agents which act as their representatives. All the sensed data is conveyed through them while being processed and cached if needed. These *virtual objects* are promising candidates for the monitoring of the perceived quality.

In this chapter, to provide a suitable context for the challenges later described, we first give an overview of the history and current state of QoS/QoE monitoring solutions. Then, we discuss the above-mentioned challenges from a perspective of researchers and practitioners in the field of QoE monitoring. Our contribution is a discussion of all these aspects with links to current research, as we believe there is a need to raise awareness of those challenges.

2.2 From Quality Models to Probes and Monitoring Systems

In this section, we describe the components of typical monitoring systems. We first focus on monitoring approaches in a historical perspective, then describe types of quality models, which are typically integrated in probes. These are devices that are placed in the network with the purpose of measuring its performance and delivering data to characterize the quality of services running on that network. Finally, we talk about different ways to organize monitoring systems.

2.2.1 From KPIs to Modern Quality Models

From a historical perspective, we can observe different developments in the approaches for network and service quality monitoring. As mentioned before, early generations of performance monitoring were technology-centric rather than user-centric, with the main focus on defining and monitoring KPIs, usually at the lower layers of the Internet protocol stack (e.g., IP layer). In this regard, the metrics defined by the IETF IPPM working group assume a key role even today. They focus on the measurement of quality, performance and reliability of protocols and services that operate on top of the IP layer [34]. Standards by the International Telecommunication Union (ITU-T) also play a role here: ITU-T Rec. Y.1540 defines parameters used for assessing the speed, accuracy, dependability, and availability of IP packet transfer services. ITU-T Rec. Y.1541 specifies performance objectives to be achieved. Finally, to implement monitoring systems, NetFlow and IPFIX are protocols that are often used to capture flows in selected points of the network [35]. While this kind of monitoring is still of great importance and widespread use today, the results obtained from IP-level measurement cannot be directly used for a precise prediction of application-level QoE. Hence, we will focus on the latter aspect in more detail.

In order to find out how certain KPIs relate to user-perceived QoE, starting with the beginning of the last decade, *QoS–QoE relationship models* have become more frequently used. Here, certain QoS parameters are mathematically related to QoE values, usually now taking into account more application-level related parameters. For example, instead of just investigating IP packet statistics to quantify the level of quality for an IPTV stream, RTP traffic can be taken into consideration too. These QoS–QoE models were mainly built for voice and video services, which means that they are lacking universal applicability for other services [33]. We will talk about these models in detail in the next section.

In the recent years, a shift from traditional technology-centric QoS models to more holistic user-centric QoE models could be observed [9], although QoS–QoE mapping models are still used. A common approach for such QoE models is to predict QoE from a multi-dimensional space. These models differ from QoS–QoE relationship models: several other parameters — which are not measurable by traditional KPIs — are taken into account. This category is referred to as "context-aware", since the (user) context has become essential in the modeling of the user-perceived quality. Algorithms based on machine learning and artificial intelligence are gaining interest in this scenario due to their potential for higher adaptability, reliability and robustness over statistical and parametric/probabilistic models [36].

2.2.2 Quality Models

Quality models automatically provide a value of quality based on a certain input, much like if we asked a human observer to rate a given stimulus (e.g., a ten-second video clip). While researchers may use them in the lab to conduct experiments "offline", they are typically implemented in network probing devices in order to allow for automated QoE calculations. We can classify QoE models based on the information they use, according to [1]: as shown in Figure 2.2, we can define *No-Reference* (NR), *Reduced-Reference* (RR), and *Full-Reference* (FR) models, depending on whether they have no, partial, or full access to a source signal. FR models compare the source signal with the received one, for example, an automated telephone call that was recorded both at the source and the receiving end. Since both signals are required for quality calculation, data transfer of the signals is often required. This makes it impractical to use these models at remote ends, especially when large video traffic has to be inspected. In order to reduce the amount of information needed, RR and NR models typically operate on the client side only. RR models receive an auxiliary channel of information, whereas NR models only inspect received signals. They therefore require less data, which typically comes at the cost of lower accuracy [37].

Another form of QoE model classification is visualized in Figure 2.3: *Signal-based models* (or *media-based models*) work on the levels of pixels and samples only; they assume

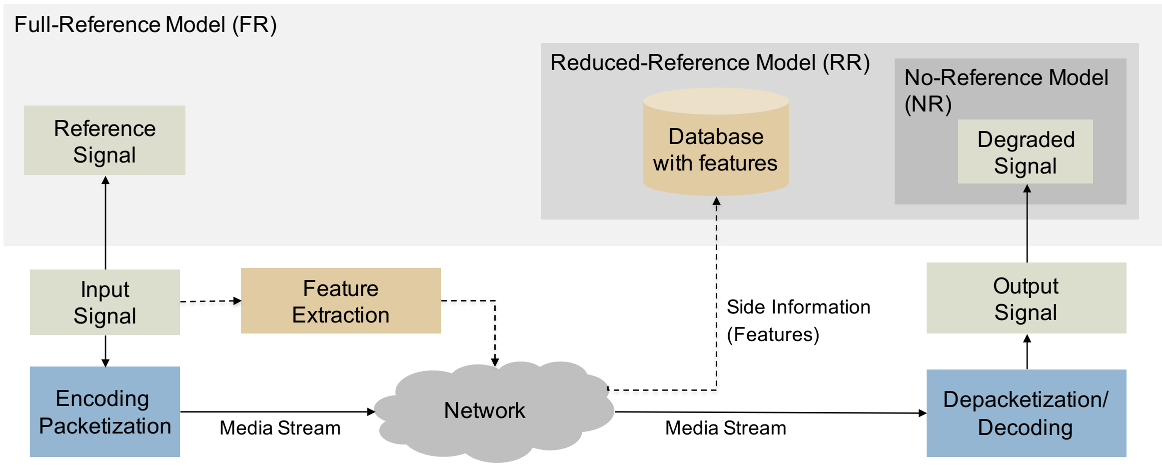

Fig. 2.2 Model classification according to information available from source.

full access to data and decoding capabilities. *Hybrid models* combine signal information with bitstream-level information, such as packet headers. *Parametric models* only operate on transmitted packet-level or bitstream-level information, or — in the case of so-called *planning models* — assume no actual transmission, but can be used for designing networks with a certain service in mind. Naturally, a model with access to more information (e.g., decoded video frames instead of just packet headers) should provide a more accurate estimation of the quality, but in practice — and as we will see later — the amount of information accessible is often influenced by several extrinsic factors beyond control of the ISP.

The International Telecommunication Union (ITU) has standardized a number of quality models for different services. A comprehensive overview of these models was already given by Raake et al. in 2011 [1], however, since then, several new models have been standardized. ITU-T Rec. G.1011 provides a reference of all ITU models for quality estimation and compares their application areas, which helps practitioners in choosing the right model for a given service. Among the most notable models for speech quality evaluation are the Full-Reference ITU-T P.862 (PESQ) and ITU-T P.863 (POLQA) and the parametric ITU-T G.107 (“E-Model”). For video quality estimation, we highlight ITU-T J.144/J.247/J.341 as FR models, and ITU-T P.1201/P.1202 as parametric models for IPTV-type transmissions. An important recent ITU development for media quality assessment is the ITU-T P.1203 family of recommendations, which describes a comprehensive, parametric bitstream-based quality model for HTTP adaptive streaming services. This model consists of an audio and video quality prediction component and a temporal integration component, the latter of which takes into account player-level indicators such as stalling events and audiovisual quality variations.

Of course, apart from standardization efforts, the scientific community and video providers have also released a plethora of quality models; their listing is beyond the scope of this

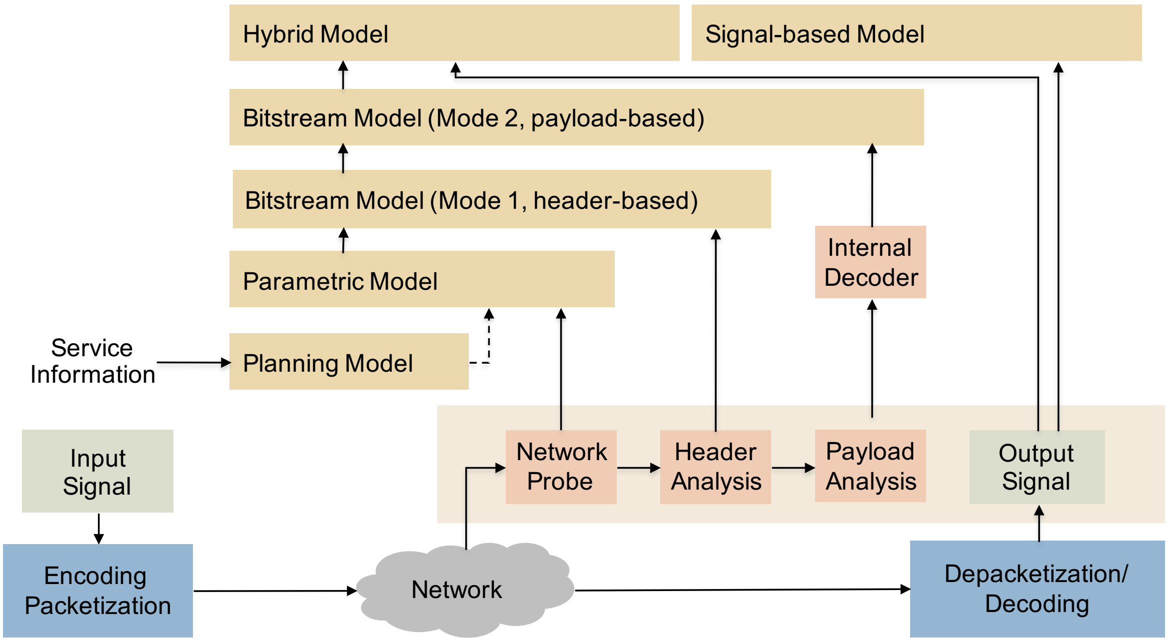

Fig. 2.3 Model classification according to information extracted from the transmission, based on [1].

chapter. Further models for video quality assessment can be found in [32]. Another recent example is Video Multi-Method Assessment Fusion (VMAF) by Netflix [38]. Still, we believe that the use of standardized models for monitoring—rather than internally developed or proprietary tools—allows for a better understanding of the obtained quality estimations; it also allows different technologies and the monitoring results of different ISPs to be compared.

Current models often assume transmission via long-established Internet protocol standards such as UDP and TCP at the transport layer, RTP, and HTTP at the application layer. However, major developments such as HTTP/2 will challenge existing models that, for example, try to predict video buffer states and web page loading times (for websites) to adapt to new transmission techniques. New protocols such as Google's proprietary Quick UDP Internet Connections (QUIC), submitted for standardization to the IETF, will require completely different approaches to modeling web traffic.

2.2.3 Probes

Probes are devices that extract and process information sent over a network (e.g., counting and forwarding packets). They may implement quality models as described above, but they can also be exclusively used for simple network traffic monitoring. We can categorize probes into *active* and *passive* ones. Active probes initiate data transfer. For example, they could start a telephone call to another probe, or request a video from a remote server using a browser running on a PC. Passive probes inspect traffic that passes through them without interfering. This distinction becomes very important in terms of how much and which kind of information can be captured, and how much control an operator has over that information. Active probes typically provide more detailed and reliable information from a user perspective, but they just capture a small sample from an entire network or a network branch. Passive probes typically provide more coarse quality predictions, but offer a better overview of the quality across a whole network or network branch. Passive probes also have the advantage of not generating additional media traffic, whereas active probes may deteriorate already bad network quality by putting additional load on the network paths they operate in. However, comparing the amount of data generated by a single probe in a certain location to the possible traffic requirements of a larger population of real customers, this impact is typically neglected in practice.

By analyzing the major Internet performance monitoring platforms, we concluded that there is not a strong preference between software and hardware probes. SamKnows, BISmark, and RIPE Atlas are examples of platforms that deploy dedicated hardware-based probes, whereas Dasu, Netradar, Portolan and perfSONAR rely on software installations for some hardware systems [34]. (It should be added that these probe systems do not implement higher-level QoE models, but are more typically used for QoS assessment.)

Hardware probes are able to gather round-the-clock measurements, whereas software measurements are more susceptible to resource contention from other applications and are harder to calibrate, but on the other hand have lower distribution costs. Performance is also an issue: surveying the market, one can find a number of offers for probes that monitor network performance on lower levels (e.g., IP). However, for estimating QoE for multimedia systems, higher processing power (e.g., for decoding UHD video streams and performing signal-based analysis) will be necessary. Here, hardware solutions are still more prevalent and practical.

2.2.4 QoS/QoE Monitoring Systems

A complete QoS/QoE monitoring system consists of one or multiple probes, placed in specific locations in a transmission chain, with the purpose of determining the quality of a service at a given point in time, for a certain location. Since probes may just forward captured information, with the quality calculation happening at a central location, we identify two distinct points [9]: 1) the monitoring point, at which information is collected, and 2) the quality estimation point, at which the information is used to calculate a quality indicator.

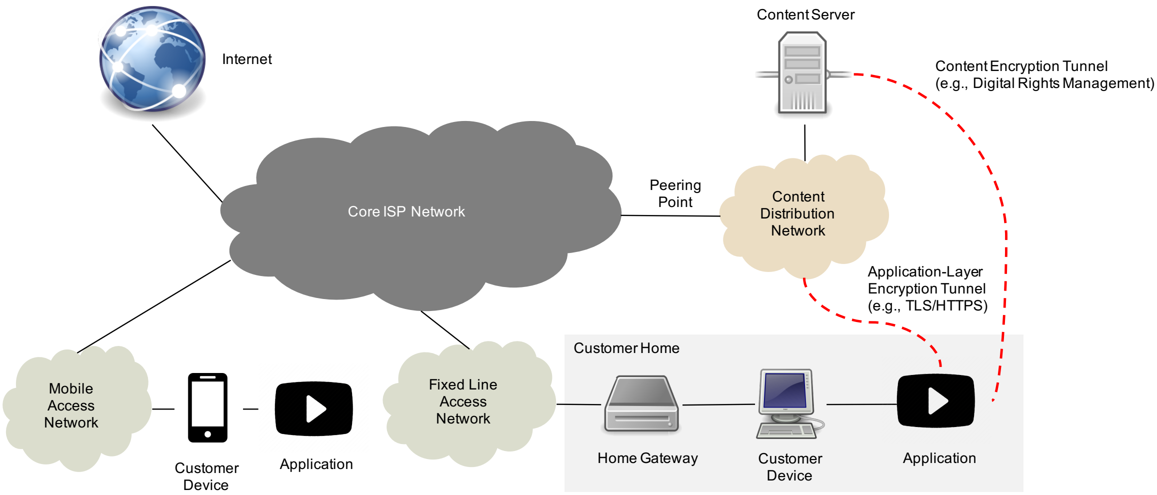

Fig. 2.4 Simplified network topology for OTT media services delivered to fixed/mobile-network consumer. Dashed lines indicate encrypted payload.

In Figure 2.4, we give an overview of a typical situation in which an OTT provides content to a customer. The OTT ingests content to a Content Distribution Network (CDN), which again is peered to an ISP's core network. The ISP provides access to the customer via an access network. Generally, we refer to client-side and network-centric monitoring. In case of client-side monitoring, the probe is placed in the customers' premises, as suggested by the name. However, network-centric monitoring is not so self-explanatory: it refers to placing the probe either in an access network, or in the core ISP network.

Figure 2.4 also shows the tunnels that may exist in case of end-to-end encryption between an OTT and a customer. Content encryption — such as video bitstreams protected by Digital Rights Management (DRM) — has its server-side endpoint at the OTT. CDNs are typically the endpoint for application-layer (e.g., TLS-based HTTPS) encryption. We will later refer to the challenges that ISPs face because of these data tunnels when it comes to QoE monitoring.

What are the advantages and disadvantages of the network- and client-centric approaches? The client-centric method places a probe either at a real customer's location or at a representative access point. For example, a probe may be set up behind the router at users' homes, if they agree to be surveyed. Such a probe is able to capture detailed information, taking into account events happening beyond the access network (e.g., low WiFi performance or faulty routers, wrongly configured DNS settings). However, a large drawback is the much lower number of streams that can be measured when compared to a network-centric probe. For each location, a new device has to be placed (although with virtualization, as we will see later, this issue can be mitigated). Setting up a probe at a real user's location however raises privacy concerns.

The network-centric approach is very efficient: with the respective equipment, multiple network streams can be analyzed in parallel. The challenge here lies in dissecting the data streams and choosing the type of information to monitor. In such a case, it is not always obvious how to take into account degradations that happen beyond the monitoring point. For example, it is more difficult to diagnose end-user equipment problems (such as the ones stated above). Here, the fact that most traffic is connection-oriented helps in troubleshooting, but more detailed traffic inspection (Deep Packet Inspection, DPI) is needed to reveal more information.

Futhermore, network-centric monitoring does not consider the multi-dimensional aspects of QoE. Beyond the network attributes, these aspects include the application, the system, the context and the user behavior itself [39]. This consideration is particularly highlighted in [40], where QoE is represented as a multi-layered vector. This concept is called the *ARCU model*, and each influencing factor is looked at as a set of measurable KPIs. DPI is not applicable in this scenario because it can only classify QoE-relevant traffic characteristics if the traffic is not encrypted. We will later discuss how encryption can be handled for QoE monitoring, however, the factors of context and user behavior will be still missing in that information [23]. The state-of-the-art therefore asks for monitoring QoE by placing the probes not only at the client side, but *in* the user terminal [41–45].

Although the work in [41, 43, 45] includes application-level KPIs in QoE monitoring — highlighting that application-level monitoring is indeed important — these approaches do not consider the context, system, and user factors. The work in [44] considers battery usage as well as application- and network- level KPIs for QoE monitoring. The work in [42] proposes a multi-layered approach for QoE monitoring by considering the application- and network-level KPIs in combination with system-level resources, user location (for the context) and user profile. The authors achieve this through a probe at the user terminal. The major challenge in this regard is the monitoring of the context. Moreover, the selection of optimal operating frequency of a user-end passive probe remains an open issue: the monitoring application at the terminal may require system-level resources which can also affect the performance of the other application [42]. Therefore it is sensible to combine both client-side and network-centric probes for a more accurate and more holistic quality estimation.

As a last thought on monitoring systems, it needs to be stated that QoS and QoE monitoring are merely subsets of general network monitoring practices. This inclusion also particularly applies to the tools and applications of such activities. However, their primary challenges — especially in case of QoE monitoring — fundamentally differ, whether we consider present or future challenges. The following section of this chapter focuses on the discussion of such QoS- and QoE-related challenges.

2.3 Future Challenges and Discussion

Today's Internet is constantly changing its shape: new media services are created, bandwidth availability and requirements are becoming higher and higher by the day, and customers' expectations are changing, adapting to these newly emerging services and novel end-user applications. In this process, ISPs will reach certain limitations and thus face challenges, examples of which we will address in this section.

2.3.1 Optimizing User Satisfaction

Currently the majority of ISPs handle Customer Experience Management (CEM) on the consumer side as a set of KPIs [46], which they consider to be sufficiently reliable representations of actual user experience. For instance, although the perception of service quality indeed correlates to some extent with the number and duration of rebuffering events, these are not the only factors affecting it. From a business perspective, ISPs still primarily focus on these KPIs, since the notion of QoE and its monitoring have just recently appeared as a determining factor among ISP goals. We expect that in the future, monitoring will (and has to) become much more user-centric.

One application of quality monitoring is to measure the compliance of the provided service with a Service Level Agreements (SLAs). These are often set up with customers such as peering partners. In this context, traditional QoS monitoring is prevalent, since in those SLAs only KPIs of the offered services are considered, such as throughput, delay, and packet loss. The ISP's paying customers at the end of the delivery chain often only see a "light" version of such agreements, stating, for example, a maximum available bandwidth. It will be interesting to observe whether an equally important but more user-oriented quality level agreement will become adopted by ISPs. For example, a "Satisfaction Level Agreement" or "Experience Level Agreement (ELA)" [22] would tailor services to the consumers, ensuring a certain level of QoE. A new form of SLAs has already appeared, known as "Next-Generation SLAs" (NG-SLA), metrics of which relate to user satisfaction via business process efficiency [46]. An exciting future trend may also be the personalization of QoE based on individual needs, and optimization considering the interaction between the users of the service.

Current and future QoE optimization goals of ISPs can be approached in several different ways. The most basic approach for optimizing would be simply maximizing mean QoE values while considering financial investments and limitations to maximize the obtainable profit. This might sound general enough to apply to all ISPs, however, optimization does not necessarily happen in this manner. A more specific goal could be to minimize the %POW (percent of Poor Or Worse, see ITU-T Rec. P.910) value in order to maximize the number of service subscribers who do not reject the provisioned quality. By doing so, ISPs reach out for a greater expected number of service subscribers and compensate potential (investment) losses during their quest for improved quality. Similarly, ISPs could aim for maximizing %GOB (percent of Good Or Better, see ITU-T Rec. P.910). However, such a form of optimization — integrated into the business model — rather applies to the future than to the present, as ISPs currently still optimize for KPIs and for best-effort service provisioning rather than a measure of QoE.

Of course the business-driven, QoE-centric optimization of user satisfaction at first sight does not seem like a network monitoring challenge per se. The above considerations reach nearly every element of the value chain. Also, QoE-based customer agreements would necessitate high-level monetary decisions. Yet it needs to be noted that it is indeed a monitoring challenge in the sense of the required data, which must be collected from networks in order to comply with service agreements and to satisfy user requirements. The potential future trend of service personalization — and generally the involvement of individual needs in optimization — however raises the question of privacy, since such solutions rely on information implicitly or explicitly provided by the users. At the time of writing this chapter, there is already an ongoing debate on the trade-offs between enhanced user experience through personalization and the protection of user data. This, on its own, is a considerable challenge for the present and the future.

2.3.2 Encryption and Privacy Aspects

The recent years have seen a rise in public awareness of privacy matters on the Internet. Since the leak of classified documents of the NSA in 2013, alleged “spying” by governments and ISPs has become a topic of public concern.[1] The IETF’s RFC 7258 says, “Pervasive monitoring is a technical attack that should be mitigated in the design of [...] protocols, where possible.” In other words, we will see an increasing protection of all traffic on the Internet. Content providers realized the importance of protecting their user data: in the past years, therefore, OTT providers have switched to application-level encryption in order to offer better privacy to their users, such as with the use of SSL/TLS for HTTP or RTP. For example, *YouTube* force-redirects most of its users to a HTTPS version of their portal. Their mobile transmissions are mostly encrypted, too.

Since the raw payload of encrypted traffic is not visible except for the end points of the connection, such an encryption scheme offers tremendously increased user privacy. To ISPs, the transmission could now only be analyzed at the TCP or UDP layer. This makes it much

[1] For more information, see the documents posted on http://www.theguardian.com/us-news/the-nsa-files.

more challenging to apply any quality model that depends on being able to operate on a level where typically, the send and receive times of application-layer packets has to be known. For example, a model for video streaming portals that estimates buffer levels and predicts video rebuffering events based on the transmitted HTTP chunks — such as proposed in [47] — will cease to operate. Similar situations can be envisioned for VoIP or IPTV transmissions using Secure RTP (SRTP). Models that use parameters from the transmitted audio and video codecs (such as resolution or bitrate) would be impossible to use, too. The ISP would only be able to see TCP/IP traffic patterns from a certain Autonomous System (AS) or a CDN to their customer and vice-versa. ISPs will have to find alternatives for monitoring, while still retaining user privacy.

In the last years, researchers have therefore studied how to estimate the user-perceived QoE from low-level parameters rather than relying on application-layer indicators and models. A prime example of a change necessitating a new view is *YouTube*, which, due to its switch to HTTPS, prevented ISPs from using existing passive monitoring tools that could directly estimate video quality by inspecting HTTP headers. As an example for new approaches based on more low-level data, Orsolic *et al.* [48] have presented a machine-learning–based architecture that estimates *YouTube* QoE from features derived from packet sizes, inter-arrival times, and throughput. They created a testbed to obtain network captures under different traffic conditions as well as ground-truth data against which their models were optimized. The authors succeeded in classifying video QoE, but only resorted to coarse prediction classes (“high”, “medium”, and “low”). A similar approach was shown by Dimopoulos *et al.* [49], who instead used real network data to predict typical QoE indicators for streaming services (e.g., played resolutions, stalling events), based on features such as round-trip times, packet loss and chunk sizes. Here, the authors also used machine learning as a promising technique for large-scale quality monitoring and prediction. Similar approaches to the two mentioned above can be envisioned for all kinds of service classes. However, changes to network architecture (e.g., the introduction of protocols like QUIC or new techniques such as HTTP/2 connection multiplexing) make it necessary to constantly adapt and re-train such models. Furthermore, the diversity of system architectures and subtle details in media player implementations mean that these models may have to be re-engineered for different service providers. Ultimately, predictions from such models may never reach the precision of bitstream-based or signal-based models that have access to the real bitstream.

Countermeasures

While we have seen that an estimation of QoE is possible via passive probing in encrypted networks, the accuracy of such models cannot reach that of models operating on the actual

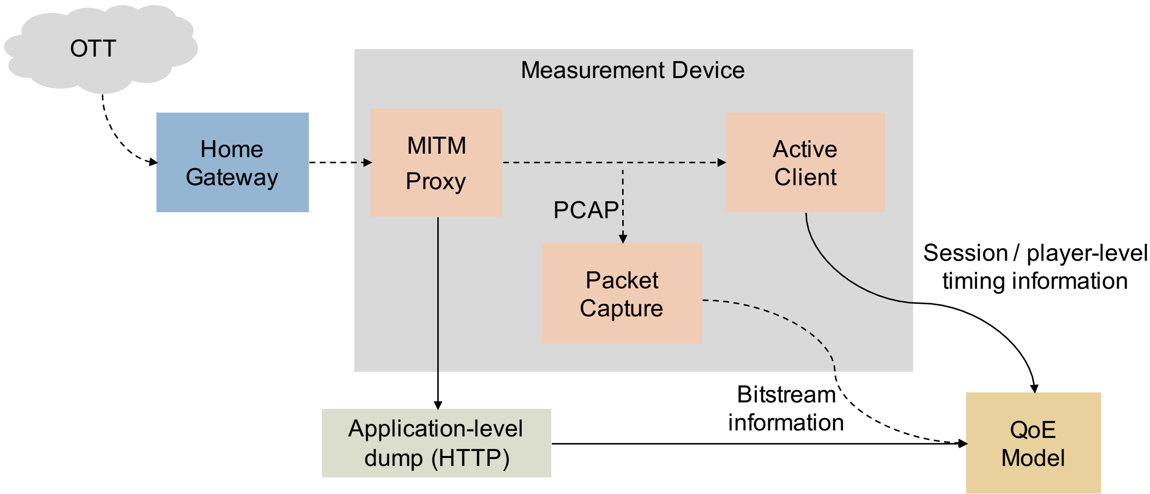

Fig. 2.5 Man-in-the-middle scheme for measuring QoE of HTTPS-encrypted transmissions. Dashed lines indicated encrypted payload.

video payload. Also, the use-case of above-mentioned models is a network-centric monitoring scenario. Thus, encryption only hinders ISPs from passively monitoring (OTT) services that are not under their control — services like VoIP or IPTV may still be monitored from end to end; an ISP could then either monitor a subset or all of their customer's media sessions. In the case where this is not possible, an active probing scenario based on a client-side measurement may still yield KPIs and KQIs to be used for a quality model (e.g., rebuffering events in video streaming). Here, the ISP would simply simulate a customer using a service according to a representative usage pattern. However, in such a case, an underlying encrypted transmission would still prevent the ISP from extracting the cleartext payload of media streams.

One way to overcome this problem is to perform a man-in-the-middle attack (MITM) on the monitoring device itself. Here, any traffic between the OTT and the client player (e.g., a web browser) is intercepted by a proxy installed on the client itself.[2] The proxy thus terminates the HTTPS connection, which gives it access to the cleartext application-level payload. Figure 2.5 shows the general scheme of a MITM measurement. The proxy captures the payload, which can then be analyzed together with player KPIs (such as video stalling events). Given the understandable concerns about user privacy and the resulting full encryption of services, the MITM approach seems promising and easy to implement. However, it can *only* be used on devices owned by the ISP and *not* on a real customer device. This is because a fake certificate authority (CA) needs to be installed on the monitoring device — something that real customers cannot be forced to do, as it would compromise their privacy and would in fact be illegal in most jurisdictions. Furthermore, the impact of

[2] An example of such a tool can be found under https://mitmproxy.org/.

MITM-based approaches on the actual measurements needs to be critically checked. Hence, this solution can only be used in special cases, with active probes, and may not yield a representative picture of QoE.

Future Approaches — QoE as a cooperation between OTTs and ISPs

As long as existing monitoring systems are still functional, there may be little incentive to change a running system or invest into new models and frameworks. Standardization bodies also have to first catch up with the rise of encryption, having to define work items for future studies, which may take years to complete. While commercial tools exist that claim to predict QoE even for encrypted streams, they would have significantly less information about the media available, and thus may be noticeably inferior in their quality predictions when compared to current models.

For a long time, there have been ongoing collaborations among OTTs and ISPs in terms of peering among their Autonomous Systems, either through Public Peering Interconnections (PPI) through Internet eXchange Point (IXPs) or through Private Network Interconnection (PNI). Those collaborations are based on mutual agreements [50], and they often involve payments from one side to another. Some ISPs are also now hosting surrogate servers provided by OTTs in their networks in order to decrease end-to-end latency for the content retrieval as well as to decrease traffic in the core network of the ISP [28, 51]. For example, YouTube and Netflix are providing both peering connections as well as surrogate servers to the collaborating ISPs for their services [52, 53]. As another example, [54] describe a protocol enabling ISP–OTT collaboration. However, these approaches are not primarily QoE-centric and may require incorporation of QoE monitoring and exchange of information among the two entities for a more QoE-oriented service delivery.

One possible solution could be an OTT–ISP collaboration for QoE-aware service delivery. Here, [55] shows an SDN-based ISP–OTT collaboration scheme for video delivery that focuses on video-specific traffic features. Another architecture for ISP–OTT collaboration is proposed in [3]. Here, the OTT monitors quality as delivered to the users on a session basis. They share QoE measurements with an ISP for a joint management of service delivery. The work in [3] shows that this approach could lower user churn.

An OTT–ISP collaboration would therefore consist of OTTs providing the ISPs access to KPIs or KQIs that would otherwise be hidden by encryption. For example, a video provider always knows when their customers experience rebuffering. An interface could be defined in which this kind of information (metadata) is made available to the ISP in an anonymized fashion. Three approaches are possible: 1) The OTT actively sends information to the ISP, 2) the ISP actively requests information from an OTT server, or 3) information is sent along the regular transmissions so that it can be passively monitored. For options 1 and 2, a dedicated server needs to be set up at either location, and a "QoE API" needs to be negotiated between OTT and ISP to exchange information. Another solution for option 1 and 2 could be the extension of the peering API such as PeeringDB API [56] where OTTs can send information as metadata or ISPs can get information related to traffic flows during the peering. Option 3 calls for more fundamental changes in how HTTPS transmissions work. A different solution to this problem [57] would consist in the metadata being carried along the TLS data in an auxiliary channel, but marked as not privacy-sensitive. Also, the data could be sent through ICMP packets in an out-of-band channel at a lower level than TCP.[3]

To give an example of how communication between network operators and OTTs can be achieved, *Server and Network Assisted DASH* (SAND) [58] is an extension of MPEG-DASH (Dynamic Adaptive Streaming over HTTP). DASH has become the de-facto standard for distributing video over the Web, alongside less frequently used solutions like Apple HTTP Live Streaming (HLS). SAND is an example of a system which goes beyond the playback client and the media server communicating over a fully encrypted connection, as typically done today. With SAND, additional network elements are introduced ("DASH-assisting network element"), which enable network operators to receive and signal information on the playback sessions (such as quality metrics), which in turn can be used by the streaming service and clients to optimize quality under the given network operator's constraints.

To summarize, OTT–ISP collaboration is a challenging field of its own, with interesting prospects for users who would gain from improved QoE. However, due to largely political arguments on both sides, realization of the aforementioned concepts is often prevented or slowed in practice.

Net Neutrality Concerns

All the above-mentioned alternatives have their benefits and drawbacks. On the positive side, in all of the above scenarios, the ISPs could utilize data provided by the OTT not only to monitor the QoE, but also to increase the QoE of that OTT service on their networks, for example by selecting more efficient peering points or routing. However, while not a core issue of network monitoring itself, the topic of net neutrality has to be discussed at this point: recent regulations in the United States and the EU have forced ISPs to treat all traffic equally, no matter from/to whom it is sent. In the lawmakers' perspective, Internet is a "good" that all users should have equal access to. Encryption helps to achieve net neutrality, since it (partly)

[3] For further position papers on this topic, the interested reader is pointed to the workshop *Managing Radio Networks in an Encrypted World (MaRNEW)*, hosted by the IETF in September 2015, https://www.iab.org/activities/workshops/marnew/

masks the content of traffic, but ISPs could nonetheless classify it according to its origin and other intrinsic properties.

OTTs may be concerned that after having determined QoS and QoE levels, ISPs throttle their traffic when they have access to QoE-related information. For example, the EU regulation 2015/2120 says that "reasonable traffic management" from ISPs is allowed when it "responds to the objectively different technical quality of service requirements of specific categories of traffic." The directive thus makes it clear that QoS requirements may vary from service to service. However, in addition, it specifies that "any (...) differentiation should, in order to optimise overall quality and user experience, be permitted only on the basis of objectively different technical quality of service requirements (...) of the specific categories of traffic, and not on the basis of commercial considerations." Thus, the EU has adopted a terminology in which QoE is explicitly considered and should be optimized — the requirement for this of course are respective QoE-based monitoring systems.

Business constraints aside, the prospect of optimizing the quality of an OTT service over an ISP's network should incentivize the OTT to supply the necessary information for the ISP to do so. However, customer privacy and net neutrality should in our view never be at stake in such a scenario, and this risk may deter OTTs from engaging in such a collaboration, notwithstanding other financial interests. Ultimately, a partnership between OTTs and ISPs should lead to a more stable or higher QoE for their customers.

How does network neutrality play such a large role in QoE monitoring? A successful collaboration requires ISPs to be able to prove to the OTTs that an efficient provisioning (which may include lowering bandwidth for certain services) does not necessarily lead to reduced QoE. This can only be proven by valid and reliable QoE models or user tests. For example, for a video streaming service, only a well-trained and evaluated QoE model should be the basis for deciding whether a reduction in bandwidth manifests in a significant change in user-perceived QoE. KPIs such as network throughput cannot be employed to drive these decisions alone. It is therefore critical to base any decisions of network management on accurate data. The scenario; however, becomes more complicated when different OTT services or service classes are involved: which ones should be prioritized? Also, different legislations make international collaborations and agreements more complicated. Obviously, the questions raised by these topics cannot be answered yet, and it is to be seen whether collaborative approaches manifest as the go-to solution for ISP–OTT relationships. For further details on net neutrality and the associated perspectives, see [59, 60].

2.3.3 Virtualization

We consider Software Defined Networks (SDN) and Network Function Virtualization (NFV) as key future technologies for network monitoring and management: first of all, they allow for systems to be programmed more dynamically. For example, it is easy to deploy a new model version via software updates and containerized solutions (such as Docker[4]), rather than providing firmware upgrades for hardware probes. This enables more rapid upgrades, higher scalability and lower maintenance costs.

Virtualized monitoring architectures can be built more flexibly and scale more easily. Additionally, virtualization technologies have a significant role in the way physical sensing devices are being deployed for collecting information about our physical world. Since the paradigm of the network monitoring and management is shifting towards virtualization — in combination with drastic growth of high-end IoT-based multimedia applications —, the complexity of monitoring QoE is increasing too. We will address these challenges in this section.

Virtualized Probes at the Network Side

SDN and NFV are becoming more widespread [61], so that core network technologies are shifting towards virtualization. This makes it harder to identify the location in which quality should be monitored and assessed, when those traditional locations do not physically exist anymore. Accordingly, the placement of the virtualized probe remains an important challenge; ISPs should place the Virtual Network Functions (VNFs) by considering effectiveness and cost [62]. For example, in the case of the virtualization of the middle-boxes between two end points, the traffic may follow the indirect paths which may cause potential packet delay hence making the selection of appropriate placement of the probe crucial for delay-sensitive multimedia traffic [62]. While in network data centers, network administrators can perform end-to-end traces of packets, in virtualized data centers, the traffic is invisible to the physical network leading to a significant challenge regarding probe placement and implementation for QoE monitoring. Thus, classic physical probes cannot be used here. Network diagnostics will instead require virtualized probes, which can communicate with both virtual and physical elements of the network for QoE monitoring. Due to a potential high load of the virtual management systems, a trade-off has to be made between accuracy and cost-effective deployment. For example, it needs to be ensured that precise timestamps are collected when needed. Preliminary solutions have been proposed, for instance a CDN

[4]https://www.docker.com/

architecture in which a virtual manager is used to manage the network based on monitoring probes with the help of utility functions [63].

When it comes to monitoring the QoE, the virtualization technologies bring significant advantages. Indeed, new services are being deployed and the existing ones are changing in the way the functionalities are offered to the final users, which has an impact on the perceived quality. This calls for new models that need to be dynamically realised and instantiated into the network. In this scenario, the use of NFVs technologies allows for adopting a plug-and-play approach, so that the collection of new parameters and the new processing operations can be performed without great management burden.

Another important aspect is related to the fact that, an open-loop approach is often used when it comes to the major algorithms and procedures currently embedded in the telecommunication networks. This is the case of some procedures that rely on an off-line estimation of network variables (e.g., quality perceived per user category), rather than on real-time measurements and direct corrective actions. Accordingly, QoE related metrics are usually observed as part of a non-real-time monitoring service as they are just included in offline diagnosis activities for long-term service improvement. This represents a strong obstacle in a scenario characterized by a large variety of supported services and rapid evolutions. In this context the big challenge is the implementation of orchestration functionalities that exploit a closed-loop approach to get the maximum benefits from the available real time feedback information to follow the high dynamicity and unpredictability of the considered scenarios [64]. This objective can be achieved with an extensive use of softwarization and virtualization technologies of the transport services.

Virtualized Probes at the Client Side

ISPs typically monitor the quality at core and access networks. Moreover, some ISPs are also introducing client-centric monitoring procedures through Customer Premises Equipment (CPEs), which are in the form of hardware devices. However, replacing CPEs with devices with more advanced quality monitoring features may be a huge cost factor for ISPs in a very competitive market, as the number of CPEs is in the order of millions or even more, depending on the market share. A more lightweight and client-centered approach to probing would be to offload the measurement into a virtual probe, which runs on traditional, already existing general-purpose hardware. In virtual CPEs [65], the probe itself is nothing more than a piece of virtualized software. It does not depend on a specific hardware, and can be instantiated quickly at different locations.

For this reason, the work in [66] puts emphasis on placing the VNFs at the network edge. However, the approach still remains network-centric rather than user-centric. To shift to a more user-oriented perspective, the works in [41–45] refer to an installation of virtual probes at user-end devices/terminals. Such an approach would not only provide for a cost-effective, robust, and flexible monitoring solution but also can offer additional information related to system, context, and user behavior [42].

However, the placement of virtualized probes for QoE monitoring at user terminals creates new challenges. These include the optimal choice of the monitoring cycle or frequency: a probe installed at the user-end device draws resources available on that device (such as CPU, RAM, and battery) [42]. Hence, continuous monitoring or detailed signal analysis is not possible without consuming users' resources.

A shift towards virtualized probes and virtualized networks will also lead to changes in the way probes are developed and marketed. Vendors of hardware probes rely on business models in which dedicated, highly customized, and costly devices are operated by ISPs, receiving software and hardware support for a longer term. If the ISP wants to expand their traditional monitoring solution, new hardware will have to be bought. Moving to virtualized monitoring architectures, ISPs will have the flexibility to instantiate software-based probes at any time. Vendors of such probing solutions will therefore have to move from one-time payment for physical devices to a subscription- or individual license-based model for the use of virtual probes.

To summarize, this shift to virtualization — no matter if done at the network or client side — may allow ISPs to introduce flexible, cost-effective and hardware-independent monitoring. Still, the implementation and deployment of QoE-centered virtualized probes introduces scalability issues in the required centralized architecture, and challenge security and reliability of the implemented protocols.

Cloud Processing — Issues of Scalability and Security

We believe that placing probes in user end devices will become much more widespread. Those probes will then be monitored centrally. Such an approach may have benefits in terms of end-to-end QoE delivery because it provides ISPs with information related to context-, system- and user-influencing factors (including the geographical location) [39], which relates to the principles of context-aware QoE monitoring. Centralized monitoring systems can utilize cloud technologies to reduce network-wide equipment costs and resolve programmability and flexibility issues. However, scalability will remain an open challenge: network overheads will increase with traffic, since a single controller will be computing all routing paths to generate a network-wide globalized routing map [67]. A real-time computation of QoE-based measurements may also result in additional network-wide delays. With respect to network security, a robust and reliable design of SDN controllers will be needed. The above mentioned issues can be addressed by introducing a semi-distributed approach leaving some freedom to SDN software-based switches to take some local traffic management decisions, while still maintaining the overall view of the network services at the control layer.

2.3.4 Virtual Objects

Virtualization technologies are impacting the way many services are being deployed and provided to the final users. This holds true especially in the field of the Internet of Things (IoT). Here, the physical and the virtual worlds merge to realize services that improve users' quality of life. ISPs that offer IoT services must be aware of the future challenges that come along with such a paradigm shift.

One of the major applications of these technologies is the *virtual object*, which is the digital counterpart of any real (human or lifeless, static or mobile, solid or intangible) entity in the IoT. These can be multimedia objects, which are those capable of acquiring multimedia contents from the physical world. They can be equipped with multimedia devices such as cameras and microphones.

Here are two simple application examples: distributed intelligent cameras could tell key facts about the occupancy of a room and/or the behavior of people in the observed environment; speed gauges, positioning systems and cameras can be used to perform remote monitoring and tutoring of practitioners when practicing with a vehicle. These technologies represent important opportunities for ISPs to provide added value services to their customers [68].

As new applications are being deployed based on these technologies, there is a need for novel approaches and models to evaluate their perceived quality. All the sensors in the described scenarios are virtualized through agents. These act as their representatives so that all the sensed data is conveyed through them while being processed and cached if needed. These virtual objects typically run in a cloud / fog computing infrastructures, as the physical devices often have constrained resources. They are then the best candidate for the monitoring of the contribution of the provided services to the perceived quality. This not only depends on the traditional QoS parameters but also on indicators related to the Quality of Information (to which extent the collected information meets users' needs for a specific time, place and social setting) and Quality of Data (i.e., accuracy of the data).

The needs and opportunities stemming from this technological change come with important challenges. Current, mostly static services like IPTV and speech do not suffer from the same problems; OTT services like Video on Demand only show a fraction of the issues that we may find with entirely virtualized IoT technologies. Here, it becomes much more difficult to create quality models for the services that we deploy with virtualized technologies. This is due to their fluidity, which contrasts with the classic approach in which we perform lab-based subjective tests with users. As soon as new models will be created in the lab and are deployed in the wild, the relevant services have already changed in the way they are delivered to the final users. Accordingly, the models will not be applicable anymore — new metrics need to be considered, or at least need to be included in a different way.

2.4 Conclusion

In this chapter, we gave an introduction to the current state and the future of QoE monitoring, based on the most important challenges we see today. We described the categorization and history of quality models, which are implemented in probes in order to be deployed in QoS/QoE monitoring systems. We also showed examples of recent research into new, user-centric modeling and monitoring approaches.

The major contribution of this chapter is the discussion of challenges that ISPs need to face when implementing QoE monitoring systems for future networks. One of these challenges stems from the fact that ISPs and OTTs will generally move from a network-centric to a more user-centric perspective. Traditional SLAs are not applicable for negotiating and communicating quality to customers, and it can be imagined that QoE will take a more prominent role here — also in the network optimization goals for ISPs. A question that ISPs must ask themselves is whether there should be a guarantee for something as complex as the level of user experience? We see this as inevitable, as with the advent of more complex and resource-intensive services, users will demand more from their ISP than just a steady Internet connection. Most importantly, calling for more detailed user-centric QoE optimization requires appropriate monitoring solutions, which are still challenging to implement.

Another one of the most important issues is introduced by the increased amount of application-level encryption. How can ISPs monitor application-level data when most OTT traffic is going to be end-to-end encrypted? Here, current works explore machine learning approaches which allow ISPs to predict QoE from lower-level network measurements. Another proposed solution would be a collaboration between OTTs and ISPs. To be more precise, a side-channel could allow OTTs can report certain KPIs and KQIs to ISPs. However, net neutrality concerns play a large role in such a scenario: although net neutrality regulations mostly affect how traffic shaping is being implemented by an ISP, the use of reliable QoE models to quantify the user experience is a precondition for that. We conclude that ISP–OTT collaboration is a highly political topic, but there are technological opportunities that offer ways for both parties to create better services for their users.

Finally, the rise of virtualized networks brings up a number of questions. These include challenges in how new probes have to be designed: they will be developed and marketed in a different manner. Furthermore, in order to allow an ISP to leverage the capabilities of those new networks, the placement of the (virtualized) probes at either client or network side becomes a critical issue. They will no longer be specific hardware devices at a particular physical location, but instead can be more flexibly scaled, which might draw more resources on nodes. We also explained how implementing probes in customer equipment — as a way to perform client-centric monitoring — is not a silver bullet. As a look into the future of the Internet, IoT technologies will fundamentally change the deployment and monitoring of services. Here, the major challenge lies in developing appropriate quality models.

The future of QoE monitoring is certainly not exclusively determined by the above challenges — and these challenges are also relevant to other areas such as QoE-based traffic management, security, and billing. These issues will be shaped by a broader adoption of new services and technologies such as Virtual Reality and cloud gaming, which are much more demanding in terms of network requirements. However, as we have seen in the chapter, even when monitoring the "classic" technologies such as video and speech, ISPs encounter challenges that they need to and will need to face. These developments raise important questions that the research community has to provide answers to, in collaboration with the industry and standardization bodies.

Chapter 3

A Multilayer Passive QoE Monitoring Solution at User Terminal

Abstract

This chapter focuses on passive Quality of Experience (QoE) monitoring at user end devices as a necessary activity of the ISP (Internet Service Provider) for an effective quality-based service delivery. The contribution of the work is threefold. Firstly, we highlight the opportunities and challenges for the QoE monitoring of the Over-The-Top (OTT) applications while investigating the available interfaces for monitoring the deployed applications at the end-device. Secondly, we propose a multilayer passive QoE monitor for OTT applications at the user terminal with ISPs prospect. Five layers are considered: user profile, context, resource, application and network layers. Thirdly, we consider YouTube as a case study for OTT video streaming applications in our experiments for analyzing the impact of the monitoring cycle on the user end device resources, such as the battery, RAM and CPU utilization at end user device.

3.1 Introduction

The provision of quality to the end users is the prime objective of the key players involved in multimedia service delivery over the Internet. Recent forecast in Internet traffic has predicted that 75% of the world's mobile data traffic will be multimedia by 2020 [69]. The drastic increase in the multimedia traffic, specially Over-The-Top services (OTTs) like YouTube, Skype, Netflix, etc., has raised the issue of the quality delivered to the end users. Moreover, the service quality is most of the time degraded by the poor network conditions and occurrence of congestion in the access network of the Internet Service Providers (ISPs) [9]. Nevertheless, the provision of the quality to the end user is an important concern for the ISPs because the quality is considered to be an important factor when it comes to the user churn [70]. Moreover, the assurance of the quality may not only decrease the user churn but may also increase the overall revenue at the enterprise level [3].

With the recent researches in the field of multimedia service delivery, the term Quality of Experience (QoE) is gaining popularity among researchers for being more user-centric rather than Quality of Service (QoS) [4]. The conducted research in the field has shown that the QoE is a multidimensional metric which depends on different factors related to human, networks, application and context [71]. This is leading to a big change in the monitoring systems, suggesting monitoring traditional Key Performance Indicators (KPIs) in combination with the Key Quality Indicators (KQIs) to make the service QoE-aware [11].

QoE monitoring for OTT applications is challenging for the ISPs because ISPs are most of the time blind to OTT application quality and traffic because of encryption and user privacy [72]. Moreover, the ISPs are not directly involved in the loop of revenue generation between users and the OTTs, which lowers ISPs incentives to invest in the new equipment for QoE-aware monitoring of the OTT multimedia applications. Nevertheless, the QoE-centric service may result in higher revenues by lowering user churn [3]. QoE monitoring of OTT multimedia applications can be considered as the first step for QoE-aware Internet service delivery by ISPs. Additionally, cloud-based services in combination with user end probes may provide an opportunity for ISPs for a cost-effective and easy deployment of the QoE monitoring for OTT applications. Nonetheless, the selection of the optimal frequency for the passive QoE monitoring is critical because of the limited end-user device capabilities and cloud-based service pricing.

The major contribution of this chapter is threefold. Firstly, with reference to the current scenario in the multimedia industry, we highlight the challenges and opportunities for monitoring the OTT multimedia applications quality with the perspectives of the ISPs. Secondly, we propose a multilayered cloud-based passive QoE monitoring solution which only relies on passive probes at the user terminal. In our approach, we do not only consider the application and network level impact on QoE but also important factors like the system, context and user profile. Thirdly, we implement the multilayered passive QoE monitor for the android devices and then we investigate the impact of the QoE monitoring cycle on the user end device by taking into account factors such as battery level, CPU, RAM. With the experiments, we highlight that higher sampling frequency of the passive QoE monitor at the user end device may utilize more device resources, which may affect the performance of the other on-device applications.

The chapter is structured as follows: Section 3.2 discusses the state-of-the-art work found in the literature, while Section 3.3 highlights the challenges and opportunities for ISPs to monitor QoE of OTT application. In Section 3.4, we propose cloud-based QoE monitoring with the user end probe, while Section 3.5 provides the implementation details of our Multilayered passive user end probe called *Qualia*. In Section 3.6 we discuss the experimental setup test conditions and results. Finally, we provide conclusion and discussion in Section 3.7.

3.2 Related Works

The work in [40] highlights the multidimensional aspect of QoE by presenting a layered model called ARCU, which considers the key influencing factors on the QoE, such as application, resource, context and the user. The represents QoE as multidimensional vector where each influencing factor is represented in different space in terms of measurable KPIs which can be included in a service to make it QoE-aware. Although the literature reports several state-of-the-art works for OTT application monitoring, most of the approaches are limited since the important factors of the system, context, and user profile are ignored. Moreover, the studies do not discuss the impact of the QoE monitoring cycle on the user end device. The rest of this section discusses past works for QoE monitoring of multimedia applications over the internet which are compared in Table 3.1 with respect to the influencing factors: network, application, system, context and user.

In [41], Aggarwal et al. investigate the utilization of passive probe-based measurements for QoE monitoring of VoIP and video streaming applications. The approach uses machine learning based algorithms for QoE monitoring with the assumption that there is a relationship between network traffic generated by the application and application KPIs. However, the work is only based on the estimation of the application quality and may not work for encrypted OTT application sessions. Moreover, the impact of the QoE Monitoring on the user end device is neglected. Similarly, the study in [43] presents passive probe-based real-time YouTube monitoring. The probe for QoE Monitoring is placed between Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN) in the core network, not in the end user device. The monitoring is based on the assumption that stalling events happen whenever the throughput become lesser than the video rate because the buffer starts to become empty. The method detects the YouTube traffic by inspection of the HTTP header; however, since the YouTube header has been changed to HTTPS and sessions are encrypted, this approach may not be able to evaluate the video streaming quality of YouTube.

Table 3.1 Comparison of the state-of-the-art work w.r.t influence factors

Influence factor	**YoMo [73]**	**Prometheus [41]**	**QoE Doctor [44]**	**Youqmon [43]**
User profile	✗	✗	✗	✗
Context	✗	✗	✗	✗
Resource/System	✗	✗	✓ Only Battery Consumption	✗
Application	✓	✓ Only estimates App. KPIs	✓	✓ Only estimates App. KPIs
Network	✓	✓	✓	✓

In [73], the authors propose a tool called *YoMoApp* for QoE monitoring of YouTube video streaming at the user end device. The proposed QoE monitor works as a normal application in the user mobile device. The work considers application level KPIs such as rebuffering events, video quality, and player state only while simulating some network test conditions. However, the tool does not account the impact of QoE monitoring on the user end device. Also, the important factors of system and context are ignored. The work in [44] presents a tool for QoE measurement named *QoE Doctor* which can analyze the QoE of an application based on a three-layer approach. The proposed QoE Monitoring technique is based on the followings: 1) User-perceived latency; 2) Mobile data consumption, because cost and user billing is associated with it; 3) Energy consumption by the network. Though the tool considers important features, it loses practicality and scalability because QoE monitoring is based on the manual monitoring of the UI changes of mobile device in the developer environment which cannot be applicable in practice, i.e., user mobile devices are not connected to the developers environment in daily life and ISPs cannot monitor UI changes manually.

3.3 ISPs' Challenges and Opportunities in QoE Monitoring

QoE monitoring of multimedia applications is challenging for the ISPs because of the complexity involved. We highlight some the key challenges and opportunities in the following.

3.3.1 Challenges

One of the major challenges for QoE monitoring is the encryption of the traffic generated by most of the OTT applications. For example, YouTube has turned from HTTP to HTTPs, where the videos are transmitted in encrypted sessions between the client's device and the streaming servers [17]. Hence, Deep Packet Inspection (DPI) may not be applicable to monitor the multimedia traffic generated by OTT applications. Even if DPI works for some multimedia applications which are not encrypted, the factor of context would still need to be monitored somehow with additional tools for providing better and accurate information about the overall QoE [23]. Moreover, the KPIs of the application cannot be known by DPI, which makes it a limited approach because with DPI only traffic can be classified. User privacy laws are another big challenge, which also restricts ISPs from DPI without user consent. In addition to the aforementioned challenges, the cost of the network equipment, maintenance and operations remain an open challenge for ISPs at the enterprise level. Nevertheless, to deal with the high dynamics of the multimedia industry, a low-cost, flexible, reliable and context-aware QoE monitoring solution is required by ISPs.

3.3.2 Opportunities

Still considering the analyzed challenges, there are opportunities for the ISPs to incorporate QoE monitoring tools in the form of cloud services and user terminal probes. The installation of the QoE monitoring probe at the user end device can leverage on hardware-free QoE monitoring solutions, which can decrease the network equipment/operations cost. Subsequently, the issue of the user privacy can be removed by taking user consent for the installation of the QoE monitoring probe in the end device. The power of software-based user end probe can also help to deal with the changing dynamics of OTTs' multimedia applications.

Referring to the mobile device specifically, most of the OTT applications are providing interfaces in the form of developer APIs, for other applications to interact with their applications. For instance, YouTube is providing a developers' API[1] by which other applications can measure application specific KPIs. However, the provision of APIs for application specific measurements always depends on the OTT service. Nevertheless, the interfaces to get application quality on the device already exist in the mobile devices Operating Systems (OS) like Android and iOS in the form of interaction among different applications, provided that the OTT service provider allows it. For example, in Android an application can interact with others by calling application development specific methods named *intent service* and *content providers* by which an application can access data from the on-device database of the other

[1]https://developers.google.com/youtube/

applications. Notwithstanding with the importance of QoE monitoring, very few application providers are providing APIs for the other applications to interact with. Furthermore, the user end probes also provide an opportunity to measure system-level KPIs: device computational power, RAM, battery level in additions to context and user related information such as the location of the user, mostly used application and currently run the application. Moreover, cloud computing services provide an opportunity for ISPs to have scalable and easy to deploy QoE monitoring solutions where the collected data from the user end probes can be stored in cloud-based databases for future analysis. However, the trade-off between the utilization of on-device and cloud-based resources is important because the latter may eventually result in higher costs for the ISP while the former may limit end device capabilities by lowering the available resources for other high-end applications.

3.4 Multilayer Cloud based QoE Monitor with User End Probe

The QoE-aware service requires different roles of the involved entities in the multimedia service delivery. In case of our proposed solution, there are the following three entities with different roles: 1) *User*– Installation of the user end probe provided by ISP; 2) *OTT*– provision of the API to ISPs for collection of application related quality indicators and; 3) *ISP*– Utilization of the information collected through the probe for QoE monitoring and management while respecting the Network Neutrality. We propose that the user end probe is based on the following five layers, differentiated on the basis of the extracted information and ordered from the top layer to the bottom layer:

1. *User Profile Layer:* The top most layer in our QoE monitoring system is responsible for the collection of the more user related information such as application usage history of the user and user status whether an active user of the service. This layer provides the individual/collective user(s) trend for OTT applications. By considering the activity frequency of the user, ISPs can predict the user churn easily because mostly inactive users are likely to become churner [74].

2. *Context Layer:* This layer is responsible for the acquisition of context related data such as location of the user (home, station, social events), currently running applications in user terminal, and user mobility. The context related information may help ISPs for better load balancing and traffic management, especially in highly populated areas; e.g., the prediction of the future traffic demands from the context-related data may improve resource allocation [23]

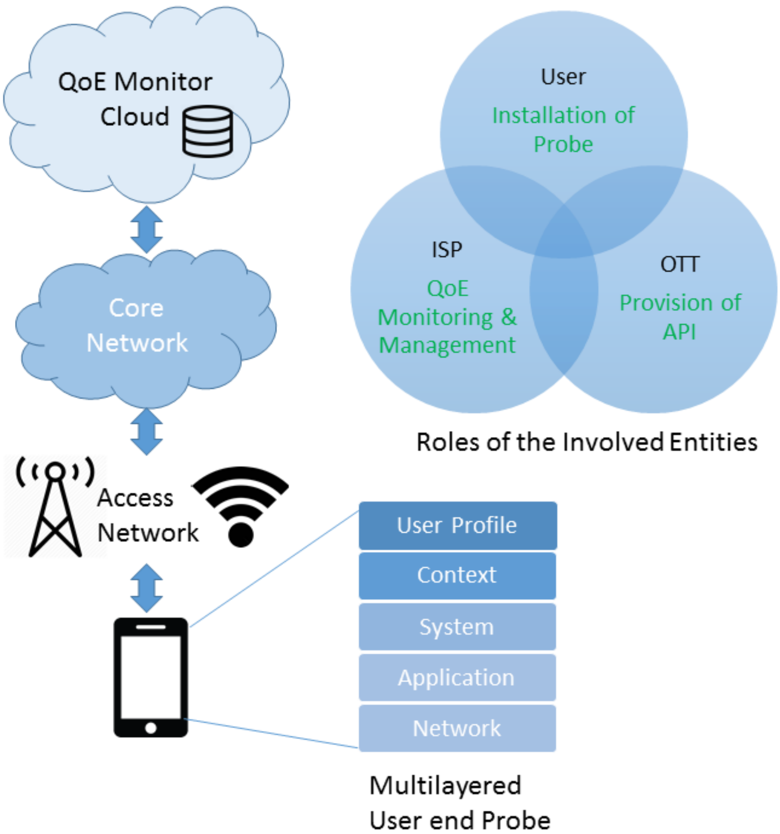

Fig. 3.1 Overview of the proposed multilayered QoE monitoring solution with the roles of the involved entities.

3. *Resource/System Layer:* This layer is being considered because the probe will be installed at the user end device and it will be important not to utilize a large amount of the resources on the user device because it can affect the quality of other applications. We consider the factors of battery power, CPU and RAM consumption by the QoE monitoring probe as well as other applications.

4. *Application Layer:* This layer includes the functionality to collect the OTT application KPIs through probes. However, the measurement of application level KPIs highly depends on whether the application provider is providing developers APIs as mentioned in Section 3.3.2.

5. *Network Layer:* This is the bottom most layer with the role of collecting network performance measurements at the user terminals, which may lead to identifying the bottleneck in the access network to drive necessary QoS management actions, included management of Physical Resource Block in wireless access.

The software based passive monitoring probe can be installed at the user terminal just like a normal application for the collection of the data required by the each layer for better analysis of QoE. The installed probe sends the data to the database in the cloud server. The users can be assigned with a unique ID which can be used to sort data from all layers. For the measurement of the quality, we propose the placement of application specific objective QoE models in the cloud to utilize cloud computing. Fig. 3.1 shows the overview of the proposed methodology and the layers of the user end probe in combination with the roles of the involved entities.

We consider that QoE is a continuous function of time which requires continuous monitoring. Therefore, we assume the monitored QoE is a continuous function of time $QoE_c(t)$ which in our case also depends on the information collected by each layer. We denote with $\alpha_{i,j}(t)$ the j-th type of information collected in i-th layer at time t through the user end probe, which is a continuous function of time. J_i is the total number of types of information in the i-th layer, and I is the total number of monitoring layers. However, we consider the fact that continuous time monitoring cannot be possible in the digital world so discrete time sampling of these functions can be represented mathematically as follows:

$$\alpha_{i,j}^{discr}(t) = \alpha_{i,j}(t) \sum_{k=0}^{\infty} \delta(t - kT) \tag{3.1}$$

where $\alpha_{i,j}^{discr}(t)$ is discrete time equivalent of $\alpha_{i,j}(t)$, T is the sampling interval, and $f = 1/T$ is the sampling frequency of the QoE monitor for gathering the data through the user end probe. The sampling interval/frequency is critical in terms of on-device and cloud services resource management, i.e., if the sampling interval is too small it may not only affect the performance of the user device by utilizing a high amount of resources, but also generate a large amount of data that may result in data handling and cloud service cost. Clearly, from the observed $\alpha_{i,j}^{discr}(t)$ values, we obtain a discrete version of the $QoE_c(t)$ function.

Section 3.6 highlights the impact of the sampling interval/frequency of the proposed QoE monitor in detail.

3.5 Implementation details

This section discusses the detailed implementation of the multilayered QoE monitor with the passive probe at the user terminal proposed in Section 3.4. The current implementation of our multilayered QoE monitor is named *Qualia*.

The proposed probe is designed specifically for Android devices due to their large market share in the electronic devices market [75]. We consider YouTube as a case study for OTT

Table 3.2 List of Quality Indicators considered in implementation

Type of Layer	**Quality Indicators**
User profile	List of mostly used applications per day
Context	Device Location, List of active applications
Resource/System	CPU, Battery and RAM Consumption
Application	Initial Loading Time, Stalling Frequency, Stalling Duration
Network	Throughput, End to end delay (RTT)

video streaming applications in our experiments, because YouTube is the mostly used video streaming application over the Internet and a developer's API to get application quality indicators is available [76]. The proposed probe installs and runs as a normal application on the device. Qualia includes a YouTube player in the foreground where the user can play and interact with the video. The YouTube player is implemented using the YouTube developer API for Android. The YouTube API for Android not only provides the functionality for embedding the YouTube player in an Android application, but also information about the occurrence of the video streaming events such as video loading, video buffering, player start time. We also considered the Android application life cycle in the implementation of Qualia, where we deliberately made the process of the computation and the collection of the quality indicators in each layer to run in the background as a *background service* in the Android device so that it may not be interrupted by the Android application life cycle system. The Android application life cycle management system stops/restricts the threads or applications which are not in the foreground. We selected Amazon Web Services[2] (AWS) as a cloud platform for the collection of the data. Since Qualia is a passive cloud-based QoE monitor, the cloud service is solely used for data collection purpose. The data generated by Qualia, which mainly consists of the userID and several quality indicators, is sent to the cloud database at selected regular sampling intervals. The rest of the section provides the detail of the quality indicators considered.

Starting from the bottom most layer of the multilayered approach, the network throughput and delays are considered for the network layer. The network throughput is calculated by computing the data transfer from the start till the end of each sampling interval. The network delay is computed by the calculation of the Round Trip Time (RTT) from the link of the played video being playing in the YouTube player. At the application layer, we consider the *initial loading time*, the *stalling frequency* and the *stalling duration* as application KPIs on the basis of the model proposed in [77]. The initial loading time is computed by observing

[2]https://aws.amazon.com/

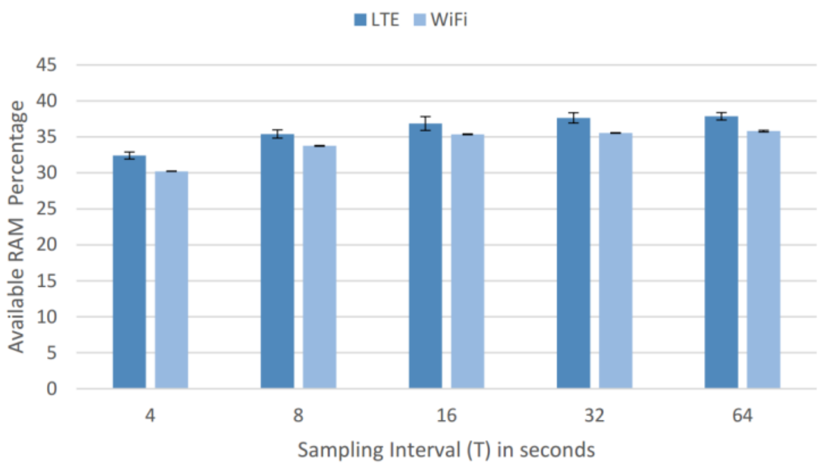

Fig. 3.2 Utilization of the on-device RAM per sampling interval

the time difference between the time when the video started to load in the player and the time when the video is loaded. The stalling frequency is calculated by counting the occurrence of rebuffering event event while stalling duration is computed by the time difference between the time when rebuffering event occurs and the time when the buffer is refilled and video plays again after buffer event. On the system layer, we considered the overall RAM, CPU and battery availability in the user end device at the time of measurements, because the user end probe may utilize more resources at the user end device and may effect the quality and performance of the other application. Moreover, the main objective of the experiments is the investigation of the impact of the passive probe and the sampling interval on the available on-device resources. The available CPU, RAM, and battery are obtained by built-in functions provided by the Android development kit in the Android Studio. In our implementation, we consider user's location and recently active applications for the context layer. For the user profile layer, we consider whether the user is an active or inactive user of the service by checking the frequency of the online status of the user. For this, we assigned each user a unique ID by utilizing the cloud service provided by AWS named as *AWS Cognito*. Table 3.2 summarized the quality indicators considered in our implementation.

3.6 Experimental Setup and Results

The objective of the experiments is the investigation of the impact of sampling interval/frequency of the proposed multilayer QoE monitoring approach on the end-user resources such as a battery, RAM, and CPU consumption. The experiments are conducted using our implementation of the multilayer passive QoE monitor *Qualia*.

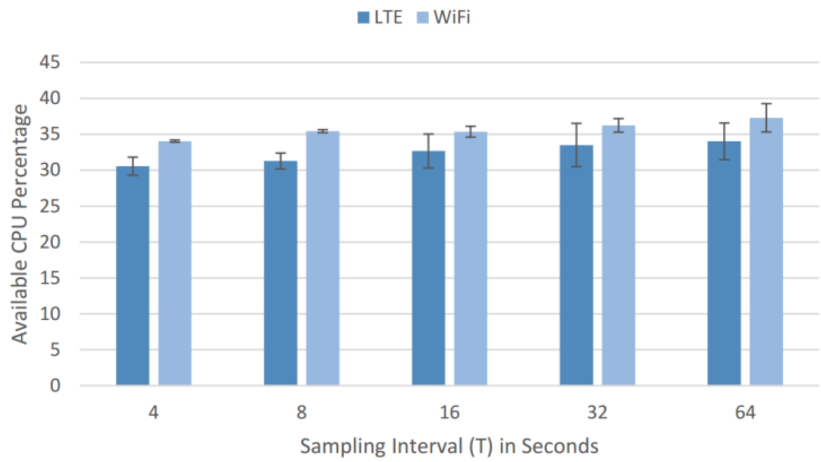

Fig. 3.3 On-device CPU utilization per sampling interval

The experiment details are as follow. The video sequence of Big Buck Bunny[3] of 1080 HD resolution with total 2 minutes and 2 seconds playtime is selected for the experiments because it is the most used video sequence found in literature. The sampling interval T has been set to 2^n seconds, where $n = 2,3,4,5,6$. We have not selected $n = 1$ because 2 seconds is a very short interval of time for the computations of all quality indicators in the mobile device and may result in redundant data. Hence, T ranges from 4 seconds to 64 seconds. The video sequence is played 10 times (10 runs) for each sampling interval. All the already installed applications on the device were stopped during the experiments to see purely the impact of passive QoE monitoring on the user terminal and avoid their possible influence on the device resources that could have lead to inaccurate results. The experiments are conducted using Samsung Note 4 at the lab. The specifications of the device used in the experiments are shown in Table 3.3. The device was fully charged before the start of the experiments. We considered all the layers in our probe. During the experiments, the user device was not in mobility; however, in order to consider the context layer, the location of the device was actively collected. Though there was no other application running on the device except our QoE monitor Qualia, we also collected that information during the experiment in order to see the impact on device resources with all layers. We considered two different scenarios; in one case the mobile device is connected to the LTE network while in the second scenario the mobile phone is connected to the WiFi network. The nearest cloud server of AWS based in Frankfurt was selected for the data collection.

All the results are the average of the 10 runs for each sampling interval. The results in Fig. 3.2 represent the average percentage of the free RAM available per sampling interval T.

[3]https://www.youtube.com/watch?v=Q1NAW7Xf1oE

Table 3.3 Specifications of the device used in experiments

RAM	3 GB
CPU	1.9GHz octa-core Samsung Exynos 5433/ 2.7GHz quad-core Snapdragon 805
Battery	3,220 mAh

The graph shows the increasing trend with the increase in the sampling interval, highlighting that with the increase in duration of the sampling interval (lower sampling frequency), more on-device memory is available to the other applications. The change in available RAM by changing the sample interval from 4 seconds to 16 seconds appears to be 4.45% (133.56MB) and 5.14% (154.28MB) in case of LTE and WiFi connection respectively. Furthermore, a small change of 1.00% (30.15MB) and 0.44% (13.23MB) is observed in free available RAM between $T = 16$ to $T = 64$ seconds for LTE and WiFi respectively. The reason of the large change in RAM utilization from $T = 4$ to $T = 16$ is that 4 seconds is quite high frequency and in case of $T = 4$ the probe has to compute all the quality indicators (all the variables involved) after 4 seconds. As the values of the variables are stored in the RAM during computation, with such a high frequency the average RAM utilization was high. Moreover, the probe appears to utilize more RAM while connected to WiFi connection with respect to LTE network. The graphs in Fig. 3.3 show the amount of CPU utilization during the different sampling intervals. Similar to the RAM, the CPU utilization of the user end probe increases with a decrease in the sampling interval, i.e., the higher the sampling frequency the higher the utilization of on-device resources. The observed change in the CPU utilization from $T = 4$ to $T = 16$ seconds is noted to be 2.11% to 1.3% for LTE and WiFi connections respectively, while a change of 1.34% and 1.93% is observed from $T = 16$ to $T = 64$ seconds for LTE and WiFi connection respectively. An important fact is that the probe utilizes more CPU while connected to the LTE network as compared to WiFi connection. Fig. 3.4 highlights that utilization of the battery for the $T = 4, 8$ seconds appears to be the same; also for $T = 16, 32, 64$ seconds, the battery drainage remains constant within the connection type. Nonetheless, the battery drainage for the lower sampling interval (higher sampling frequency) is higher than for higher sampling interval. Table 3.4 summarizes the changes in on-device resources observed w.r.t the change in sampling interval.

3.7 Conclusion

In this chapter, we proposed a multilayered QoE monitor at the user terminal for measuring the quality of multimedia applications. The proposed approach considers the perspective

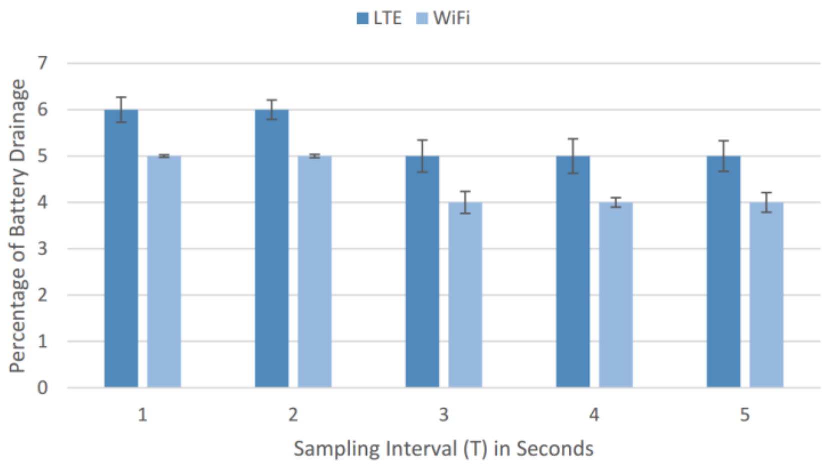

Fig. 3.4 Battery drainage per sampling interval

Table 3.4 Percentage changes in the available on-device resources w.r.t change in sampling interval T in seconds

Connection Type	RAM T=4-16	RAM T=16-64	CPU T=4-16	CPU T=16-64	Battery T=4-16	Battery T=16-64
LTE	4.452%	1.005%	2.114%	1.346%	1%	0%
WiFi	5.142%	0.441259	1.307%	1.935%	1%	0%

of ISPs for QoE monitoring. The proposed approach provides a cost-effective and scalable solution for measuring multimedia application quality which is not only flexible but also easier to deploy with the current network infrastructure without bringing any change into the existing network architecture. In this chapter, we also provide the novel concept of the discrete time QoE monitoring at user terminal where we also present the investigation of the impact of the multilayer user end probe in terms of the utilization of the device resources by changing the sampling interval of the proposed approach. Some interesting observations were possible with respect to the change in the connection type on the proposed approach.

Chapter 4

A QoE Monitoring Approach based on Quality Degradation

Abstract

This work is based on the investigation of the passive Quality of Experience (QoE) monitoring of Over-The-Top (OTT) applications at the user terminal with the perspective of the Internet Service Provider (ISP). The contribution is multi-fold. Firstly, we investigate the advantages and drawbacks of QoE monitoring at the user terminal by highlighting the roles of the different entities involved in multimedia service delivery, i.e., users, OTTs and ISPs. Secondly, a survey is conducted related to user's privacy and preferences with reference to the installation of monitoring probes at the terminal. The analysis of survey results highlights the applicability of the approach in terms of users' willingness to have probes at their terminals in exchange for higher quality, discounts or extra data usage. Thirdly, we propose a multi-layered passive monitoring approach at the user terminal which considers the activation of the probes only in case quality degradation is revealed (in an event-based fashion), in contrast to conventional passive monitoring solutions whose probes send monitored data to the collection centers on regular intervals basis (constant frequency). As a case study, YouTube is considered as video streaming service in our experiments to investigate the effectiveness of the proposed approach.

4.1 Introduction

In the past decade, multimedia traffic has emerged to be the prevalent Internet traffic and by 2020 it is forecast to be at 80% of all IP traffic [78]. This is mainly due to the rising popularity of handheld devices (e.g., smartphones and tablets) and Over The Top (OTT) services such as YouTube, Netflix, Facebook, and Skype, which provide their services over the networks of Internet Service Providers (ISPs). Thus, ISPs have to upgrade their networks to deal with the issue of quality delivery to the end users because most of the time the quality is degraded due to bottlenecks in the ISPs' networks caused by the drastic increase of multimedia traffic. Moreover, within the multimedia service delivery chain, ISPs are the most affected entities by the user churn because most of the users who perceive inadequate quality switch their Internet connection to another ISP [3].

The end-to-end encryption adopted by OTTs does not allow the ISPs to monitor the quality of the OTT multimedia services to apply the basic principles of network traffic management. Even if Deep Packet Inspection (DPI) may be applicable for some multimedia services which are not encrypted, user-related factors such as application, human and context would still need to be monitored somehow with additional tools for providing better and accurate information about the overall quality perceived by the users, i.e., the Quality of Experience (QoE) [4]. Hence, the ISPs must incorporate new monitoring solutions able to measure network parameters as well as to inspect user-related factors which would allow for obtaining measures closer to the QoE.

In this chapter, we consider the opportunity for the ISPs to incorporate passive QoE monitoring tools in the form of cloud services and user terminal probes, with the objective to provide a QoE-aware service delivery of OTT applications. The contribution of this chapter is multifold. Firstly, we investigate the advantages and drawbacks of QoE monitoring at the user terminal by highlighting the roles of the different entities involved in multimedia service delivery, i.e., users, OTT and ISP, as well as by comparing this approach to other QoE monitoring approaches found in the literature. Secondly, we conduct a survey asking people specific questions regarding the monitoring at their terminals to understand privacy issues and user's preferences for the betterment of delivered quality. The analysis of the conducted survey highlights the applicability of the approach in terms of users' willingness to have probes at their terminals for QoE-aware service delivery. Thirdly, we propose a passive monitoring approach at the user terminal which considers the activation of the probes only in case quality degradation is revealed. This approach is opposite of conventional passive monitoring solutions, whose probes use to send monitored data to the data collection centers on the regular intervals (constant frequency) basis. Also, this could lead to extensive utilization of computational and power resources of the user terminal. Instead, with the proposed monitoring approach, the QoE is predicted at the user terminal as a function of monitored network, application and context parameters. Depending on the service considered, specific QoE models proposed in the literature may be used for QoE prediction. Hence, the probe at the user terminal is activated only if the degradation in the quality is found beyond a certain threshold, i.e., in an event-based fashion. Thereafter, the monitored data is sent to the cloud database of the ISP, where a management system will take the optimal decisions for resolving the issue related to quality degradation over the ISP's network. We consider YouTube as a case study for video streaming application in our current implementation. Finally, we provide a comparison of the proposed approach with state of the art monitoring approaches, such as constant frequency-based monitoring approaches, on the basis of utilization of device resources such as battery lifetime, RAM, and CPU as well as the amount of data generated by the monitoring probe.

The chapter is structured as follows. Section 4.2 discusses the state-of-the-art work found in the literature. Section 4.3 describes the motivations behind the monitoring at the user terminal as well as how this will be implemented. Section 4.4 presents the results of a survey conducted to investigate users' privacy issues regarding the monitoring at their terminals. In Section 4.5, the proposed monitoring approach is presented whereas Section 4.6 provides the implementation details of the monitoring solution. Section 4.7 provides experimental results and finally Section 4.8 concludes the chapter.

4.2 Related works

The QoE is influenced by various different factors, such as system, human and context factors [4]. In [40], the multidimensional aspects of QoE is highlighted by presenting a layered model called ARCU, which considers key influencing factors on the QoE, i.e., Application, Resource, Context, and User. QoE is represented as a multidimensional vector where each influencing factor is represented in a different space. A layered model considering the impact of different influence factors on the QoE of multimedia applications is also proposed in [71]. Though the literature is full of the state-of-the-art work for monitoring of OTT applications, however, most of the techniques found in the literature do not consider the important factors such as system, context, and user.

In [41], the utilization of passive probe-based measurements is investigated for QoE monitoring of VoIP and video streaming applications. The approach uses machine learning based algorithms for QoE monitoring with the assumption that there is a relationship between network traffic generated by the application and application KPIs. However, the work is only based on the estimation of the application quality and may not work for encrypted OTT application sessions. Moreover, the impact of the QoE Monitoring on the user end device is neglected. Similarly, the study in [43] presents passive probe-based real-time YouTube monitoring. The probe for QoE Monitoring is placed between Serving GPRS Support Node

(SGSN) and Gateway GPRS Support Node (GGSN) in the core network, not in the end user device. The monitoring is based on the assumption that stalling events happen whenever the throughput become lesser than the video rate because the buffer starts to become empty. The method detects the YouTube traffic by inspection of the HTTP header; however, since the YouTube header has been changed to HTTPS and sessions are encrypted, this approach may not be able to evaluate the video streaming quality of YouTube.

In [73], the authors propose a tool called *YoMoApp* for QoE monitoring of YouTube video streaming at the user end device. The proposed QoE monitor works as a normal application in the user mobile device. The work considers application level KPIs such as rebuffering events, video quality, and player state only while simulating some network test conditions. However, the tool does not account the impact of QoE monitoring on the user end device. Also, the important factors of system and context are ignored. [44] presents a tool for QoE measurement named *QoE Doctor*, which analyzes the QoE of an application based on a three-layer approach. The proposed QoE Monitoring technique is based on the followings: i) user-perceived latency; ii) mobile data consumption, which is associated with cost and user billing; iii) energy consumption by the network. Though the tool considers important features, it loses practicality and scalability because QoE monitoring is based on the manual monitoring of the UI changes of mobile device in the developer environment which cannot be applicable in practice, i.e., user mobile devices are not connected to the developers environment in daily life and ISPs cannot monitor UI changes manually.

4.3 Monitoring at User-Terminal: The Why, What and How?

The QoE reflects the quality as perceived by the user and depends on system factors (QoS) as well as on user-related factors such as human, application, and context [4]. Although the ISP is able to monitor network QoS parameters, the end-to-end encryption adopted by OTTs does not allow the ISP to monitor the quality of the OTT multimedia services to apply the basic principles of network traffic management [17]. However, even if Deep Packet Inspection (DPI) may be applicable for some multimedia services which are not encrypted, user-related factors would still need to be monitored somehow with additional tools for providing better and accurate information about the user perceived QoE [23]. Hence, the ISP must incorporate new monitoring solutions able to measure network parameters as well as to inspect user-related influence factors that would allow for obtaining measures closer to the QoE. However, the measurement of these factors is not straightforward because of their strong dependence on the user. Thus, there is the need to monitor such factors as close as possible to the user side, i.e., at the user device.

The advantage to run monitoring probes at the user device is essential as it is possible to access to parameters regarding the installed applications as well as to information regarding user behavior. However, the QoE monitoring at user terminal requires the efforts by all the involved entities, i.e., user, OTT, and ISP, which have different roles: 1) *User*–should install the monitoring application provided by the ISP in the user device and allow the application to monitor different influencing factors in accordance with one's privacy concerns; 2) *OTT*–should provide specific APIs through which application level KPIs can be monitored by the ISP and: 3) *ISP*–should monitor the quality delivered to the user through its application considering the user's privacy concerns.

For example, it is possible to measure network parameters (e.g., throughput, packet loss rate, delay) through the network interface as well as application parameters (e.g., buffer state, bitrate) by querying applications through specific APIs (if provided by OTTs, e.g., YouTube provides the developer API[1] through which application level information can be retrieved by other applications). Also, terminal sensors' can be used to monitor context factors, such as the GPS for acquiring the user position, whether the user is moving or at home, etc. Further context information such as the price paid for the service and the terminal specifications may be acquired from the user profile and the terminal information respectively. The main issue regarding the monitoring at the user terminal is the user's privacy, since the user may not be willing to provide some personal information. This aspect will be discussed in Section 4.4. The monitoring probe, in this case, is a software running on the user terminal and it is specifically programmed as an application compatible with the operating system running on the device. In this work, we implemented an application for Android devices. This application should be allowed to access device and user information in order to measure and collect parameters for QoE evaluation.

4.4 Privacy Issues and User's Preferences: A Survey

A survey is conducted to investigate users' preferences and privacy concern to allow ISPs to install monitoring probes in the user terminals. The survey comprises of 224 responses, where the gender ratio (male/female) is 70.6% / 29.4%. The 95% of participants are $21-35$ years old, mostly from Italy, Pakistan, and Greece. The 93.7% of participants use multimedia services over the Internet on the daily basis.

[1] https://developers.google.com/youtube/

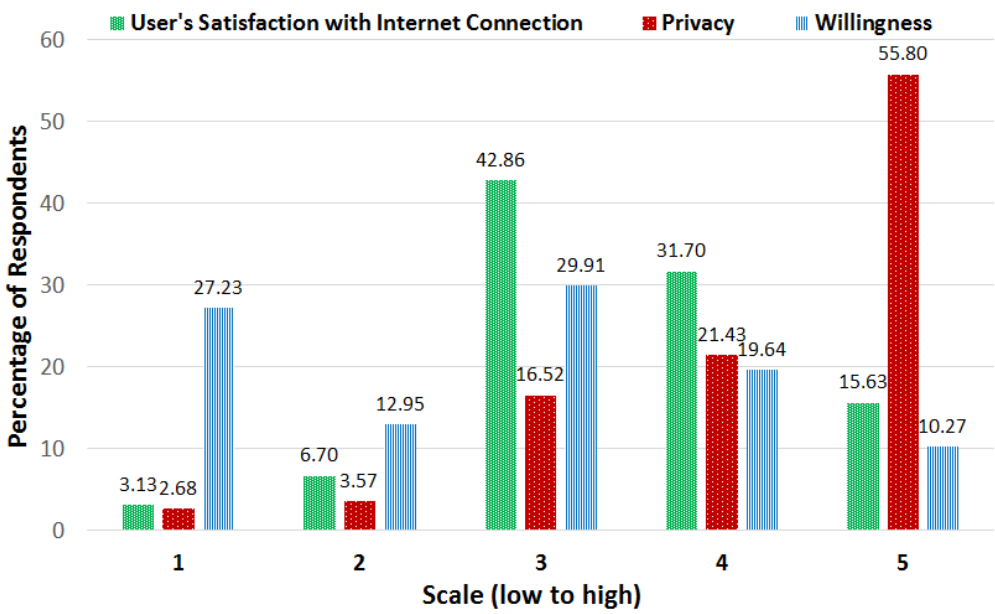

Fig. 4.1 The results of the user's preferences and privacy concerns.

In the survey, we asked people to answer, with a scale from 1 (lowest value) to 5 (highest value), three specific questions: 1) *How much are you satisfied with your Internet Connection?*; 2) *How much privacy matters to you?*; 3) *In exchange of some benefits, will you allow ISP to install an application in your terminal to monitor the quality of your multimedia applications?*. The results of the survey are illustrated in Fig. 4.1, where *User's satisfaction*, *Privacy*, and *Willingness* identify respectively questions 1, 2, and 3. Users are mostly satisfied with their Internet connection: 47.33% of the population rated their Internet connection between 4 and 5 whereas 9.83% rated between 1 and 2. The others, the 42.86%, gave a 3. With regard to user privacy, it is evident that for most of the respondents the privacy is very important. Indeed, 77.23% of respondents rated between 4 and 5, whereas the 6.25% rated between 1 and 2 so do not care about their privacy. The 16.52% showed medium level of privacy concern giving a 3. Although for most of the users their privacy matters a lot, the 29.91% is highly willing to install a monitoring probe in their terminal in exchange for some benefits (rates between 4 and 5). However, the 40.18% is not willing to install the probe (rates between 1 and 2) whereas the 29.91% showed medium level of willingness (3).

Finally, we asked the question *In which condition will you allow the ISP to monitor the quality of the applications you are using in your device?* to investigate which benefits may be preferred by users. It resulted that the 28% do not want to install any probe under any condition, whereas 72% agreed on the probe installation in exchange for better quality (25%), discounts on the services (10%), extra bandwidth/data usage (11%) and a combination of these 3 benefits (26%). Hence, on the basis of these answers it appears to be feasible for the ISP to install the probe in the users' terminals by satisfying their benefits preferences for a majority of users.

4.5 Quality Degradation based monitoring at user-terminal

We propose a multi-layered QoE monitoring solution with the probe installed at the user-terminal. In contrast with the traditional monitoring techniques, the passive monitoring probe is activated when quality degradations occur. The information collected by the user-end probe is classified into five different layers, as follows:

1. *User Profile Layer*–This layer is responsible for the collection of user-specific information such as user preferences of the service, mostly used applications in the terminal, the frequency of service usage (to classify whether the active user or not) and user's subscription of the service.

2. *Context Layer*–The context layer monitors the context related information such as the location of the user, user mobility and currently running the application in the terminal. This information may help the ISP to perform better load balancing and resource allocation in highly populated areas [23].

3. *System Layer*–This layer collects the information related to user-end device resources such as CPU, RAM, battery level and device capabilities because if the installed probe at user terminal utilizes more resources on the user device, it may affect the performance of the other applications.

4. *Application Layer*–The application level KPIs are collected by this layer on the basis of the objective parametric QoE models and currently running applications in the terminal.

5. *Network Layer*–This layer collects the network level KPIs such as throughput, jitter, packet loss rate, delay.

An overview of the proposed quality degradation based QoE monitoring solution is shown in Fig. 4.2.

The software based passive monitoring probe is installed in the user terminal to collect the aforementioned information, which is then sent to ISP's cloud server where it is stored with reference to user ID. We propose the placement of application specific objective parametric QoE models also on the user terminal for two reasons: 1) real-time measurements of the

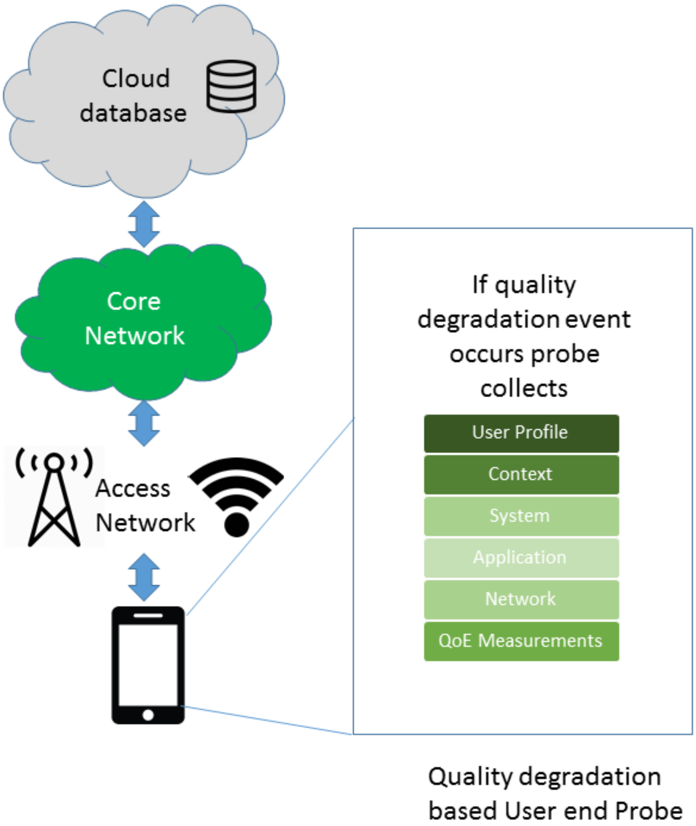

Fig. 4.2 Overview of the proposed quality degradation based QoE monitoring solution.

quality delivered to the user and; 2) activation of the passive monitoring probe to send the data to the ISP's cloud server for the QoE management. The monitoring probe runs as background service in the user device and computes the QoE of the currently running application in the user terminal so as to monitor the application specific quality degradation events. For example, in the case of video streaming OTT services such as YouTube, the number of stalling events have a higher impact on the user perceived quality than video quality [79]. So in that case, the probe will be listening to the stalling events and whenever the stalling event happens, the probe is activated to collect the aforementioned information in the different layers in order to send it to the cloud database server owned by the ISP. In the case of Voice over IP (VoIP) application and web browsing, call drop and a certain threshold of response time can be respectively considered as events for activation of the probe. According to the model proposed in [80] for web browsing, if the response time increases above 5 seconds, the Mean Opinion Score (MOS) decreases below 3. Therefore, for the web-browsing, we propose that when the response time is increased above 5 seconds, the probe is activated.

4.6 Implementation Details

This section provides the implementation details of the multi-layered quality degradation based monitoring solution proposed in Section 4.5. The current implementation of the proposed monitoring solution is named *Zeus*.

We developed the proposed monitoring solution particularly for the Android system because of the huge market share of the Android devices [75]. In our implementation, we consider YouTube as a case study for the QoE monitoring of video streaming application because YouTube provides the developer's API through which application level KPIs such as initial loading time, stalling frequency, stalling duration can be obtained [76]. The current implementation of the probe installs as a normal Android application in the user device. *Zeus* has a standard YouTube player in the foreground of the application through which videos can be played and the user can interact with the YouTube player. The probe collects the different information in the background as *background service* to consider the Android application life cycle: if the application is not running in the foreground, Android application life cycle stops/restricts all the processes/threads associated with that application. *Zeus* listens to stalling events in the background. Whenever the stalling event happens the data is collected, QoE measurements are computed and all the information is transferred to the cloud database. For the cloud database, the Amazon Web Services[2] (AWS) is selected to collect the data acquired through the user-end probe. The data generated by *Zeus* is solely composed of the quality indicators in each layer with respect to the userID which is sent to the cloud server if the quality degradation event occurs (in this case when stalling event occurs). The rest of the section discusses the details of the quality indicators collected by the *Zeus*.

Starting from the Top layer, the data collected from the user profile layer is composed of the list of most used application. Moreover, we also consider whether the user is an active or inactive user of the service by checking the frequency of the online status of the user. For this purpose, we assigned each user a unique ID through *AWS Cognito*. In the context layer, user's location and currently running applications are collected through built-in functions provided by the Android development kit. For the system layer, the probe collects the information related to CPU and RAM utilization as well as the battery level of the user device because the main objective is the investigation of the resource utilization by the passive probe at the user terminal. The system level information is also collected by the built-in functions provided by the Android development kit. On the application level, *stalling duration* and *number of stalling events* are collected based on the model proposed in [81]. The number of stalling events are computed throughout the video playback by counting the occurrence of

[2]https://aws.amazon.com/

Table 4.1 List of Quality Indicators considered in implementation

Type of Layer	Quality Indicators
User profile	List of mostly used applications per day
Context	Device Location, List of active applications
Resource/System	Utilization of CPU, RAM and Battery
Application	Number of Stalling events, Stalling Duration
Network	Throughput, End to end delay (RTT)

the re-buffering events while the stalling duration is calculated as the time difference between the occurrence of rebuffering event and the video playback event after the rebuffering. The QoE measurements are also computed using the model in [81] when the stalling event occurs during the video playback. On the network layer, throughput and end-to-end delay are considered as quality indicators. The throughput is computed by the calculation of the data transfer rate at the time of stalling event while the end-to-end delay is calculated from Round Trip Time (RTT) from the link of the video being played in the player. The Table 4.1 provides the summary of the quality indicator considered in our implementation.

4.7 Experiments and Results

The objective of the experiments is the evaluation of the proposed monitoring solution (event-based) in comparison with the conventional constant frequency-based monitoring techniques, in terms of utilization of the resources on the user terminal as well as cloud database utilization. The experiments are conducted with our monitoring implementation, i.e., *Zeus* and regard the monitoring of the quality of the YouTube video streaming application.

For the experiments, the video sequence of Big Buck Bunny[3] with 1080 HD resolution is selected. The video sequence is 2 minute and 2 second long. The aforementioned video sequence is selected because it is the most used video sequence in the literature. *Zeus* monitors the stalling events throughout the video playback: if the stalling event happens, the probe measures the QoE as well as collects the data in the different layers and sends it to the cloud server. To make a comparison of *Zeus* with the constant frequency-based probe approach, we consider the activation of the probes at the user terminal at regular intervals (either every 4 or 8 seconds).

The video sequence is played 10 times (10 runs) in each configuration (Zeus, 4 and 8 seconds sampling intervals). All the installed applications are stopped before the experiments

[3] https://www.youtube.com/watch?v=Q1NAW7Xf1oE

Table 4.2 Specifications of the device used in experiments

RAM	3 GB
CPU	1.9GHz octa-core Samsung Exynos 5433/ 2.7GHz quad-core Snapdragon 805
Battery	3,220 mAh

in order to solely evaluate the impact of the proposed monitoring solution on the user-end device resources such as RAM, CPU and battery level. The experiments are conducted on the *Samsung Note 4* in the lab environment. Table 4.2 provides the details of the device used in the experiments. The device is fully charged before the start of experiments. Though the device was not in mobility during the experiments but information related to the location was acquired from the probe to consider the context layer. For the experimental purposes, all the application in the terminal were stopped in order to investigate the impact of only user end probe on the on-device resources.. For the collection of the data, the AWS cloud server based in the Frankfurt is selected for being the nearest cloud server provided by AWS. Furthermore, the experiments are conducted with the two different type of wireless networks: WiFi and LTE in order to consider two different scenarios.

All the results are the average of 10 runs in each configuration and scenario. Fig. 4.3 highlights the battery usage in each configuration per scenario. The graphs show that battery utilization in case of quality degradation based activation of the probe is less than as compared to the constant frequency-based approach. However, the difference is negligible because the battery drainage is mostly related to device screen luminosity level which is kept constant for all configuration and scenarios. Moreover, the LTE connection appears to consume more battery than the Wi-Fi connection. Fig. 4.4 provides the comparison among each configuration per scenario with regard to the available RAM. The difference of available RAM percentage between the *Zeus* and constant frequency based approach 4 s and 8 s are found to be respectively 6.26% (187.8 MB) and 2.72% (81.6 MB) for Wi-Fi connection. While for LTE connection a greater difference is observed, i.e., 9.89% (296.7 MB) and 6.89% (206.7 MB), respectively. Fig. 4.5 represents the comparison of *Zeus* with the constant frequency-based approaches in terms of utilization of CPU on-device. The difference in CPU utilization between *Zeus* and constant frequency based approach with 4 seconds frequency is 6.30% and 6.77% for Wi-Fi and LTE connections respectively. Whereas the difference in utilization between *Zeus* and 8 s constant frequency based approach are 4.92% and 6.05% for Wi-Fi and LTE connections respectively. Table 4.3 represents the summary of the results.

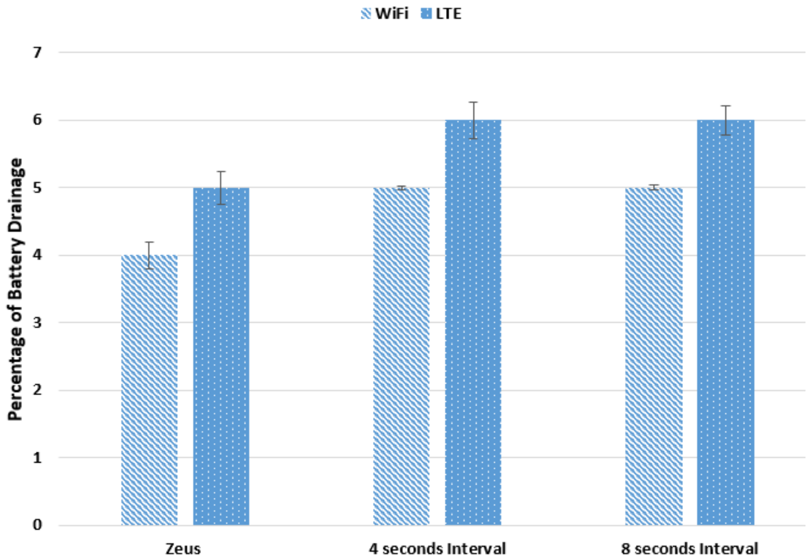

Fig. 4.3 On-device battery drainage.

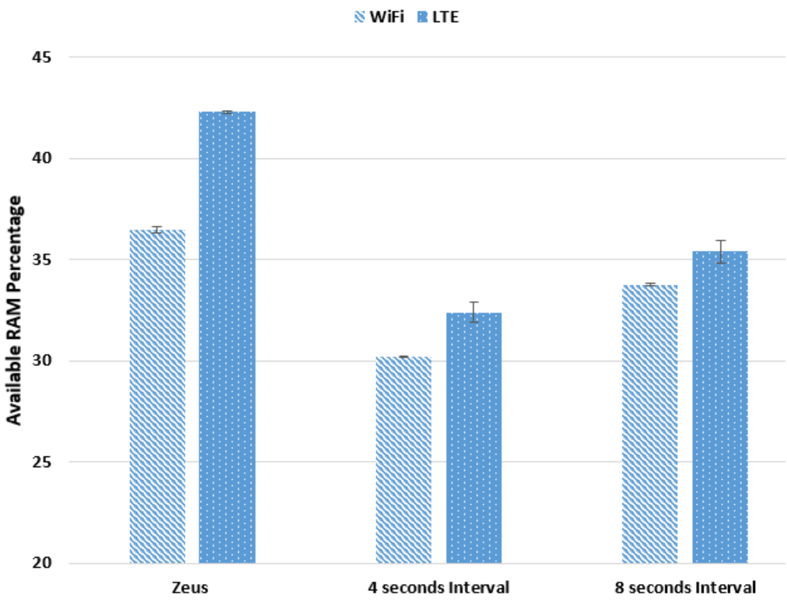

Fig. 4.4 On-device RAM utilization.

4.8 Conclusion

In this chapter, we proposed a multilayer passive monitoring solution where the activation of the probe is based on the quality degradation event. From the results of a survey regarding user's privacy and preferences, we motivate the user's willingness for the installation of the probe in the terminal. The proposed approach is not only cost-effect to be incorporated into the current infrastructure of ISPs' network but also scalable. The experimental results

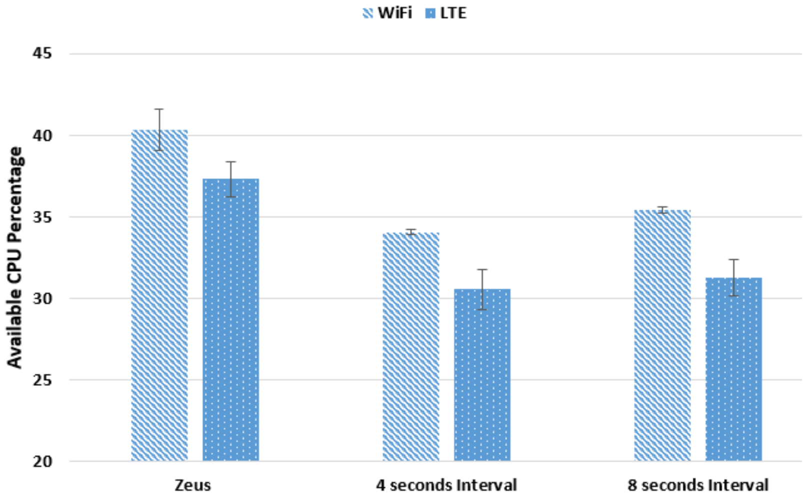

Fig. 4.5 On-device CPU utilization

Table 4.3 Percentage changes in the available on-device resources in each configuration

Connection Type	RAM Zeus-4s	RAM Zeus-8s	CPU Zeus-4s	CPU Zeus-8s	Battery Zeus-4s	Battery Zeus-8s
WiFi	6.26%	2.72%	6.30%	4.92%	1%	1%
LTE	9.89%	6.89%	6.77%	6.05%	1%	1%

highlight the novelty of the approach in terms of an on-device resource such as CPU, RAM and battery level. The proposed approach outperforms the traditional constant frequency based monitoring approach while utilizing less on-device resources. The future research will be focused on the investigation of cloud service utilization w.r.t to the activation of probes as well as consideration of different layers in the monitoring solution.

Part II

QoE Management

Chapter 5

A Joint-Venture Approach for Collaborative QoE Management

Abstract

The provisioning of the quality to end users is a major objective for the successful deployment of multimedia services over the Internet. It is more and more evident from past research and service deployments that such an objective often requires a collaboration among the different parties that are involved in the delivery of the service. This chapter specifically focuses on the cooperation between the Over-The-Top (OTTs) and the Internet Service Providers (ISPs) and propose a novel service delivery approach that is purely driven by the Quality of Experience (QoE) provided to the final common users. Initially, we identify the need of the collaboration among the OTTs and the ISPs where we not only highlight some of the enterprise level motivations (revenue generation) but also the technical aspects which require collaboration. Later, we provide a reference architecture with the required modules and vertical interfaces for the interaction among the OTTs and the ISPs. Then, we provide a collaboration model where we focus on the modeling of the revenue, whose maximization drives the collaboration. The revenue is considered to be dependent on the user churn, which in turn is affected by the QoE and is modeled using the Sigmoid function. We illustrate simulation results based on our proposed collaboration approach which highlight how the proposed strategy increases the revenue generation and QoE for the OTTs and the ISPs hence providing a ground for ISP to join the loop of revenue generation between OTTs and users.

5.1 Introduction

Internet traffic has evolved over the past decade from web traffic to multimedia traffic due to the widespread use of smartphones as multimedia content generators and significant advancement in multimedia services over the Internet. Recent studies on trends in Internet traffic have predicted that more than 75% of the world's mobile data traffic will be multimedia by 2020 [78]. Such a drastic increase in the use of multimedia services requires more resources at the Internet Service Providers (ISPs) end to assure the required level of quality to the users, although ISPs are not in the loop of revenue generation between the providers of the multimedia services (i.e., the OTTs - Over-The-Top providers) and the users. Indeed, the ISPs, as well as the OTTs, are affected by the reaction of the users to low service quality as they are more and more quality aware. Then, both entities must face the risk of user churn which may result into decrease of market share and reputation which is unavoidable in this era of strong competition in this domain.

The recent researches over the past years have revealed that quality perceived by the users not only depends on quality delivered by the network but also on application parameters and subjective factors. Indeed, the Quality of Experience (QoE) is a multidimensional concept in which several influencing factors are involved, such as: human, context, price and application aspects. Accordingly, the eco-system for QoE delivery analyzed in [82] shows that application and network providers are key players involved in QoE delivery and both contribute to the final quality level delivered to the users.

Lately, the research in the field of QoE has been conducted separately with the different prospects of the OTTs and the ISPs which gave birth to two different areas in the field of QoE: application-aware networks and network-aware applications. The network-aware applications [7] aim to adapt the delivery of multimedia contents on the best effort over the network by inducing change in the application parameters, whereas application-aware networks [8] focus on effective management of network according to application requirements. However, the drawback of the above mentioned research stream is that OTTs have no control over network for enhancing users' QoE, whereas ISPs have no availability of application model neither encrypted content nor users' privacy allow them to go for deep packet inspection [9]. Hence, both the ISP and the OTT cannot deliver the best QoE to their valued customers, which results into user churn as well as decrease in market shares.

On the basis of these considerations, in this chapter we focus on the investigation of the impact of a QoE centered OTTs-ISPs collaboration for QoE based service delivery to end users. At the first we discuss some of the technical aspects and impacts of collaboration highlighting the need of OTT-ISP collaboration for QoE based service delivery. Later, on the basis of the possible roles of the OTTs (QoE monitoring and application optimization) and the ISP (QoS monitoring, revenue maximization and network-wide operations), we propose a reference architecture which defines the interfaces and modules required for their interactions providing a baseline for continuous exchange of information/service between the two entities. Then we propose the QoE centered collaboration approach which is driven by the maximization of the revenue based on different factors, such as the user churn (which is modeled as affected by the QoE using the Sigmoid function), pricing and marketing actions. The collaboration is guided by ISP which maximizes the revenue as a function of the delivered QoE with the provision of better network resources on the basis of application QoE model while the OTTs perform the context-aware QoE monitoring and provide the ISP with the information about the class of service per user as well as about application parameters. Finally, with simulations we highlight how the proposed collaboration approach increases the revenue generation and the QoE for both the ISP and the OTT.

The chapter is structured as follows. Section 5.2 discusses the state-of-the-art related works , while Section 5.3 presents the reference architecture, whereas Section 5.4 discusses the proposed collaboration approach. Section 5.5 provides the simulations based on our proposed approach and finally Section 5.6 concludes the chapter and discusses future work.

5.2 Past works

This section reviews the works that propose new algorithms for QoE-centric service delivery, those that focus on OTT-ISP collaboration, and those that address the user churn modeling.

5.2.1 QoE-Centric Service Delivery

The delivery of quality in accordance with end user perception is only possible if the service delivery is QoE centered, i.e., with the inclusion of application specific QoE models in the service delivery process. Accordingly, some of the works found in literature defined QoE centered approaches. In [15], the authors presented a QoE monitoring model based on network and application parameters in Long Term Evolution (LTE) architecture. The provided results highlighted that different applications require different level of network resources on the basis of their QoE models. Similarly, the work presented in [16] proposed QoE based scheduling algorithm for LTE networks where higher scheduling priority is given to packets of mostly used application based on QoE models. The results shown that the VoIP and video streaming required high level of network resources in order to deliver better quality. The case of wireless LAN is addressed in [83].

In [84], Varela et al. highlighted that QoE provision to end user cannot be done with the current Service Level Agreements (SLA) but rather Experience Level Agreements (ELA) would be required to deliver guaranteed QoE. Whereas in ELA the change in Service Level Objectives (SLOs) from mean time to failure or mean time to recovery in QoS parameters to minimum assurance of Mean Opinion Score (MOS) was proposed. The work also proposed agreement between OTTs and ISPs on QoE based SLOs.

Other works addressed the pricing strategies between ISPs and OTTs. In [85], the authors investigated the cases of QoS sold by the ISP to the OTT or to the users. The impact of different QoS pricing strategies were modeled analytically and analyzed with numerical results. It resulted that the ISP may sell QoS to users at a lower price than when QoS is sold to the OTT. Similarly, the studies in [86] proposed a coalition model for CDNs and ISPs based on QoS where CDNs will pay ISPs for better provision of QoS to their traffic. In [87, 88], the authors propose a pricing model based on the network architecture similar to the Paris Metro Pricing (PMP) method proposed in [89]. The PMP aims at partitioning the main network into logically separate channels where each channel has fixed fraction of network capacity and associated price. There would be no guarantees of QoS because packets are always delivered on a best-effort basis. However, the channels with higher prices are expected to be less congested than those with lower prices, resulting in provision of better quality to customers who pay more. The study in [87] demonstrates pricing for the network with two service classes for any number of competing ISPs. From their analysis, they concluded that a network with two service classes is socially desirable, but it could be blocked due to unfavorable distributional consequences, i.e., violation of network neutrality principle. Furthermore, they demonstrated that in the absence of regulation and considerable ISP market power (small), a sizable fraction of the current network users will experience a surplus loss with two service classes. In [88], the PMP method has been integrated with QoE aspects giving birth to PARQUE (Pricing and Regulating Quality of Experience). PARQUE considers two different types of applications (web traffic and video traffic) implying higher QoS requirements for video traffic than for the web traffic. For both the types of application the users' QoE expectations are considered together with the user willingness to pay for the service.

From the results provided by these studies, it can be stated that providing different classes of services to the users on the basis of their willingness to pay can improve quality as well as the revenue. However, when studying network resource allocation among different applications an important factor must always be considered, i.e., the network neutrality (also called Net Neutrality or NN). Although there is no standard definition yet, Net Neutrality principle states that in order to preserve the openness of the Internet, the end users should have equal access to all the content on the Internet, and the ISP should be prohibited from discriminating/blocking the content from any of the application providers [90]. For such principle, the network should deliver traffic in a best effort manner, but lower levels of neutrality violation can be accepted as intrinsic prioritization, load management and blocking of illegal content [90]. In [91], the authors discussed the Net Neutrality with social, economical and technical prospects where authors classify Net Neutrality as a threat to future innovation and technology which may eliminate ISPs incentives to invest in the network.

5.2.2 OTT-ISP Collaboration: Technology Oriented Aspects

Although the collaboration among OTTs and ISPs is catching the eyes of researchers working in QoE-oriented service management, still only few works have really addressed this aspect in the literature. The collaboration between networks and applications in the future Internet is addressed in [92], where the importance of the collaboration between network providers and applications is highlighted by discussing a scenario in which applications give more information about their needs and network usage so that ISPs can allocate network resources more efficiently or even open their network so that applications can dynamically invoke some network services. Two existing collaboration techniques are discussed: the ALTO (Application-Layer Traffic Optimization) [93] and the CINA (Collaboration Interface between Network and Application) [94]. The ALTO initiative allows P2P networks and ISPs to cooperate in order to optimize traffic being generated by P2P applications and transported over the ISP's infrastructure. However, the application-ISP interaction in ALTO only concerns network information provided by ISPs and processed by applications, i.e., the ISP is blindfolded to the services which their customers subscribe to. These limitations are addressed by the CINA interface, which not only allows applications to retrieve information about the network, but also offers the possibility to instantiate network services such as multicast service, caching nodes, and high capacity nodes. Nonetheless, these works are specific for P2P applications and the collaboration between network and application is limited. Furthermore, business aspects are not investigated.

5.2.3 User Churn

According to the study conducted in [70], quality and pricing are considered as major causes for a user to become churner. Nowadays, the users' satisfaction related to a particular service plays an important role in the growth of market share of any company dealing with multimedia services and it has high cross correlation in the prediction of users' churn as well. However, to the best of authors' knowledge, no works can be found in literature regarding users' churn model in terms of quality perceived by the user. In fact, most of works propose utility functions which model the QoE on the basis of network and application parameters. For example, in [95] the Sigmoid function is used to model user satisfaction as a function of QoS parameters, such as delay and error rate, for Internet Protocol Television (IPTV). In [96], the IQX hypothesis is presented, i.e., a generic exponential relationship between user-perceived QoE and network-caused QoS. This relationship has been proved to be valid for some case studies, such as: voice quality as a function of loss and jitter; cancellation rates of web surfer as a function of access link bandwidth. Indeed, these and other related works provide a QoE measure in function of specific QoS and application parameters, i.e., they can be useful for the monitoring of end-to-end system parameters. However, what is missing in the state-of-the-art is a model which is able to estimate the influence of the QoE in causing customer churn for telecommunications services.

5.3 Reference architecture for Collaboration

The reference scenario is composed of an ISP which provides network infrastructures and services, and different OTTs that provide over-the-top applications. The major aspect that links the OTTs with the ISP is the QoE delivered to the final users, which can be selected as the core component for building collaboration strategies towards service delivery. As a matter of fact, the OTT is aware of users' expectations and the level of quality they are experiencing, thanks to the control of the software at the application level and a *close* relationship with the user. Indeed, through the application software it can monitor application parameters (such as buffer occupancy and playout delay in video streaming applications) and context parameters (such as the type of device and the position of the user), and can even ask the user to fill surveys about quality satisfaction. However, it cannot have any control on the network. On the other hand, the ISP is more focused on QoS and controls network resources provided to all of its users; however, not always better ISP provided QoS has a positive effect on QoE.

Therefore, since the OTT is the entity which is more QoE-oriented, a collaboration between the OTTs and the ISP can help the ISP to implement a QoE-aware network management for the provisioning of adequate QoE to the end-users. Fig. 5.1 sketches the reference architecture of the collaboration scheme we focus on. We provide a high-level architecture which defines a set of functional requirements that must be provided by OTTs and ISP for making possible the collaboration approach. Since it is a functional architecture, we do not provide any specification about how to implement the functional blocks nor recommendations are given about the network interfaces to be used for information exchange. We assume that multiple OTTs decide to collaborate with a single ISP. The OTTs monitor the QoE of their users using QoE models which are specific for the application they are providing to their users. This is the role of the *QoE monitoring* block, which measures the QoE as a function of application parameters and context parameters (extracted from user profile information) such as user location, user's device, user's expectations, etc. The QoE measurements are then conveyed to the ISP trough a dedicated interface, together with the information about the class of service of the users. In fact, a dedicate communication channel is established between each OTT and the ISP, to allow for the transmission of information between OTTs and ISP. Such a channel is interconnected at both the ISP and OTT ends with a functional block implementing Authentication, Authorization and Accounting (AAA) functions for a secure information exchange.

On the other hand, the *QoS monitoring* block of the ISP monitors the QoS of the network through which all the OTT applications are provided to the end-users. QoS and QoE measurements are then received and collected by the *Network management* block, which runs a QoE-aware network management algorithm which aims at controlling the QoE by looking at its impact on the user churn and then on the revenue. Specifically, on the basis of a model of the revenue, the best combination of price and QoE level is computed using (5.5). Then, relevant requirements are taken for providing network resources to quality-demanding users for maximizing the revenue. However, these decisions must take care of not putting other users at a disadvantage neither discriminating other OTTs (for assuring Net Neutrality).

5.4 Collaboration model

We assume that the multiple OTTs services are passing through an ISP network and they agree upon the collaboration on the basis of the roles and interfaces defined in the reference architecture proposed in Section 5.3. The reason for consideration a single ISP is to simplify the treatment. However, the proposed solution can be extended to multiple ISPs scenario without any issue of scalability. Nonetheless, the proposed collaboration requires a common ground among ISP and OTTs in the form of key models related to: revenue generation, QoE, user churn, pricing and marketing. The way these models are used in our proposal is discussed in the following subsections in the following order: the pricing model, the revenue model for the collaboration, the user churn modeling and the revenue maximization approach.

5.4.1 Pricing Modeling

An important point of the collaboration between the ISP and the OTTs is the definition of the pricing model, i.e., the economic rules which define how much the users should pay for

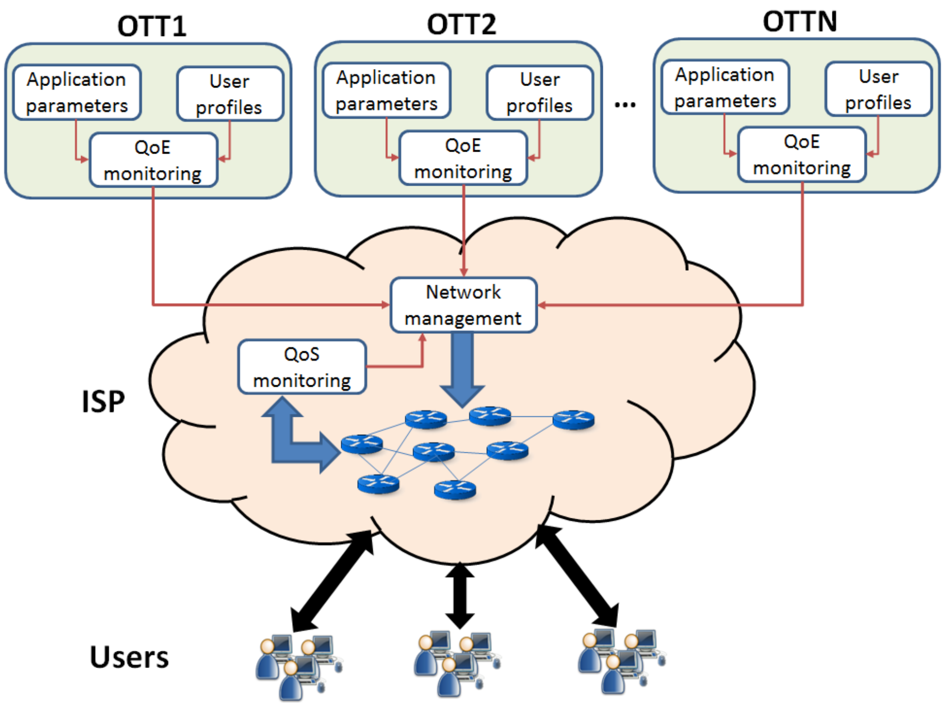

Fig. 5.1 Reference architecture for the collaboration between ISP and OTTs.

accessing the combined ISP-OTT services. As analyzed in [97], network congestion brought to the birth of Smart Data Pricing (SDP), i.e., a suite of pricing and policy practices that have been proposed by operators as access pricing options instead of the traditional flat-rate model. One application of SDP is the pricing for end-user QoE, i.e., pricing strategies for matching the operator's cost of delivering bits at the consumer's QoE needs for different application types at the price the customers are willing to spend. SDP approaches are mainly classified into static and dynamic models depending on whether the prices are changed in real-time or on a longer timescale. In this chapter, we rely on a static model: the Paris Metro Pricing (PMP) concept for Internet pricing proposed in [89]. Although the PMP model charges different prices for different network channels, there are no guarantees of QoS for the users which pay more. However, it is expected that the channels with higher prices would be less congested than those with lower prices, resulting in higher quality provided to customers paying more.

For our pricing model, we propose an enhanced version of the PMP model which assures a minimum guaranteed quality to the users depending on the money they pay. Accordingly, we assume that the services provided by the ISP-OTTs collaborations are being offered in J different levels of quality (with $j = 1, 2, 3, \ldots, J$ indexing the different levels) and different prices, and that the higher the price the better is the expected (and provided) quality. The users are assigned to a specific class in function of their willingness to pay W^n, where n indexes the user. For simplicity, as in [88] we consider normalized W^n and normalized prices so that $W^n \in [0,1]$ and $P_{i,j} \in [0,1]$, where $P_{i,j}$ is the price to be paid to subscribe to the j-th class of service of application i. As a general example, a user n will subscribe to the service class j if $P_{i,j} \leq W^n < P_{i,j+1}$.

Till here, this is a QoS-based pricing model which aims at providing higher system performance to users paying more, i.e., different Service Level Agreements (SLAs) are defined between the providers and the users as a function of their willingness to pay. However, SLAs are difficult to be understood by the users and are not directly related to their perceived quality. Therefore, inspired by the concept proposed by Varela et. al in [84], we define the quality provided by each class of service in terms of Experience Level Agreements (ELA). An ELA is defined as *a special type of SLA designed to establish a common understanding of the quality levels that the customer will experience through the use of the service, in terms that are clearly understandable to the customer and to which he or she can relate*. Therefore, we decided to represent the quality provided by each class of service with a star rating (from 1 to 5 stars), where the stars have the same meaning of the rating values defined by the Mean Opinion Score (MOS), i.e., 1 star means “Bad quality", 2 stars mean “Poor quality", 3 stars mean “Fair quality", 4 stars mean “Good quality", and finally 5 stars mean “Excellent quality". However, any other representation method can be used as an alternative.

At this point, a question arises: how can quality levels be defined? Specifically, how can the collaborating ISP-OTT providers decide which QoS and application parameters provide the users with a certain quality level? This is an important point as we know that the QoE depends on many different factors, ranging from objective QoS and application parameters to more subjective factors such as the context in which an application is used (used device, environment, time of the day, social factors, etc.) and human factors (user’s expectation and experience, user’s sensitivity, etc.). In this chapter, we assume to use existing (and future) QoE models depending on the considered applications, e.g., VoIP and video streaming, which investigate how the QoE perceived by the users varies in function of network and application impairments. Some use cases are addressed in the simulation section.

5.4.2 Revenue Modeling

Recall that our collaboration is driven by the maximization of the revenue for both the service providers, for which we need to define an appropriate model. According with the price model proposed in the previous section, here we provide a model for the revenue computation. The OTT-ISP revenue clearly evolves over the time due to several factors, such as the price and the QoE. We then consider the revenue as a discrete-time process where t_x $(x = 0, 1, 2, ...)$

indexes the time instants at which the revenue is computed and corrections to the system are introduced. Furthermore, we define the time window $T = t_{x+1} - t_x$ as the period of time during which the prices of the classes of service and the number of users belonging to each class are static.

The combined revenue for the i-th OTT and the ISP (we are not separating the revenue) can be computed as follows

$$R_i^x = \sum_{j=1}^{J} N_{i,j}^x \cdot P_{i,j} \tag{5.1}$$

where $i = 1, 2, 3, ..., I$ indexes the OTTs collaborating with the ISP and $N_{i,j}^x$ is the total number of users belonging to the j-th class calculated at time t_x for the i-th OTT service. $P_{i,j}$ is the price to be paid for subscribing to the j-th class of service of the i-th OTT application. Accordingly, the total revenue generated by the collaboration between the ISP and all the OTTs can be calculated as

$$R^x = \sum_{i=1}^{I} R_i^x = \sum_{i=1}^{I} \sum_{j=1}^{J} N_{i,j}^x \cdot P_{i,j} \tag{5.2}$$

The evolution of $N_{i,j}^x$ over the time depends on the churn effect, i.e., the process of users leaving the service. Hence, we consider that the users being represented by $N_{i,j}^x$ are the active users of both the i-th OTT and ISP, i.e., the user continuing the services in j-th class. This number then evolves over time due to the churn and due to the activation of new contracts, as follows

$$N_{i,j}^x = N_{i,j}^{x-1} \cdot U_{i,j} + \zeta_{i,j} \tag{5.3}$$

where $U_{i,j}$ is the user churn function that is defined in the Section 5.4.3 and $\zeta_{i,j}$ is the number of users joining the j-th class of collaborative service of i-th OTT through advertisement. Indeed, studies conducted in [98] emphasis that mostly the companies gain their customers by effective marketing/advertisement campaigns, which is something considered in our modeling but not controlled by our strategy. Specifically, the study in [99] emphasized that the Poisson distribution can be utilized to predict the increase in the market share in telecommunication. Hence, we consider $\zeta_{i,j}$ as a stochastic process which follows a Poisson distribution depending upon marketing strategies, socio-economic factors and product discounts.

5.4.3 Churn Modeling

The user satisfaction to a service plays an important role in the reputation of any service provider in the market. Lowering the level of user satisfaction may result into high level of user churn, i.e., reduction of the number $N_{i,j}^{x}$ of active users. Notwithstanding the importance of this phenomenon, only limited works exist about the study of the impact of QoE on the user churn. One major obstacle is that to predict/model user churn in terms of QoE requires data over long periods of observation from both OTT and ISP. Still, to go ahead with our analysis, we consider that there is a high cross-correlation between user satisfaction and user churn and we build a user churn function based on the Sigmoid function [95]. Indeed, it is one of the mostly used activation function in Multi-layered Perceptron Neural Networks in the field of artificial intelligence to model human perception into machine [100, 101]. We consider the user churn function as upward criterion function, i.e., the function increases with the increase in QoE, which means that more users will be continuing the service if higher QoE is provided, and vice versa. The user churn function can be defined mathematically in terms of QoE as follows

$$U_{i,j}(QoE_{i,j}) = \frac{1}{1+e^{-z(QoE_{i,j}-QoE_{i,j}^{m})}} \tag{5.4}$$

where $QoE_{i,j}$ is the quality delivered to the j-th class of service of the i-th OTT service, whereas $QoE_{i,j}^{m}$ is the quality level at which half of the paying users leave the service in the class j (i.e., $U_{i,j}(QoE_{i,j}^{m}) = 0.5$). Moreover, the sensitivity of the users with regard to the price paid is represented by z. In fact, users who pay more expect to receive a better quality than those who pay less, and the users keeping the service for the former class of service must be lower than that of the latter, for the same value of QoE perceived. Hence, the higher the price paid the smaller the z, i.e., higher the sensitivity of the user with the quality. Fig. 5.2 shows an example of the user churn function for different values of $QoE_{i,j}^{m}$ and z. The user churn function ranges in the interval $[0,1]$ where 1 means that the 100% of the users are keeping the service. The QoE is measured as for the MOS in the interval $[1,5]$ where 1 means minimum quality and 5 maximum quality. In the example shown in Fig. 5.2 there are two different groups of curves: the continuous curves refer to the lower class of service whose users have lower QoE expectations and for this reason although the perceived QoE is 2.5 half of the users will be keeping the service. On the other hand, dotted curves refer to the higher class of service, whose users are paying more and therefore have higher QoE expectations. In fact, in this case half of the users will be keeping the service for a MOS at least of 4, which means that half of users for being satisfied and keeping the service expect a

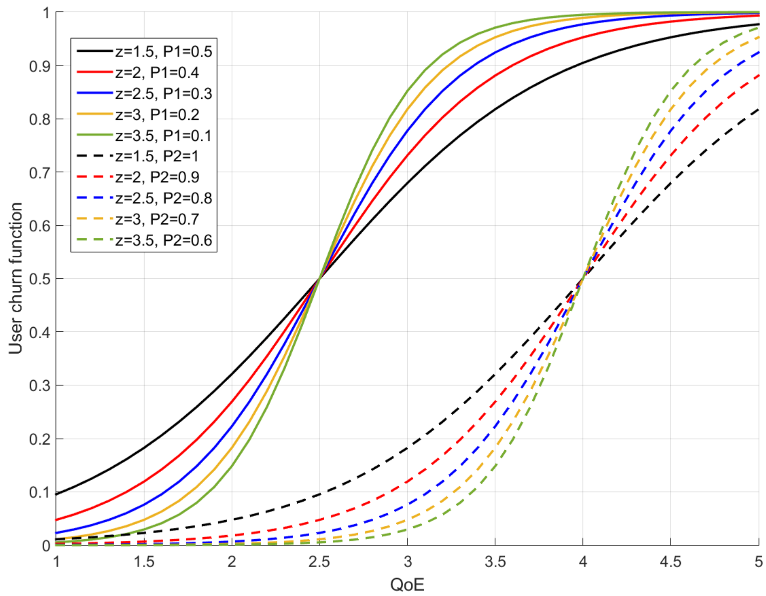

Fig. 5.2 User churn function for different values of $QoE^m_{i,j}$ and z. Continuous curves refer to $QoE^m_{i,1} = 2.5$ whereas dotted curves refer to $QoE^m_{i,2} = 4$.

more than good QoE. The different values of z identifies the different sensitivity of the users and depends on the price paid to be subscribed to that class of service.

It is important to note that we defined the user churn function following the recommendations in [102]: 1) the user churn function follows the characteristics of the users' QoE; 2) the user churn function does not change drastically with small changes in the QoE. Moreover, mathematically the proposed user churn function is valid in accordance with the law of diminishing marginal utility, which implies three conditions:

1. Concavity of $U_{i,j}(QoE_{i,j}), \forall QoE_{i,j} \geq QoE^m_{i,j}$
2. Convexity of $U_{i,j}(QoE_{i,j}), \forall QoE_{i,j} \leq QoE^m_{i,j}$
3. $U'(QoE_{i,j}) \geq 0$

We also want to stress that the proposed user churn function can be better calibrated when data about user behavior is available.

5.4.4 Revenue Maximization

With the complete modeling of the revenue we can now achieve the target of its maximization with a coordinated control of OTTs and ISP. Specifically, they target the maximization of an average revenue computed during a reference period $t_{X_1} - t_{X_2}$ as follows

$$\bar{R}^* = \max_{QoE_{i,j}, P_{i,j}} \left[\sum_{x=X_1}^{X_2} \left(\frac{\sum_{i=1}^{I} (\sum_{j=1}^{J} (N_{i,j}^{x-1} \cdot U_{i,j} + \zeta_{i,j}) \cdot P_{i,j}}{t_{X_2} - t_{X_1}} \right) \right] \tag{5.5}$$

where the influence of $QoE_{i,j}$ is taken into account by the user churn function $U_{i,j}$. Even if not explicitly highlighted here, the different combinations of the $QoE_{i,j}$ need to be considered under the available network resources. Then, the collaboration between the OTT and the ISP is fundamental for delivering adequate QoE to the users. The OTT is able to know QoE expectations of the user and to measure the QoE delivered, while the ISP supports the OTT by providing the needed network services. The output of this maximization are the $QoE_{i,j}^*$ levels and prices $P_{i,j}^*$ for the different service classes. This maximization will be done at the ISP side in order to assure better quality to the end users in accordance with the reference architecture provided in Section 5.3, whereas through the interface defined between the OTTs and the ISP the maximization results can be shared with the respective OTTs involved in collaboration.

Herein, we want to highlight that how the OTT and the ISP decide to divide the revenue is out of the scope of the chapter. However, a joint venture approach can be considered for the revenue sharing where an involved enterprise can get the share of revenue proportional to the amount of investment made by that enterprise over the total amount of investment for the service delivery. For an ISP, the investment may occur in the form of maintenance, up-gradation and operations of the access/core network infrastructure while an OTT can make investment in specialized data centers, multimedia streaming servers, content delivery networks and data-center networks. Additionally, it is important to note that this approach can be implemented without violating the Net Neutrality principle in several ways: better QoS could be provided by hosting OTT content in ISP nodes and using underutilized network areas without affecting the other OTTs' traffic; traffic could be prioritized if it does not affect the final QoE of the applications of which the traffic flows could be delayed.

5.5 Simulation

The objective of the conduced simulations is an analysis of the potential of the proposed collaboration approach. Specifically, we consider an ISP and 2 OTTs and we investigate their revenue generation for two different approaches: No Collaboration (NC) and Joint Venture (JV). The former does not consider a collaboration between the OTTs and the ISP so that the OTTs deliver their contents through the best effort service over the ISP network. The JV is the collaboration approach described in Sections 5.3 and 5.4, i.e., the OTTs collaborate with the ISP with the objective of maximizing the revenue. Without loss of generality, we focused on two specific OTT services: video streaming and VoIP. In Section 5.5.1 we discuss the QoE models used to evaluate the QoE perceived by the end-users whereas in Section 5.5.2 we present the simulation settings and results.

5.5.1 QoE models

As a use case, we consider two different OTT applications: video streaming and VoIP. The reason for the selection of the aforementioned applications is in accordance with the studies conducted in [15, 16] that considered video streaming and VoIP as the most sensitive multimedia applications with reference to network resources usage. For the evaluation of the QoE we based on the model proposed in [103] for the video streaming application and on the E-Model for the VoIP application [104].

The model proposed in [103] is a parametric packet-layer model for monitoring the video quality of IPTV services, which measures the QoE provided by HD (1440x1080) videos encoded with the H.264 codec at different bitrates and corrupted by packet loss:

$$QoE_{video}^{HD} = 1 + \left(v_1 - \frac{v_1}{1 + \left(\frac{BR}{v_2}\right)^{v_3}} \right) \exp\left(-\frac{PLR}{v_4} \right) \tag{5.6}$$

where $v_1 = 3.8$, $v_2 = 4.9$, $v_3 = 3.6$ and $v_4 = 3.5$ are the coefficients of the model while BR and PLR are the source coding rate of the video and the packet loss rate of the network, respectively. We consider this model for evaluating the QoE provided by HD videos because it considers both application (source coding rate) and network (PLR) parameters and because the cross-correlation factor computed between the proposed model and subjective QoE results is greater than 0.9 with 99% confidence interval.

In [105], the authors extended the model in [103] considering also the movement of the video content, the MPEG-2 codec and different video resolutions. The model is as follows:

$$V_q = 1 + 4K \left(1 - \frac{1}{1 + \left(\frac{b \cdot BR}{v_5}\right)^{v_6}} \right) \tag{5.7}$$

$$K = 1 + k_1 \exp(-k_2 \cdot b \cdot BR) \tag{5.8}$$

where, for videos with medium content movement, encoded with the H.264 codec at SD resolution, $v_5 = 0.67$, $v_6 = 1.4$, $b = 1$, $k_1 = 1.36$ and $k_2 = 1.93$. Also the model in eq. (5.7) can take into account the effect of the PLR if multiplied for the exponential factor of eq. (5.6). Therefore, we consider the model in eq. (5.9) for evaluating the QoE provided by SD videos:

$$QoE_{video}^{SD} = 1 + 4K\left(1 - \frac{1}{1 + \left(\frac{BR}{v_5}\right)^{v_6}}\right)\exp\left(-\frac{PLR}{v_4}\right) \tag{5.9}$$

Both the models in eq. (5.6) and eq. (5.9) measure the QoE with values ranging from 1 (Bad quality) to 5 (Excellent quality) as the MOS.

The E-Model is a planning parametric model defined by the ITU for VoIP applications, which measures the voice quality in terms of the R-factor, i.e., a quality index ranging from 0 to 100, where 100 is the best quality. The R-factor is defined in terms of several parameters as follows

$$R = 100 - I_s - I_d - I_{ef} + A \tag{5.10}$$

where I_s is the signal-to-noise impairment, I_d is the impairment associated to the mouth-to-ear delay of the path, I_{ef} is the equipment impairment associated with the losses within the codecs and A is the advantage factor which allows for compensation of impairment factors when the user benefits from other types of access to the user. The study in [106] presented an adapted version of the E-Model (see eq. (5.11)), which emphasizes the effect of sources of quality degradation observed over data networks, namely one-way delay, packet loss ratio, and coding scheme. The adapted model is

$$R = 94.2 - I_d(d) - I_e(CODEC, PLR) + A \tag{5.11}$$

where I_d and I_e capture the quality degradation caused by delay and equipment impairment factors, respectively. d is the mean one-way delay of played voice packets during an assessment interval, *PLR* is the packet loss ratio, and *CODEC* is the used speech encoding scheme.

The quality degradation caused by one-way delay when echoes are perfectly removed are calculated as

$$I_d(d) = 0.024 \cdot d + 0.11 \cdot (d - 177.3) \cdot H(d - 177.3) \tag{5.12}$$

where

$$H(x) = \begin{cases} 1, & x < 0 \\ 0, & x \geq 0 \end{cases} \tag{5.13}$$

On the other hand, the quality degradation caused by equipment impairment factors are calculated as

$$I_e(CODEC, PLR) = a_1 + a_2 \cdot ln(1 + a_3 \cdot PLR) \tag{5.14}$$

where a_1, a_2 and a_3 are coefficients obtained through a logarithmic regression analysis depending on the used speech codec. For example, for the G.729a codec $a_1 = 11, a_2 = 40$ and $a_3 = 10$ whereas for the G.711 codec $a_1 = 0, a_2 = 30$ and $a_3 = 15$.

With regard to the advantage factor, the default value of A in case of conventional wirebound communication system is $A = 0$. The maximum values of A are provided in [104] for different scenarios. For example, $A_{MAX} = 5$ in case of mobility by cellular networks in a building and $A_{MAX} = 10$ in case of mobility in a geographical area or moving in a vehicle.

Furthermore, in [106] is also provided an equation for converting the R-factor with values between 1 and 5 as the MOS:

$$QoE_{VoIP} = \begin{cases} 1, & R < 0 \\ 4.5, & R > 100 \\ 1 + 0.035R + 7 \cdot 10^{-6} R(R-60)(100-R), & 0 < R < 100 \end{cases} \tag{5.15}$$

We consider the model in eq. (5.15) for evaluating the QoE provided by the VoIP application, where R is computed with eq. (5.11).

5.5.2 Simulation results

The simulation scenario considers two OTT applications which are delivered to their users through a network owned by an ISP. For simplicity we assume that the users are stationary and located in the same area, where the Internet access is provided by the ISP. For both the approaches (NC and JV) and for both the applications (video streaming and VoIP), the users can choose between two different plans: standard plan (service class 1) at price $P_{i,1}$ and premium plan (service class 2) at price $P_{i,2}$, with $P_{i,1} < P_{i,2}$. The subscript i identifies the OTT application. We consider normalized prices so that $P_{i,1}, P_{i,2} \in [0, 1]$. Each user subscribes to one of the two proposed plans on the basis of his/her willingness to pay $W^n_{i,j}$, where n indexes the user. We assume that the user is at least a standard user, then $W^n_{i,j} \in [P_{i,1}, 1]$. As a consequence, if $P_{i,1} \leq W^n_{i,j} < P_{i,2}$ the user is a standard user, while if $W^n_{i,j} \geq P_{i,2}$ the user is a premium user. Therefore, for the application i, there will be $N_{i,1}$ users subscribed to the standard plan and $N_{i,2}$ users subscribed to the premium plan, while the total number of users N_i will be $N_i = N_{i,1} + N_{i,2}$.

On the application side, with regard to the video streaming application, standard users can watch videos only at Standard Quality (SD), i.e., with a resolution of 720x480 pixels, whereas premium users can watch videos at HD quality, i.e., with a resolution of 1440x900 pixels. On the other hand, both standard and premium VoIP users have access to standard VoIP services (calls, phone conferencing, etc.) whereas only premium users can have access to extra services such as recording functions, voicemail, etc. We selected the G.729 codec for VoIP simulations because it provides good performance and requires a low bandwidth ($31.2kbps$ [107]).

On the network side, there is a difference between the NC and JV approaches. In fact, while for the NC approach the applications are delivered on the best effort network, for the JV approach the ISP provides different network resources to the standard and premium users of the applications. Specifically, for the video streaming application, a minimum bandwidth of $2Mbps$ and $5Mbps$ is guaranteed to standard and premium users, respectively. In fact, generally a HD video is encoded at a bitrate ranging from $1.5Mbps$ to $4Mbps$ whereas a SD video is encoded at a bitrate ranging from $500kbps$ to $2Mbps$ [108]. Furthermore, a PLR lower than 0.3% is guaranteed to premium users whereas for standard users the maximum PLR will be 1.5%. These PLR values are selected on the basis of the study in [103] where the influence of the PLR on the QoE for video streaming has been investigated. With regard to the VoIP application, on the basis of the study in [106], a one-way delay lower than $100ms$ and a PLR lower than 1% are guaranteed to premium users whereas for standard users the maximum one-way delay and PLR will be $350ms$ and 5%, respectively. As discussed in Section 5.4.1, we express the quality of the service classes in terms of ELA and therefore we assume that with these network and application parameters the JV approach can provide at least a quality of 3 (Fair quality) to standard users and of 4 (Good quality) to premium users.

For the NC approach, the total bandwidth is divided in equal parts to each user with no guarantee of minimum bandwidth provided. Furthermore, for PLR and delay we consider the same maximum values selected for the standard users of the JV approach.

We conducted simulations with the Matlab software setting a starting number of users $N_{VoIP} = N_{Video} = 100$ and considering a total bandwidth of $500Mbps$. Since we based on the PMP pricing model, $P_{i,1}$ ranges from 0.1 to 0.5 while $P_{i,2}$ ranges from 0.6 to 1.0 with a step of 0.1 [87]. We consider $P_{VoIP,1} = P_{Video,1}$ and $P_{VoIP,2} = P_{Video,2}$. For each combination of the prices $P_{i,1}$ and $P_{i,2}$, we randomly assign a willingness to pay (uniform distribution between 0 and 1) to each user and on the basis of this value the user is assigned to the standard or premium classes of the VoIP and video streaming applications. We want to highlight that for simplicity we considered the willingness to pay uniformly distributed between 0 and 1. This way, the higher the price for joining the class of service the lower the number of users joining

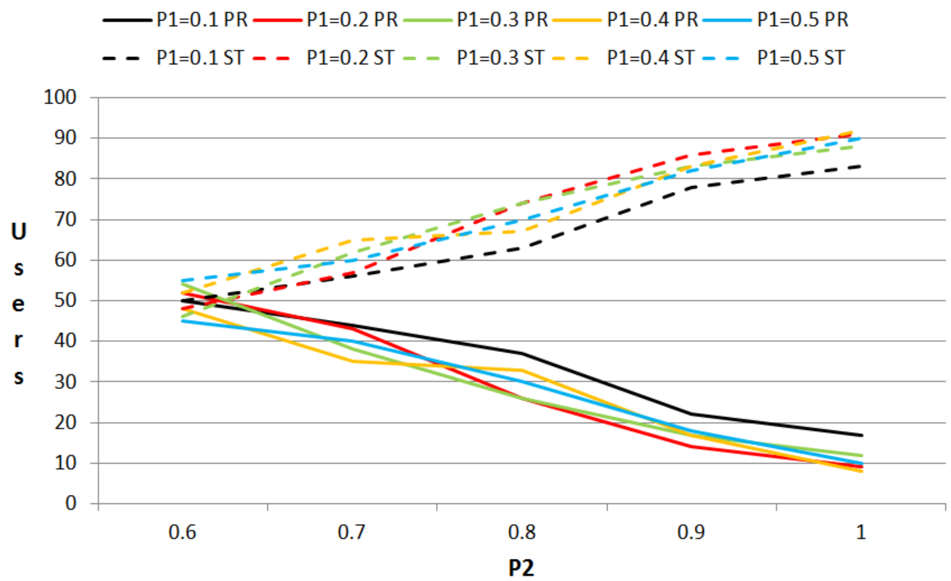

Fig. 5.3 Number of starting users for standard (ST) and premium (PR) classes of service as a function of the prices $P_{i,1}$ and $P_{i,2}$.

that class. As an example, in Fig. 5.3 we show the starting number of users in function of the prices $P_{i,1}$ and $P_{i,2}$. For example, within the population of 100 users, fewer users will join the premium class and more user will join the standard class as $P_{i,2}$ approaches 1. However, the willingness to pay distribution only influences the starting number of users joining the classes of service while the number of users keeping or leaving the service in next months depend on the user churn model based on the user's QoE. Therefore, we expect that using different willingness to pay distributions will bring to the same revenue results in the long period.

Once the starting number of users for each class of service are assigned, we compute the starting video and VoIP revenue with eq. (5.2). We want to investigate the revenue of the 2 OTTs for the following 24 months by using eq. (5.5) for the revenue maximization and the QoE models discussed in the previous section for the QoE evaluation. With regard to $U_{i,j}$, we set $QoE^m_{i,1} = 2.5$ for the standard service and $QoE^m_{i,2} = 4$ for the premium service, since premium users have greater QoE expectations than standard users. In Fig. 5.2, we show the user churn function for different values of z and $QoE^m_{i,j}$. With regard to $\zeta^x_{i,j}$, we computed it as a random number from the Poisson distribution with the mean equal to the 5% of users belonging to the j-th class and application i at time $(x-1)$. The time range x in this case is a month. Within each month, we compute 100 QoE measurements and we use the resulting average QoE for maximizing the revenue in eq. (5.5).

Fig. 5.4 shows the revenue obtained with the two approaches by the two OTTs in the first 2 years as a function of the prices $P_{i,1}$ and $P_{i,2}$. We did not provide the graphs of all the

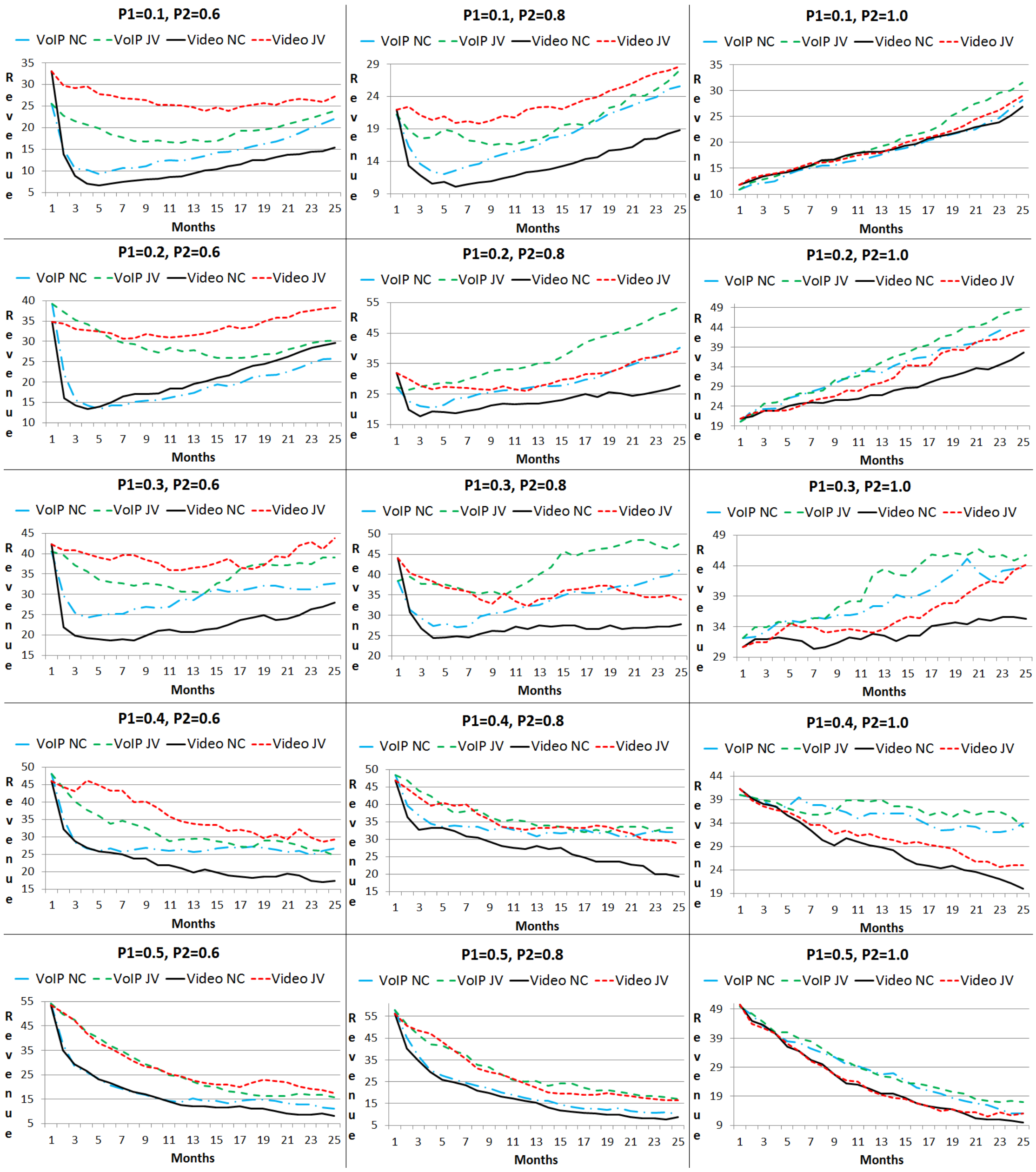

Fig. 5.4 Revenue obtained with the two approaches by the two OTTs in the first 2 years as a function of the prices $P_{i,1}$ and $P_{i,2}$.

combinations of prices to save space, but the graphs provided allow to understand how the revenue evolves with the time for major scenarios. The most evident result is that for each prices combination the revenue obtained with the JV approach is always greater than that obtained with the NC approach for both the video and VoIP applications. This is mainly due to the fact that with the JV approach the OTTs collaborating with the ISP are able to satisfy the QoE expectations of both the standard and premium users. Specifically, the premium users are those who contribute to the revenue difference between the two approaches. In fact, standard users are less QoE demanding and they are the main contributors to the revenue generation in the case of the NC approach. Indeed, from the graphs, it is evident that when the price for the premium service is accessible to many users ($P_{i,2} = 0.6$ and $P_{i,2} = 0.8$), the NC approach fails to satisfy premium users, resulting in a great revenue drop, which is balanced over the time only thanks to the revenue provided by the standard users. When the price for the premium service reaches the highest value (i.e., $P_{i,2} = 1.0$), the standard users are prevalent with respect to the premium users and the difference between the two approaches is less evident although the JV approach provides quite higher revenue for both the applications.

From Fig. 5.4, it can also be noticed another interesting result concerning the z parameter, which represents the sensitivity of the user to the price, as shown in Fig. 5.2. In fact, with the increasing of $P_{i,1}$, the standard users become more QoE demanding and are more likely to leave the service if the QoE provided is not adequate. Indeed, for $P_{i,1} = 0.4$ and $P_{i,1} = 0.5$ the revenue is not more increasing over the time as for the lower values of $P_{i,1}$, but is decreasing because not all the users are satisfied by the quality of the perceived service.

Fig. 5.5 shows the QoE provided by the video and VoIP applications for the NC and JV approaches. “ST" and “PR" stand for standard and premium service, respectively. With regard to the VoIP application for the NC approach there is no distinction between standard and premium services because the considered QoE model is a function of the only network parameters and in the case of NC approach the ISP does not guarantee any network parameter to premium users. Then, the same QoE is provided to standard and premium users. The difference between standard and premium users in this case are the extra application features which cannot be evaluated with current QoE models.

The QoE values are the average QoE computed over all the simulation cycles, and the error bars show the minimum and maximum QoE values provided. It is evident that the JV approach is able to provide a great and stable QoE to the premium users of both the applications, which results in a significant revenue generation as discussed before. On the other hand, the NC approach fails in this objective, providing to premium users a QoE even lower than that provided to standard users. With regard to the standard users, the two approaches provide comparable QoE for both the applications.

Concluding, with regard to the JV approach, the best trade-off is obtained for $P_{i,1} = 0.3$ and $P_{i,2} = 0.6$, with a quite constant revenue with an average of 40 and 35 for video and VoIP application, respectively. On the other hand, for the NC approach the most convenient prices are $P_{i,1} = 0.3$ and $P_{i,2} = 1.0$, with an increasing revenue with an average of 35 and 45 for the video and VoIP application, respectively. However, with these prices and considering the low

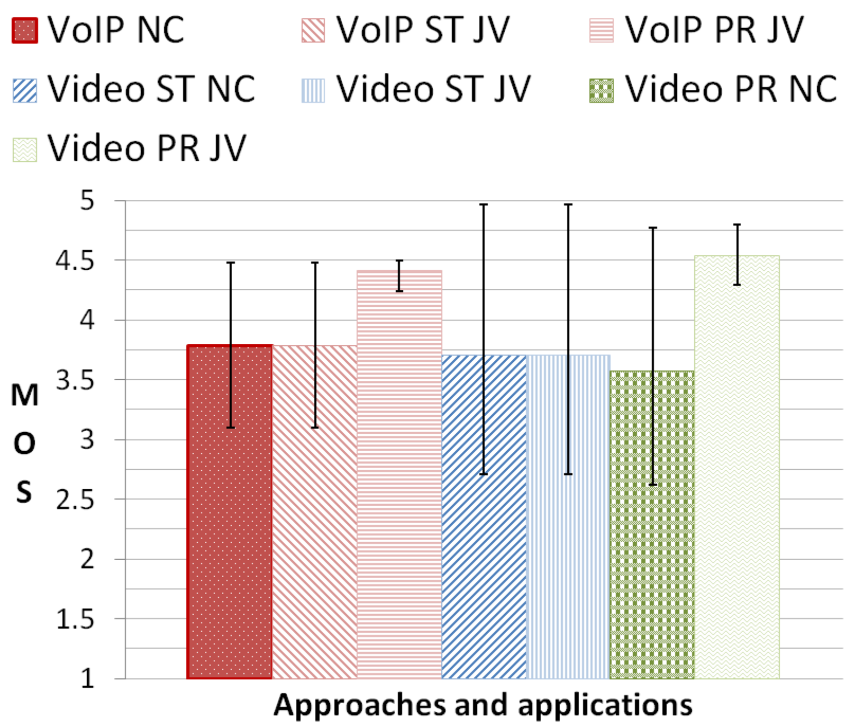

Fig. 5.5 QoE provided by the video and VoIP applications for the NC and JV approaches.

QoE provided to premium users, it does not make any sense to offer two service classes to the users but it would be better to restrict to the only standard service.

5.6 Conclusion and Future Work

In this chapter, we proposed a novel approach for collaborative QoE management between OTTs and ISPs. Differently from QoS based collaboration models found in literature as ALTO and CINA, the proposed collaboration approach is completely QoE centered based on maximization of revenue. Moreover, the proposed model also takes into account important factors such as the user churn, pricing and marketing, making it novel. Also, with the consideration of the different QoE models for different applications we investigated the flexibility and adaptability of our collaboration model which has proven to be robust, reliable and adaptable with respect to any change in QoE model. Furthermore, with simulations we highlighted that if ISPs and OTTs adopt the proposed collaboration approach they will not only increase the revenue but will also provide better QoE to their users with relatively lower prices.

Though the QoE based service delivery requires the collaboration among OTTs and ISPs, the research in this domain is suffering from key challenges. One of these is that no inter-operable interface exists to date which contributes towards scalability of the approach. Hence, it will not only require standardized interfaces among OTTs and ISPs to exchange QoE based information but it will also require standardized interfaces among ISPs like peering connections or exchange points to share QoE related information. Therefore, the future research should focus on the provision of QoE-centric interfaces between OTTs-ISPs to enable them QoE based service delivery. The Software Defined Networks (SDN) and Network Function Virtualization (NFV) can provide an opportunity in this regard because of their programmability and flexibility. However, scalability and security remains as an open issue in collaboration even if centralized SDN controller will be used for the QoE management. The computational complexity for the QoE measurements may appear to be another issue contributing to the scalability of the collaborative approach and complexity may increase with the increase in the number of OTT applications and customers. Moreover, there will be the requirement of storing data related to QoE, user churn and revenue generation which may increase the cost of network planning and operations as well.

Additionally, although big efforts have been conducted in QoE modeling, most of these are developed for providing an estimation of the perceived quality for very short periods of time. This aspect raises a practical issue when applying the resulting models to the considered scenario where the QoE affecting the churn should provide the level of experience quality resulting from longer periods of service consumption. Hence, the development of a robust and reliable QoE models valid for longer periods of time is essential for QoE based service delivery. Notwithstanding the importance of user churn, no model has been proposed which can correlate user churn with QoE which is important at the enterprise level. Nevertheless, the creation of user churn prediction model will be requiring the real customer data and analysis of that data over the significant periods of time. Moreover, the Network Neutrality and user privacy is also another future challenge for collaborative QoE based service delivery. All these challenges need to be taken into account in the future research, so that QoE based service delivery can be possible.

Chapter 6

A Customer Lifetime Value Based Approach

Abstract

In this work, we propose a QoE-aware collaboration approach between Over-The-Top providers (OTT) and Internet Service Providers (ISP) based on the maximization of the profit by considering the user churn of Most Profitable Customers (MPCs), which are classified in terms of the Customer Lifetime Value (CLV). The contribution of this work is multifold. Firstly, we investigate the different perspectives of ISPs and OTTs regarding QoE management and why they should collaborate. Secondly, we investigate the current ongoing collaboration scenarios in the multimedia industry. Thirdly, we propose the QoE-aware collaboration framework based on the CLV, which includes the interfaces for information sharing between OTTs and ISPs and the use of Content Delivery Networks (CDN) and surrogate servers. Finally, we provide simulation results aiming at demonstrating the higher profit is achieved when collaboration is introduced, by engaging more MPCs with respect to current solutions.

6.1 Introduction

The recent developments in the multimedia industry, especially Over-The-Top services (OTTs) such as YouTube, Netflix, Skype, Facebook, have changed the Internet traffic to multimedia traffic. Indeed, with the recent findings, the Internet traffic is predicted to be 75% multimedia traffic by 2020 [78]. Such a drastic increase of multimedia traffic requires Internet Service Providers (ISPs) to upgrade their networks to deal with issues regarding the quality delivered to end users because most of the time this is degraded due to bottlenecks in the ISPs' networks. Moreover, within the multimedia service delivery chain, ISPs are the most affected entities by the user churn because most of the users switch their Internet connection to another ISP. Consequently, this leads to decrease in market share and reputation as well as lower revenue for an ISP.

Since Quality of Service (QoS) is a system centered measure of service quality, in the last decade the research has been shifted to Quality of Experience (QoE), which is a function of QoS but reflects the quality as perceived by the user. Indeed, the recent research reveals QoE as a multidimensional metric based on several influencing factors such as human, economic, network, application and context [4]. Subsequently, this shift of research paradigm towards QoE is leading to a big change in network monitoring, modeling and management techniques from Key Performance Indicators (KPIs) to Key Quality Indicators (KQIs) [9, 22]. However, although network and application providers share the common objective of providing better service quality to their customers, to date research in this field is divided into two separate paths: on the one hand network-aware application management approaches adapt application parameters on the basis of the monitored network status; on the other hand, application-aware network management approaches adapt network resources on the basis of the services to be delivered through the network. Delivering adequate QoE to their customers has become fundamental for network and application providers in order to decrease the user churn and consequently increase their profits. For this reason, in our previous work, we investigated the implementation of a QoE centered collaboration approach, between OTT and ISP, driven by revenue maximization [3].

The major contribution of this chapter is multifold. Firstly, we investigate the different perspectives of ISP and OTT regarding QoE management and why they should collaborate. We focus on quality delivery and technical reasons as well as economic reasons. Secondly, we provide some examples of ongoing collaborations in the industry regarding multimedia service delivery. Thirdly, we propose a QoE-aware collaboration approach between OTT and ISP based on the maximization of the profit by considering the user churn of Most Profitable Customers (MPCs), which are classified in terms of the Customer Lifetime Value (CLV). We define the technical aspects of the collaboration, i.e., the interfaces for the information exchange between OTT and ISP and the possible solutions for a QoE-aware management of network and applications such as the utilization of Content Delivery Networks (CDN), surrogate servers, best path routing, etc. Finally, we provide simulation results aiming at demonstrating the higher profit achieved by proposed collaboration approach with respect to current solutions.

The chapter is organized as follows. In Section 6.2, we highlight the different perspectives of ISP and OTT with regard to QoE management. Section 6.3 presents the economic reasons which drive the collaboration approach. Section 6.4 discusses the current ongoing collaborations between OTTs and ISPs for multimedia service delivery. In Section 6.5 we present our proposed collaboration approach whereas Section 6.6 presents simulation results. Finally, Section 6.7 concludes the chapter.

6.2 Why is the collaboration needed: different perspectives of OTT and ISP

ISP and OTT have different perspectives with regard to QoE management, which are mainly influenced by their roles in the service delivery chain. The ISP is the owner of the network through which its customers are allowed to connect to the Internet as well as to subscribe to OTT services. Nowadays, users are more quality demanding and the ISP has the difficult role in managing the enormous amount of traffic generated by OTT services (especially multimedia services) to provide adequate QoE to the users. Though, the ISP has a network centered view of the quality provided and manages the network on the basis of QoE predictions provided by QoS-to-QoE models in function of monitored KPIs, i.e,

$$QoE^{ISP} = f_1(KPI_1, \cdots, KPI_i, \cdots, KPI_I), \tag{6.1}$$

which means that the QoE predicted by the ISP is a function of I network KPIs. Examples of QoE-based network management approaches are provided in [8, 17]. However, the QoE predicted with Eq. (6.1) does not represent the actual quality perceived by the users. The main reason is that the ISP does not have direct access to users' devices nor application information can be acquired by inspecting network traffic due to the data encryption by the OTTs. Even if the Deep Packet Inspection (DPI) works for some of the OTT services, the human and context influence factors will be still missing and more complex context-aware QoE models for specific applications cannot be integrated into the service delivery [23].

On the other hand, the OTT has access to some types of user-related information (e.g., application usage, preferred application settings, etc.) having its application running on the users' devices. Although the users' needs and expectations are known to OTT, it is limited in providing a good QoE to the users because most of the issues occur in the network and the OTT has no control over the network. The only optimization of the quality which OTT can perform is the adaptation of service parameters in function of the available network conditions in order to maximize the QoE, i.e.,

$$QoE^{OTT} = f_2(KQI_1, \cdots, KQI_j, \cdots, KQI_J, /QoS), \tag{6.2}$$

which means that the QoE predicted by the OTT is a function of J KQIs which in turn depend on the network QoS. Hence, the optimization on the OTT side is done on the basis of application quality indicators only. Examples of QoE-based application management approaches are rate adaption algorithms for HTTP Adaptive Streaming (HAS) [24].

From Eqs. (6.1) and (6.2), it is clear that not only the OTT and ISP have different perspectives in terms of QoE measurements but they also have different roles in the management and optimization of the delivered quality. However, optimal end-to-end QoE-based service delivery can only be possible if OTT and ISP collaborate and agree upon a common QoE-based management approach and exchange of the QoE-oriented information. Since the OTT is more QoE-oriented, its role in the collaboration may be the monitoring of the QoE perceived by the users. By sharing this information with the ISP, this may be able to manage the network in a more QoE fashion than by using QoS-to-QoE models. By following a joint service management paradigm, both the entities may overtake their limitations in QoE management: the ISP can manage its network in a QoE fashion and the OTT can base on the ISP if better network resources are needed for service delivery.

6.3 Economic View

It is a matter of fact that the proper management of QoE brings to direct economic advantages; in fact, if a customer does not receive adequate QoE, it is more likely to become a churner. Indeed, survey results showed that around 82% of customer defections are due to frustration over the service and inability of the operator to deal with this effectively [109]. Moreover, frustrated customers will tell other people about their bad experiences, resulting in negative publicity for the provider. It is also important to note that most of the customers do not complain before defecting but they simply leave once they become unsatisfied. Therefore, it is very important that providers investigate user's expectations and constantly monitor delivered quality to prevent the user churn. From another survey, 20% of respondents reported that they would immediately leave a company because of poor service experience [110]. But most importantly, it resulted that the number of customers who leave because of poor customer experience is significantly higher than the number of those who leave a business because they found a lower price elsewhere: 68% versus 53%. For this specific case, service quality has more influence on the user churn than price.

Therefore, delivering good QoE may help to contrast the user churn and increase the revenue. Usually, the most used parameter for revenue measurement is the Average Revenue Per User (ARPU), which is calculated as

$$ARPU = \frac{TotalRevenue}{TotalNumberOfSubscribers} \tag{6.3}$$

However, this metric does not explicitly represent the impact of a single user on the total revenue, it makes no consideration of the service profitability and does not consider the cost of managing the customer service within the network. Differently, for profit maximization, it is important for the providers to understand the 'value' of each customer in terms of profit, i.e., the revenue (price paid by the customer for the services received) minus the costs (costs for providing the services and managing the customer). Indeed, for the providers, some customers have more value than others and the identification of these customers allows the operator to drive retention actions to the most profitable customers.

Therefore, we consider the Customer Lifetime Value (CLV) as the metric for customer selection because it is defined as "the total value of direct contributions and indirect contributions to overhead and profit of an individual customer during the entire customer life cycle, from the start of the relationship to its projected ending" [111]. CLV is basically computed as

$$\begin{aligned} CLV = (&TotalCustomerRevenue)\times \\ &(NumberOfLoyalYears)\times \\ &(CompanyProfitMargin) \end{aligned} \tag{6.4}$$

Several modified versions of CLV formula are provided in the literature. For example, in [112] the CLV of the customer i, for the horizon h from the period t is defined as

$$CLV_{i,t} = \sum_{k=1}^{h} \sum_{j=1}^{q} \frac{\pi_j x_{i,j,t+k}}{(1+r)^k} \tag{6.5}$$

where r is the discount rate and the net cash flow yielded by the transaction on product j is computed as the product of the product usage $x_{i,j,t}$ and π_j, i.e., the marginal profit by a unit of product usage. This is a marketing equation for a general product but it may be used, for example, for computing the CLV of the customers of call services, for which the profit depends on the service usage. However, this equation cannot be applied in the case of flat rates because of the pricing structure that charges a single fixed fee for a service, regardless of usage. Since flat rates are common in telecommunication and multimedia services, we provide a CLV equation for this pricing strategy as follows

$$CLV_{i,t} = \sum_{k=1}^{h} \sum_{l=1}^{L} \frac{\pi_l s_{i,l,t+k}}{(1+r)^k} \tag{6.6}$$

where l is the level at which a service s is provided and the marginal profit π_l depends on the level l. Indeed, generally, more than one fixed fee can be chosen by the customers, which depends on the level of usage of the service they need. If this level is overtaken, additional fees are applied.

Only the customers which have already been acquired are involved in churn modeling and the CLV helps to identify these customers. One of the major objectives of ISP and OTT is to maximize their profit and to do this we focus on minimizing the churners.

6.4 Current Trends in Multimedia Industry

The Internet service delivery chain involves several entities including ISPs, regional/global transit providers, OTTs, Content Delivery Networks (CDNs) and Internet eXchange Points (IXPs). The role of each entity is different and services are delivered based on mutual agreements [50]. Currently, in the multimedia industry collaboration agreements are QoS-based and the most common form of collaborations between OTTs and ISPs can be classified into two categories: peering and surrogate servers. The rest of this section briefly highlights some of the ongoing practices used in the multimedia industry for the delivery of quality to the end users.

6.4.1 Peering

Currently, most of OTT services use different CDNs for the decentralization of the contents so that these are located near to the users on a regional basis in order to lower the content retrieval latency. A recent form of collaboration between OTT and ISP can be seen as the ISPs peering with the CDNs of the OTTs. On the basis of peering policies, peering agreements between OTT and ISP can be classified as Settlement-Free-Interconnection (SFI) and Paid-peering. The former is the mostly used peering strategy found in the industry while the latter is considered as a violation of Network Neutrality [50]. Technically, the peering can be categorized as 1) Direct or Private Network Interconnection (PNI) and; 2) Public Peering Interconnection (PPI) through the IXPs. To understand the usage of these options, we analyze the cases of YouTube (Google), Skype (Microsoft) and Netflix as some of the popular OTT services.

YouTube video streaming is mostly delivered by Google CDNs where the CDNs are connected to the Google core data centers [52]. Google provides both PNIs and IXPs based peering to ISPs based on SFI peering policy with two different Google Autonomous Systems (ASs), which provide a complete set of Google services (common peering option) and a subset of Google's most popular content (available at a small number of locations) [52]. Similarly, for Skype, Microsoft is providing SFI based PNIs and IXPs peering connections where the peer should declare all the routing paths and Microsoft carries traffic optimization for the Skype generated traffic on the basis of network latency [113]. In the case of Netflix, there are several SFI based open peering PNIs and IXPs connections for the ISPs; however ISP ready for the peering agreement needs to join Netflix Open Connect Program, whose purpose of peering is not only to provide better connection between Netflix's CDN and ISP but also to move the most popular content near to end users inside the host's network with Netflix's surrogate servers called Embedded Open Connect Appliances (OCAs) [53].

6.4.2 Surrogate servers

It is also found to be a normal ongoing practice by the OTTs to provide the hosting ISP with surrogate servers, which can be installed inside the host network to provide the contents near to the end users. The selection of the content is based on the usage trends which are known to the OTT. The surrogate servers provide mutual benefits for the two providers: by decreasing the transit requests, the ISP decreases transit costs as well as the network overload in the backbone network; furthermore, it allows to deliver the OTT contents with less latency than just a simple peering. Most of the surrogate servers based collaborations found in the industry are settlement-free where none of the parties pay each other but they agree on mutual benefits. An important case is the one of Google that is providing Google Global Cache (GGC) edge nodes to ISPs with more Google's applications (YouTube, Google Play, Google Search, Google Maps etc.), which are capable of treating from 60% to 80% of the traffic generated by Google's application [52]. Another is the case of Microsoft that is providing surrogate servers to ISPs for Skype services [113], whereas Netflix is providing the collaborating ISPs with OCAs surrogate servers under Netflix Open Connect program [53].

6.5 Proposed Collaboration Approach

As analyzed in the previous sections, OTTs and ISPs are collaborating on a QoS basis through peering and surrogate servers to deliver better service quality rather than considering the user perceived quality. To address this limit, we propose a QoE-aware collaboration approach

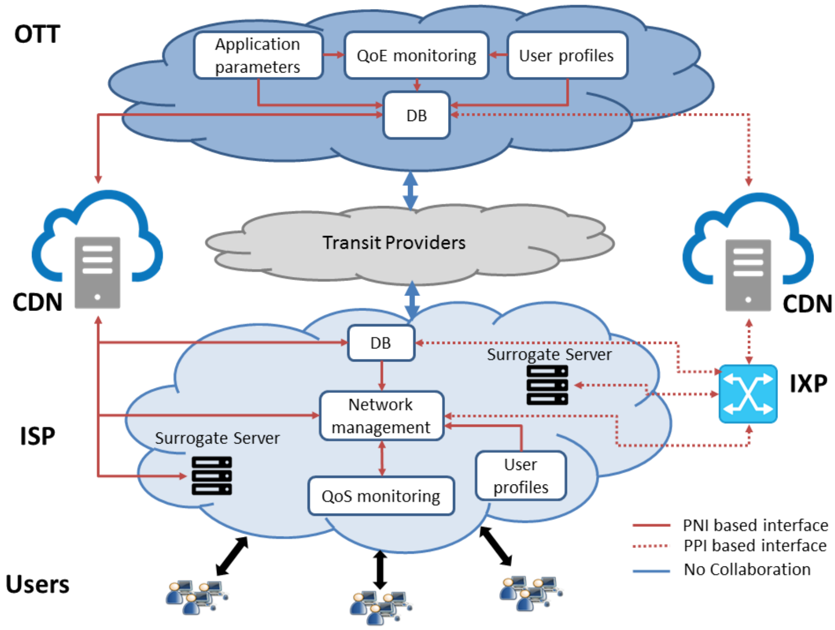

Fig. 6.1 Reference architecture for the proposed OTT-ISP collaboration approach.

where optimization on both network and application sides is done jointly on the basis of economic factors expressed in terms of the Customer Lifetime Value (CLV), as highlighted in Section 6.3. The proposed collaboration approach is inspired by the settlement-free peering policy where none of the partner organizations pay each other. The main reason is that the collaboration with the OTT may reduce the transit costs for the ISP; indeed, if contents are provided by CDNs and surrogate servers, the ISP does not have to pay transit providers. Moreover, the collaboration with the ISP may decrease the user churn for OTTs because thanks to the information provided by the OTT, the ISP can deliver multimedia services with better QoE. As demonstrated in [3], the higher the QoE the lower the user churn and the higher the profit for OTT and ISP. Therefore, the collaboration brings advantages to both the parties.

The reference architecture of the proposed collaboration approach is illustrated in Fig. 6.1. The OTT monitors the QoE of the users as a function of application and network parameters and customer profiles using specific QoE models for the delivered services. A survey on parametric QoE estimation models for popular services is provided in [79]. The QoE is measured per user session each time the user utilizes the application. The measured QoE is stored in the OTT database together with the type of application and the IP address and CLV of the customer. Similarly, the ISP has its own database in which the IP address and the CLV of each customer are recorded. IP and CLV values are updated only if they change. The OTT provides the ISP with APIs which allow the ISP to get all the OTT information from the OTT database. As highlighted in Fig. 6.1, information exchange between OTT and ISP can occur through both PNI based interface (information is directly shared by PNI peering interfaces via OTT's CDN) and PPI-based interface (peering and exchange of information is done through publicly available IXPs, e.g., PeeringDB [1]).

On the basis of the IP addresses, the ISP creates a table composed of the customers in common between the ISP and the OTT. These customers are then sorted in function of their CLV. Specifically, the CLV considered is the sum of the CLV computed by the OTT with the CLV computed by the ISP so that, for a customer i:

$$CLV_i^{TOT} = CLV_i^{OTT} + CLV_i^{ISP} \tag{6.7}$$

The higher the CLV_i^{TOT}, the most profitable is the customer for both the ISP and the OTT. Note that the OTT may just provide normalized CLV data not to disclose the real absolute values to the external world still allowing the collaboration to work properly.

Getting customers' information from the OTT, the ISP has a total overview of common customers, from their network resources to their perceived QoE as well as their economic impact. The objective of the ISP-OTT collaboration is to implement a QoE-aware management of the network to minimize the user churn of customers with the highest CLV_i^{TOT} so to maximize the profit. Then, the first step is to classify customers into two classes depending on their CLV_i^{TOT}. Specifically, a threshold is chosen to distinguish between most profitable customers (MPCs) and standard customers (SCs). The threshold depends on the difference of CLV_i^{TOT} among the customers. However, since resources are limited and costly, the percentage of MPCs should not be too high. For example, a 20% or 30% could be a reasonable percentage. Though, these decisions should be evaluated case by case.

The second step is to identify, within the whole territory controlled by the ISP, the location zones (LZs) where most of the MPCs are situated. Let $n = 1, 2, \ldots, N$ indexes the LZs controlled by the ISP. These LZs are ranked on the basis of the number of MPCs belonging to each of them. Following this ranking, more network and application resources are provided to the highest ranked LZs to optimize QoE of these customers and lose as few MPCs as possible.

Since the focus is on multimedia services, for QoE optimization several actions may be possible, such as i) network management implemented by the ISP for traffic prioritization on the basis of the type of service to be delivered [8]; ii) the places where surrogate servers should be located must be as close as possible to the highest ranked LZs; iii) the contents replicated on the surrogate servers must be the most viewed by MPCs. Indeed, as discussed

[1] https://www.peeringdb.com/

in [25], CDN basically deals with the issues of identifying where to place surrogate servers and identifying what content to replicate on the servers. With the proposed collaboration approach, these issues are solved with a major focus on profit maximization. In fact, if surrogate servers will be situated closer to the highest ranked LZs, customers belonging to these LZs will perceive better QoE due to the improvement of service delivery.

6.6 Simulation and Results

The objective of the simulations is the evaluation of the potentialities of the collaboration approach proposed in Section 6.5 in comparison with the No collaboration in terms of the QoE delivered and the user churn for MPCs and SCs. As a use case, we focus on video streaming services which are currently the most popular and resource consuming services. Therefore, the considered scenario consists in an ISP and a video streaming service (OTT), which collaborate with the aim to increase the profit. The ISP's network covers a certain region divided in different LZs, and the objective is to provide better resources to the LZs where most of MPCs are located.

For the simulation purposes, two cases are considered, i.e., No Collaboration (NC) and Collaboration (Col): with the former approach, services are provided to all users in all LZs over the best effort Internet, while the latter is based on our proposed approach, i.e., the surrogate servers (SRs) are distributed closer to the LZs on the basis of the number of MPCs for both OTT and ISP per LZ.

The simulations are performed using the MATLAB software platform. We consider 5 different LZs in the ISP network through which the users are connected to Internet to use OTT video streaming service. The number of users for each LZ is as follows: 600 for LZ_1, 980 for LZ_2, 922 for LZ_3, 950 for LZ_4 and 896 for LZ_5. The assigned CLV_i^{TOT} to each user is randomly distributed on the normalized scale from 0 to 1: the users with the CLV_i^{TOT} greater than 0.8 are classified as MPCs whereas the others as SCs.

For prediction of the QoE, in terms of the Mean Opinion Score (MOS), the model proposed in [81] is chosen because it is designed for the HTTP video streaming:

$$MOS = \alpha \cdot e^{-\beta(L) \cdot N} + \gamma \tag{6.8}$$

where L and N are stalling duration and number of the stalling events, respectively, while α, γ and $\beta(L)$ are some formula parameters that have been set as follows: $\alpha = 3.5$, $\gamma = 1.5$ and $\beta(L) = 0.15 \cdot L + 0.19$.

Table 6.1 KQIs assumed to be provided with regard to the number of SRs for LZ for Col and NC approaches.

Approach	**Number of SRs**	L **(s)**	N
NC	0	0–4	1–5
Col	0	0–2	1–3
	1	0–1.5	1–2
	2	0–1	1
	3	0-0.5	0

Table 6.1 shows the KQIs, N and L, which are assumed to be provided considering different number of SRs (from 0 to 3) situated closer to a LZ. As discussed in [25], SRs contain replicas of video segments and not the whole videos. Accordingly, the customer's video client can select the SR which provides each video segment with best performance. Therefore, the higher the number of SRs closer to a LZ, the better the performance provided to the video streaming customers (both MPCs and SCs) belonging to that LZ. The values of L and N are chosen on the basis of the influence they have on the perceived QoE [81]. The difference of values between NC and Col approaches for the case of 0 SRs is due to the fact that with the Col approach the latency for the content retrieval is lower because of peering provided by CDN.

Moreover, we consider the user churn function proposed in our past work [3] for the calculation of the user churn:

$$U_i(QoE_i) = \frac{1}{1+e^{-z(QoE_i - QoE_i^m)}} \tag{6.9}$$

where $i = \{MPC, SC\}$ identifies the CLV class, $U_i(QoE_i)$ is the user churn of the user belonging to i-th CLV class, QoE_i is the mean quality delivered over the time period of a month, and QoE_i^m is the quality level at which 50% of the users leave the service. Moreover, z is the sensitivity of human perception to the perceived quality. The QoE_i^m in Eq. (6.9) is set to 3 and 2.5 for the MPCs and SCs users respectively while $z = 3.5$ for both. The user churn is calculated using mean of 100 iterations of the QoE delivered per LZ over a month.

We considered the distribution of 5 SRs to the 5 LZs. Fig. 6.2 shows how the number of SRs per LZ influences the user churn of MPCs and SCs, respectively. It is evident that the higher the number of SRs, the higher the QoE perceived by the customers and the lower is the user churn. However, 2 SRs are more than enough to provide a satisfying level of QoE because customers are not leaving the service. It is important to note that we are not

Table 6.2 Simulation results.

SR distribution	LZ ranks					N. of leaving MPCs (%)	N. of leaving SCs (%)	Average QoE
	$1(LZ_5)$	$2(LZ_3)$	$3(LZ_4)$	$4(LZ_1)$	$5(LZ_2)$			
N. of SRs per LZ (D_1)	3	2	0	0	0	24.73	6.89	**3.86**
N. of SRs per LZ (D_2)	2	2	1	0	0	18.61	4.89	3.77
N. of SRs per LZ (D_3)	2	1	1	1	0	**16.37**	**3.71**	3.76
N. of SRs per LZ (D_4)	1	1	1	1	1	20.97	4.34	3.69
N. of SRs per LZ (NC)	0	0	0	0	0	95.05	73.48	2.59

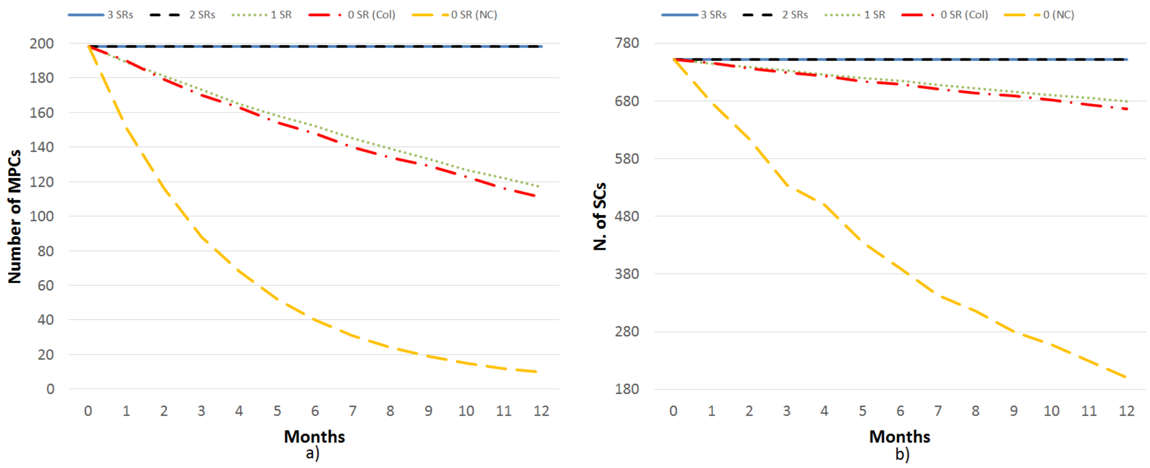

Fig. 6.2 Number of MPCs (a) and SCs (b) leaving the service as a function of the number of SRs closer to their LZ.

considering new customers joining the services but only the number of customers leaving the service with respect to customers present at the time of service's activation.

We considered 4 different distributions for simulations and the results are shown in Table 6.2. The best results are obtained with the distribution D_3, which allows to lose the lower number of MPCs and SCs. However, with respect to the NC approach, all the distributions considered for the Col approach provide better results, both in terms of user churn (reduced from 95% to 16% for MPCs and from 73% to 3.71%, using D_3) and QoE (increased from 2.59 to 3.76, using D_3). Even though the D_3 is not providing the highest average QoE as compared to D_1 but the user churn is lower than the other distribution because the resources are more distributed among LZs in accordance with MPCs.

6.7 Conclusion

In this work, a QoE-aware collaboration approach between OTT and ISP is proposed, based on the maximization of the profit. The objective is to improve the service delivery to the LZs with the highest number of Most Profitable Customers (MPCs) (classified in terms of the Customer Lifetime Value (CLV)) by distributing Surrogate Servers (SRs) closer to these LZs. With simulation results we show that with the proposed collaboration approach the user churn of MPCs and SCs is reduced respectively from 95% to 16% and from 73% to 3.71%. This is due to the better QoE perceived by these customers, which is increased (in terms of MOS) from 2.59 for no collaboration to 3.76 for the collaboration approach.

Chapter 7

Timber: SDN based Collaborative QoE Management Platform

Abstract

In this chapter, we present an open source Software-Defined Networking (SDN) based emulation platform called *Timber*. It is aimed at providing the research community with a tool for experimenting new Quality of Experience (QoE) management procedures and tools in multimedia service delivery. *Timber* is developed on the top of Mininet SDN emulator and Ryu SDN controller, which provides the major functionalities of the traffic engineering abstractions in SDN environment. Moreover, the platform provides an actual complete video streaming application including the implementation of the server side and client side probes for QoE measurements which have functionalities to store the quality measurements into the cloud database accessible to the SDN controller application. In this chapter, we first discuss the general architecture and framework of *Timber*. Secondly, we provide the implementation details and major functionalities of the platform. Thirdly, we provide experimental results to highlight the major functionalities of *Timber* by 4 different scenarios which include traffic shaping through DiffServ and dynamic resource allocation by queuing strategies.

7.1 Introduction

Today the Internet traffic is mainly composed of multimedia traffic due to significant advancement in relevant technologies that are now exploited by the Over-The-Top service providers (OTTs) [114]. Such a huge rise in consumption of the multimedia services over Internet has raised an issue of delivering quality to the end users, i.e., the Quality of Experience (QoE).

Hence, it is a matter of fact that QoE-aware service management approaches are gradually replacing classical QoS-aware service management to consider the subjective perception of the users. Accordingly, many studies proposed QoE-aware management approaches for different services and especially video streaming services [8, 9, 17].

The rise of Software-Defined Networking (SDN) provided new insights into network and service management. Being based on traffic engineering abstractions and virtualization techniques, SDN allows to simplify the implementation of traffic engineering techniques and the communication between network and service providers. SDN provides the ease to maintain a global view of the network to perform the network-wide management operations with the centralized programmable control plane separated from the data plane. SDN architectures are dynamic, manageable, cost-effective, and adaptable, and for these reasons have been widely used in the last years for the QoE-aware management of multimedia services [115, 116].

Currently, different open source SDN based network management tools are available such as Ryu controller[1]/Mininet[2]; however their use for QoE monitoring and management is not straightforward as those tools are not particularly designed for QoE monitoring/management purposes. Moreover, the learning curve of those tools is quite steep. Though some SDN-based QoE management solutions particularly designed for the specific purposes can be found in literature, on the top of which new algorithms cannot be evaluated. Hence, there is a need for an open source and general purpose SDN-based platform, which may assist the research community to evaluate QoE management and monitoring strategies.

In this work, we propose *Timber*, an open source SDN based emulation platform to provide the research community with an emulation tool for experimenting new QoE management procedures and techniques. *Timber* is developed on the top of the Mininet SDN emulator and Ryu SDN controller, which provides the major functionalities of the traffic engineering abstractions in SDN environment. Furthermore, *Timber* provides a complete video streaming application: the server side implementation includes the content hosting whereas the client-side implementation provides probes at the user end for the collection of QoE measurements, which are stored in a cloud database which is accessible to the SDN controller application for network-wide QoE management. The current implementation of *Timber* is equipped with QoE-aware traffic engineering techniques in SDN environment which include traffic shaping, dynamic resource allocation and prioritization of flows by DiffServ implemented through SDN controller applications. Finally, we provide experimental

[1]https://osrg.github.io/ryu/
[2]http://mininet.org/

results to highlight the major functionalities of *Timber* by 4 different scenarios which include traffic shaping through DiffServ and dynamic resource allocation by queuing strategies.

The chapter is structured as follows: Section 7.2 presents some of the most relevant QoE-aware service management studies considering SDN based approaches. Section 7.3 highlights the motivations behind the implementation of *Timber* and the considered use cases. Section 7.4 presents the proposed SDN based architecture including the implementation details, whereas in Section 7.5 preliminary experimental results are presented. Section 7.6 concludes the chapter.

7.2 Background

SDN based network management approaches are widely used for QoE optimization. In [115], an SDN-based function is presented, which is supposed to enable OTT players to monitor QoE and ensure QoE requirements by imposing policies/rules by the SDN controller in specific points of the network, as the means to perform traffic management. However, besides the proposed architecture, no implementation/experiment results is provided. [117] leveraged on SDN to implement an in-network QoE Measurement Framework (IQMF) that provides live network-wide QoE measurements. IQMF offers network measurements and the respective analysis as a service to the network provider or content distributor via an API, which can use it for QoE-aware service optimization. A combined architecture of LTE and SDN is proposed in [118], where an SDN controller was included into an LTE network as the QoS manager. APIs were implemented to allow the SDN controller to interact with the PCRF entity so that dynamic network changes could be made in accordance with LTE policy rules.

Most of the SDN-based architectures are focused on video streaming services. In [119], an SDN-based architecture is adopted for dynamic bandwidth allocation across multiple competing HTTP Adaptive Streaming (HAS) streams. A video QoE optimization application (VQOA) is implemented to collect information from various points in the network, by interacting with the SDN controller, to analyze the network state and provide a more accurate estimate of the end user QoE. Also, the VQOA is in charge of coordinating rate adaptation decisions in HAS clients and bandwidth allocation at the bottleneck link. Similarly, [116] proposed an intelligent streaming architecture, called SDNHAS, which leverages SDN capabilities of assisting HAS players in making better adaptation decisions. SDNHAS formulates the bitrate and quality decisions and then uses a per-cluster decision algorithm (Q-learning based) to successfully avoid sub-optimal decisions and to maximize QoE without any bandwidth violations and affecting other players QoE. To mitigate bottlenecks and improve video QoE, [120] leveraged an SDN platform from OTT video service provider's point of view. The proposed SDN application is designed to monitor network conditions of streaming flow in real time and dynamically change routing paths using multi-protocol label switching (MPLS) traffic engineering to provide reliable video watching experience. By selecting better routing paths to provide higher resolution video during a download, simulations results show that the proposed QoE-aware mechanism allows for a 55.9% improvement in enhancing the viewing experience, especially during busy hours. In [121], a QoE-driven dynamic task allocation scheme is proposed, which makes use of SDN/NFV technologies to find and provide the best combination of network nodes that can cooperate during the execution of the tasks and in the same time to improve the overall QoE level of the end users. Preliminary results based on the Mininet network emulator and the OpenDaylight controller have shown that the proposed approach can significantly improve the QoE of a transmitted video by selecting the best path with normalized QoS values. [122] used an SDN-controller for a video streaming service to perform several tasks within the ISP network: i) communicate with the network devices in order to enable monitoring of the required data; ii) monitor the network; iii) analyze the data to perform the MOS estimation for the different users; iv) define an action plan to improve the quality of the delivered video.

However, all the aforementioned studies are particularly designed for specific objectives. Our objective is to provide the research community a platform for experimenting QoE management techniques in a realistic environment. To the best of author's knowledge, the only study proposing a framework to facilitate rapid experimentation of QoE models in SDN environments is [123], which is still in its early stage.

7.3 Motivations and use cases

Though some SDN-based QoE management solutions can be found in literature which are particularly designed for the specific purposes, as highlighted in Section 8.2, one of the major drawbacks is that new algorithms cannot be evaluated on the top of these QoE management solutions. Hence, there is a need for an open source and general purpose SDN-based platform, which may assist the research community to evaluate QoE management and monitoring strategies.

When defining a QoE-aware management approach, the following functionalities should be considered: 1) *QoS monitoring*– collection of network-related parameters and measurement of the network state; 2) *QoS management*– traffic engineering actions aimed at optimizing network load and traffic flows; 3) *Service management*– optimization of service-related parameters; 4) *QoE monitoring*– collection of quality perceived by the user, user-related information and user's opinions and feedback; 5) *QoE prediction–* mathematical model which predicts user's QoE as a function of user-related parameters, network state and service state and; 6) *QoE controller:* software that, on the basis of the current predicted QoE, decides whether correcting actions should be taken on network and service parameters to optimize QoE.

It is evident that the QoE-aware management of multimedia services is not an easy task, especially due to different information and services in the hands of different providers. Indeed, whereas QoS monitoring and QoS management are in the hands of network providers, service providers may count on QoE monitoring and prediction. But the QoE controller must take decisions on the basis of the combined information owned by both the providers to efficiently optimize end users' QoE [3]. However, joint service management may be impeded by different un-interoperable systems which may discourage the providers towards collaborative approaches.

The proposed SDN based platform, *Timber*, may provide the solution to the need of collaboration among the providers to implement a QoE-aware joint service management. Indeed, thanks to traffic engineering abstractions and virtualization techniques applied by SDN, it is possible to abstract the different system's hardware and to define virtual functions aimed at the monitoring and sharing of information as well as combined controlling decisions. *Timber* has been mainly implemented to test different algorithms and collaboration approaches aimed at QoE-aware management in an open source emulated SDN environment. Hence, the proposed platform is based on the Mininet to create virtual networks that emulate real networks, as well as on the Ryu SDN controller, which allows the application of the traffic engineering solutions in the emulated SDN environment. Also, we implemented further modules to permit joint QoE-aware service management, such as client-side probes for QoE measurements, which have functionalities to store the quality measurements into the cloud database accessible to the SDN controller application and to both the providers. *Timber* is presented in Section 7.4.

7.4 Proposed SDN based platform: Timber

In this Section, we present our proposed SDN based platform called *Timber*. First, we describe the *Timber* architecture and then the implementation details.

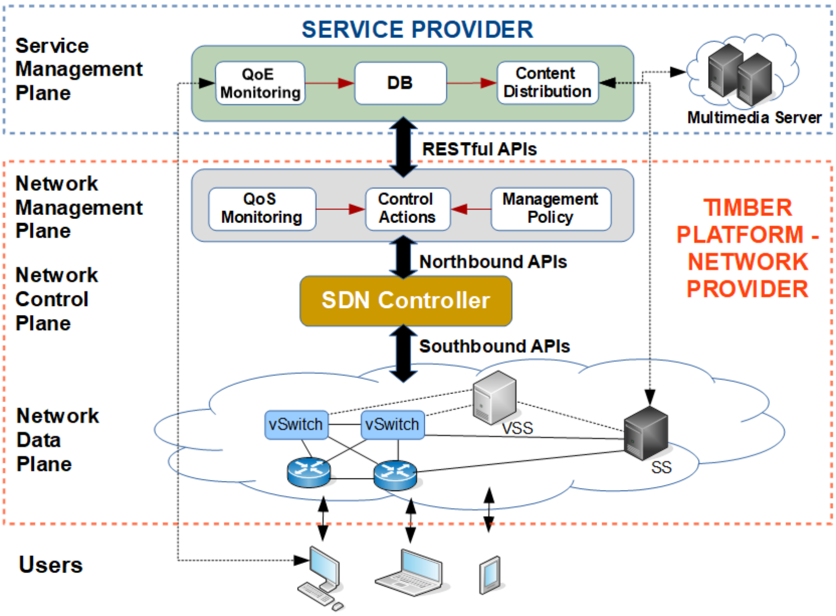

Fig. 7.1 Reference architecture.

7.4.1 Architecture

Fig. 7.1 represents the reference architecture, which is composed of four planes: the Network Data Plane, the Network Control Plane, and the Network Management Plane are the planes concerning *Timber*, whereas the Service Management Plane is a cloud space owned by the service provider.

In the following, we provide the details for each plane.

- *Network Data Plane*: the physical and virtual network, which includes all network devices (switches, vswitches and routers) as well as Surrogate Servers (SSs) and Virtual Surrogate Servers (VSSs) for content replication applied by Content Delivery Networks (CDN), which can be owned by both the network and the service providers.

- *Network Control Plane*: the SDN controller through which it is possible to manage the network devices.

- *Network Management Plane*: the application layer of the SDN architecture controlled by the network provider. It includes: a QoS Monitoring module to monitor the performance of the network; a Management Policy module to take into account Service Level Agreements (SLA); a Control Actions module that decides, on the basis of specific algorithms and objectives, which network control actions should be implemented by the SDN controller to optimize the network resources and improve the quality.

- *Service Management Plane*: a cloud space owned by the service provider. It includes: a QoE Monitoring module to estimate the user's QoE on the basis of service parameters acquired at the client side; a DB where QoE measurements are stored and may be shared with the network provider; a Content Distribution service to deliver multimedia contents with a CDN.

In the proposed approach, the role of the service provider is to perform the QoE monitoring of the delivered service by acquiring service information from the client side based on passive measurements of service-related parameters. On the basis of these measurements, specific QoE models may be used to predict the user experience. Moreover, having its application installed in the user's device, the service provider may predict the user's perceived quality also as a function of human and context parameters which are not available to the network provider. The QoE measurements of active clients sessions are then stored in the DB, which can be also accessed by the network provider through RESTful APIs.

The network provider makes use of SDN technologies to dynamically manage the network in collaboration with QoE measurements performed by the service provider. The SDN application is responsible for QoS monitoring and management to implement the network-wide policies by communicating with the controller through the Northbound RESTful APIs. Furthermore, the SDN controller performs network-wide operations such as link-aggregation, the addition of new flows, traffic reroute and placement of the VSSs. For example, in case of network congestion events, the SDN controller may perform traffic engineering techniques on the congested part of the network. The SDN controller may also make use of Virtual Network Functions (VNFs) to place the VSSs on the network edge near to locations where cache miss handling/flash crowd occurrences are reported by the service provider. The use of Network Function Virtualization (NFV) allows for dynamic resource allocation in terms of content retrieval by enhancing the total capacity of surrogate servers to serve more clients in case of a flash crowd. Moreover, the content retrieval time is also reduced because the client is served by the VSSs instead of the peered CDNs outside the network. Hence, cache miss handling can be easily avoided.

7.4.2 Implementation details

Timber is based on several open source software packages which are listed in Table 7.1. The implementation is broadly categorized into three categories: 1) Server side implementation; 2) Client-side implementation and; 3) Virtual network and SDN controller. The rest of the section provides the implementation details of the aforementioned categories.

Table 7.1 Software packages used in the current implementation of Timber.

Software Package	**Purpose**
Mininet	Creation of virtual networks
Ryu Controller	SDN controller for network management
Apache HTTP Server	Server side implementation of content hosting server
MongoDB 3.4	Implementation of cloud database for exchange of information between network and service providers
Video.js	Server and client side implementation of the video streaming application

Server side implementation

The server-side video streaming service is implemented by the combination of the Apache HTTP server[3] and Video.js[4], which implements the HTML5 video player. The videos encoded in different resolutions and the web pages to play the hosted video contents are hosted on the Apache server. MySQL server and PHP libraries for apache2 server are also installed as dependencies for the web/content hosting service.

Client-side implementation

The client-side implementation includes the client-side video player and the monitoring probe. The client-side video player is implemented by the combination of the Video.js HTML5 player, Javascript library, Jquery, and Ruby script. The major functionality of the back-end side is the implementation of the user end probe which is inspired by the works in [26, 124]. The back-end implementation is mainly responsible for: 1) Loading the video from the server to the front-end player; 2) Front-end player state observation; 3) Computation of QoE application level Key Quality Indicators (KQI) through the probe and; 4) Storing of the results of the probe in the MongoDB[5] database (shared between network and service providers). The computation and storage of video player parameters in the MongoDB database are performed by the Ruby script. Video player parameters include *initial buffering time*, *stalling duration* and *number of stalling events*. These parameters have been considered due to higher influence on the quality perceived by the user as investigated by the studies [125]. The initial

[3]https://httpd.apache.org
[4]http://videojs.com
[5]https://www.mongodb.com/

buffering time is computed as the time difference between the video loading time and the starting time of the video playback. The number of stalling events is computed throughout the video playback by counting the occurrence of the buffering events while the stalling duration is calculated as the time difference between the occurrence of a buffering event and the video playback event after the buffering. The computed video parameters are sent to the MongoDB where they are stored in the database.

Virtual network and SDN Controller

Timber is built on the top of Mininet emulator, which is used to create the virtual network environments. The virtual host of the Mininet acts as a client by running the client-side application in the Google Chrome browser. The Mininet Open vSwitches (OvS) are connected to the Ryu controller through the Southbound interface with the OpenFlow 1.3 protocol in the bridged configuration. The Ryu controller is used as the SDN controller to perform the network-wide monitoring and management tasks through the OpenFlow 1.3 protocol, which provides the following functionalities: 1) Configuration of OvSs through *ovs-vsctl* while keeping the global view of the network and; 2) Monitoring and administration of flow statistics and flow tables through *ovs-ofctl* respectively. The SDN controller application runs on the top of the Ryu controller which communicates through Northbound RESTful APIs. In the current implementation, the controlling operations are divided into two categories based on traffic engineering concepts: 1) Dynamic resource allocation for flows by setting the queues and; 2) Flow prioritization by setting DSCP prioritization classes and remarking of the flows (DiffServ). The SDN application is connected to the cloud database (implemented by MongoDB 3.4) through the RESTful APIs which enable the SDN application to acquire the QoE oriented information (stored by the client side probe) from the cloud database. The SDN application retrieves the information from the cloud database through the secured HTTP GET response after the authentication. The retrieved information follows the JSON format which is later sorted by the SDN controller application. The SDN controller application performs the network level operations on the basis of retrieved information. Fig. 7.2 shows the implementation details of *Timber* as well as the Mininet topology used for the experiments conducted in Section 7.5.

7.5 Experiments Setup and Scenarios

The objective of the experiments is to highlight some of the major functionalities of *Timber* for the QoE management of multimedia services in the SDN paradigm. Note that *Timber*

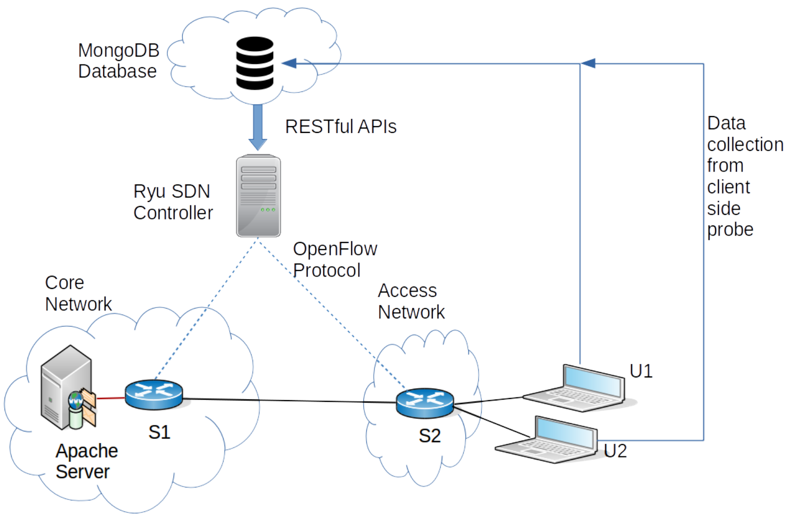

Fig. 7.2 *Timber* implementation details and Mininet topology used for the experiments.

platform is not limited to only these experiments and the focus of experiments is better QoE provisioning to video streaming.

The experiments are conducted on the *Timber* platform built in the Linux Ubuntu server 14.04, 64 bits environment installed on the Virtualbox with 3 GB RAM, 8GB Hard drive and core 2 Duo @2.4 GHz CPU resources. The Big Buck Bunny 60-second long video sequence encoded with H.264 codec in 1280×720 resolution with 30 fps is used for all the experiments. The web page for video streaming contains only simple HTML5 player without any other content. The web page and video content are hosted on the Apache server. The Apache server is run as a service in one of the Mininet hosts. All the experiments follow the same topology illustrated in Fig. 7.2. For the experiment purposes, we fix the maximum link capacity of $S1-S2$ to 4 Mbps, as the 720p HD video streaming requires $1.5-4$ Mbps throughput[6]. Hence, the $S1-S2$ link can provide a good video streaming service only if one HD 720p video streaming session traffic is going through it. In order to create a network bottleneck at $S1-S2$ link, HD video streaming sessions are started in the $U1$ terminal while TCP protocol based web traffic of 4 Mbps is generated through the $U2$ terminal in all experiments settings. The video streaming session is started in the $U1$ terminal by visiting the link of a video streaming web page hosted by Apache server. We considered four different scenarios to highlight major functionalities of *Timber* to improve quality in case of network congestion:

[6]https://support.google.com/youtube/answer/2853702?hl=en

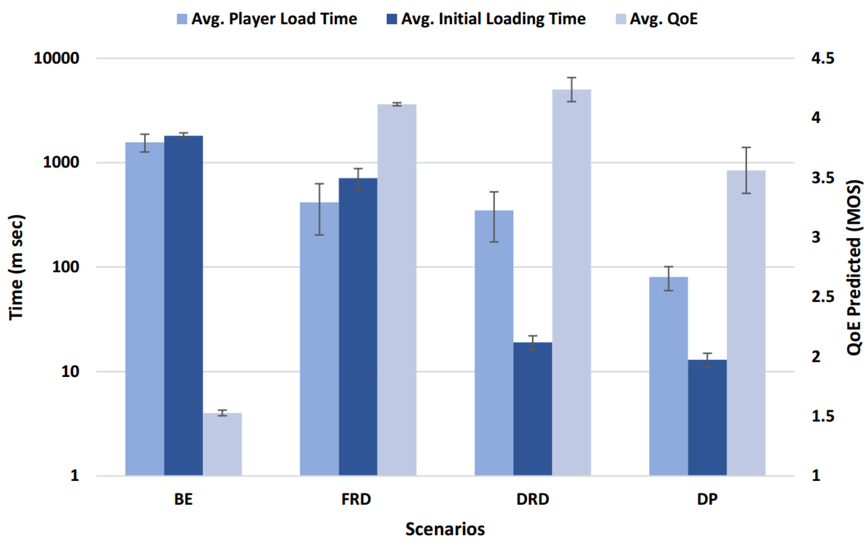

Fig. 7.3 Player loading, initial loading time and predicted QoE for each scenario.

- *Best Effort (BE)*– In this scenario no optimization/traffic engineering concepts are applied, i.e., the network delivers the traffic on the best effort.
- *Differentiated Resource Distribution (DRD)*– This scenario provides different network level resources to video streaming and normal web traffic. The queues are set up in $S1$ and $S2$ OvS switches by the controller application to avoid QoE degradation by congestion. The queues have the throughput variance $2.5-3$ Mbps and $1-1.5$ Mbps for video streaming and web traffic respectively. The SDN controller application assigns the traffic flows generated by the video streaming and web traffic to the specific queues through flow matching. These queues provide the minimum guaranteed throughput of 2.5 Mbps and 1 Mbps to video streaming and web traffic respectively.
- *Fair Resource Distribution (FRD)*– This scenario is similar to DRD scenario in terms of traffic engineering concepts; however the throughput is divided among the video streaming and web traffic by setting two different queues of the same throughput, i.e., throughput of each queue is 2 Mbps.
- *Diffserv Prioritization (DP)*– In this scenario, the MPLS network is created between $S1$ and $S2$. The video streaming flow is prioritized as DSCP class $AF11$ at $S1$ while marking is done at $S2$ by setting the flow matching criteria through controller application. The web traffic is served as best effort in this scenario.

In all scenarios, the video sequence is played 5 times (5 runs). The Fig. 7.3 represents the average page loading time and initial loading time of the video in milliseconds (ms) on a logarithmic scale in each scenario on the primary vertical axis. The *BE* scenario shows the highest average page loading time 1568 ms while *DP* has the minimum average page loading time 80.33 ms. The *BE* scenario has the highest average initial loading time 1809 ms among all scenarios while the *DP* has the minimum 13 ms. It is worth noticing that significant decrease in page loading time and initial loading time is observed after by QoE-aware traffic engineering techniques performed through *Timber* controller application as compared to BE case.

Fig. 7.3 shows the average predicted QoE for all scenarios on Mean Opinion Score (MOS) scale (1-5) on the secondary vertical axis. For the computation of the QoE, we considered the model (which is validated by subjective assessment) presented in [125]. Note that the selection of a particular QoE model is not the prime focus of the work and it can be replaced with the more complex models. The *DRD* scenario has the highest average predicted QoE 4.23 in all scenarios while *BE* has the lowest score 1.52. Hence, significant improvements in the predicted QoE is observed even in the case of network congestion by QoE management through *Timber*. Fig. 7.4 represents the throughput utilization behavior of video streaming and web traffic in one of the 5 runs observed by traffic engineering application of Timber platform in each scenario. Note that Fig. 7.4 does not represent the average throughput utilization of the scenarios in 5 runs. This is because to avoid the curve smoothness due to aggregation. The throughput fluctuations in *BE* scenario represent that client player continuously fetched the video segments from the server due to limited throughput in order to fill the buffer while in *DRD* scenario most of the video segments are downloaded till 38 s because video traffic is provided dedicated queue. In case of *FRD*, the video traffic is shaped to 2 Mbps which is not enough for 720p video so continuous download of video can be observed while in case of *DP*, the effect of priority on video and web traffic can be observed. The Table 7.2 provides the comparison summary for all the scenarios in terms of the measured KQIs and average predicted QoE.

7.6 Conclusion

In this work, we provide an SDN based emulation platform called *Timber* for QoE management experimental research purposes. *Timber* is developed on the top of Mininet SDN emulator using Ryu controller. *Timber* has the capability to not only monitor the KQIs of video streaming session through client side (user end probe) but also to perform QoE-aware traffic engineering tasks through SDN controller application. This work provides the architectural and implementation details of *Timber*, which can be used by the research community to evaluate future QoE management and monitoring strategies. Moreover, we

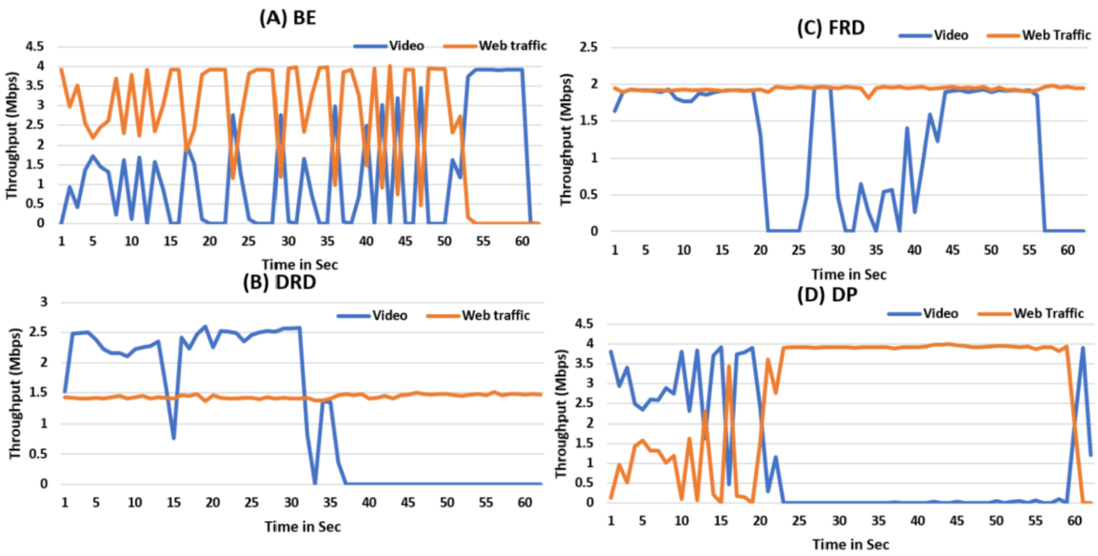

Fig. 7.4 Throughput utilization behavior for each scenario in one run.

Table 7.2 Comparison summary for all scenarios in terms of KQIs and predicted average QoE.

Scenario	**Avg. Player Loading Time (ms)**	**Avg. Initial Loading Time (ms)**	**Avg. Stalling Duration (ms)**	**Max. No. of Stalling Events**	**Min. No. of Stalling Events**	**Avg. QoE (MOS)**
BE	**1568**	**1809**	**4758.08**	**7**	**4**	**1.528**
FRD	416.33	711.33	677.33	1	1	4.114
DRD	350	19	**372**	1	1	**4.237**
DP	**80.33**	**13**	2264.66	1	1	3.560

provide experimental results to highlight the major functionalities of *Timber* by 4 different scenarios which include traffic shaping through DiffServ and dynamic resource allocation by queuing strategies. The future work will be focused on adding more multimedia services in the platform.

Chapter 8

Comparative Analysis of the Collaborative QoE Management Strategies

Abstract

It is a matter of fact that the Quality of Experience (QoE) has become one of the key factors that determine whether a new multimedia service will be successfully accepted by the final users. Accordingly, several QoE models have been developed with the aim of capturing the perception of the user by considering as many influencing factors as possible. However, when it comes to adopting these models in the management of the services and networks, it frequently happens that no single provider has an access to all the tools to either measure all the influencing factors related parameters or control over the delivered quality. In particular, it often happens to the Over The Top (OTT) and Internet Service Provider (ISP), which act with complementary roles in the service delivery over the Internet. On the basis of this consideration, in this chapter we first highlight the importance of a possible OTT-ISP collaboration for a joint service management in terms of technical and economic aspects. Then, we propose a general reference architecture for a possible collaboration and information exchange among them. Finally, we define three different approaches, namely: joint-venture, customer lifetime value-based, and QoE-fairness-based. The first aims to maximize the revenue by providing better QoE to customers paying more. The second aims to maximize the profit by providing better QoE to Most Profitable Customers (MPCs). The third aims to maximize QoE fairness among all customers. Finally, we conduct simulations

to compare the three approaches in terms of QoE provided to the users, profit generated for the providers and QoE fairness.

8.1 Introduction

The current continuous increase of Internet traffic is mostly due to the developments in the multimedia industry conducted by Over The Top (OTT) service providers (e.g., YouTube, Netflix, Skype, Facebook) as well as to the widespread use of powerful multimedia mobile handheld devices, such as smartphones and tablets. Indeed, with the recent findings, the Internet traffic is predicted to be made for 78% by video traffic by 2021 [126]. Such a drastic increase of multimedia traffic has required Internet Service Providers (ISPs) to upgrade their network infrastructures and invest in the deployment of new technologies (e.g., optical fibers, 4G/5G mobile networks) to provide the means to handle such a huge amount of traffic.

However, network oversizing is not the right way to deal with the problem of traffic delivery because various are the types of services delivered over the Internet with different bandwidth requirements: from messaging services, which require few tenths of bytes per second, to virtual reality streaming that may require even hundreds of Mbps. Also, it would be difficult for the ISP to transform the investment in upgraded networks to revenue because the ISP is on a disadvantaged economic position with respect to OTTs, for four main reasons: i) ISPs are not in the loop of revenue generation between OTTs and users; ii) OTTs are in the market of traditional business of ISPs, e.g., voice and messaging services are also provided by OTTs such as WhatsApp and Skype; iii) although on average ISP's profit per customer is greater than OTT's profit per customer, OTTs are able to reach a much larger amount of potential customers over the globe; iv) ISPs are the most affected entities by the user churn because most of the time the user perceived quality is degraded due to bottlenecks in the ISPs' networks and the users decide to switch their Internet connection to another ISP. Consequently, this leads to decrease in market share and reputation as well as lower revenue for an ISP.

For these reasons, ISPs have to rethink current technical and business approaches, which are linked by the Quality of Experience (QoE), i.e., the quality as perceived by the user [127]. Indeed, if users perceive a high QoE, they tend not to leave the service and to continue to contribute to the ISP's profit; but the provision of high QoE for each type of service requires the ISP to adopt novel QoE-aware traffic management approaches that must be aware of the type of traffic to be managed [8, 9, 128–130]. Additionally, OTTs need to provide high QoE to their customers to increase their business: this has allowed for the growth of third entities (e.g., Akamai, Limelight, L3) that provide services, such as Content

Delivery Networks (CDNs), Web Acceleration, Caching, which help to provide multimedia services with better quality. However, emerging multimedia services include high quality videos, cloud applications and highly interactive applications, which require increasingly network resources. For these reasons, OTTs and CDNs are trying to place their platform for service provision inside the ISPs' networks to be as close as possible to the end users.

But acting independently it is difficult for ISPs and OTTs to deliver adequate QoE to their customers, because QoE is a function of some parameters on the hands of the ISPs (QoS) as well as some other parameters on the hands of OTTs (application parameters, human and context factors). Therefore, a collaboration among the OTTs and ISPs seems to be the natural solution which can increase their profits with better provision of quality to their customers. However, to date research in this field is mainly divided into two separate paths: on the one hand OTTs make use of network-aware application management approaches, which aim to adapt application parameters on the basis of the monitored network status [24]; on the other hand, ISPs make use of application-aware network management approaches, which aim to manage network resources on the basis of the services to be delivered through the network [8, 17]. Some approaches towards collaboration between networks and applications for improving QoE are discussed in [131].

In this chapter, we define a general architecture to allow for collaboration among ISPs and OTTs. We then propose three different collaboration approaches among OTTs and ISPs and we compare them in terms of QoE provided to the users, profit generated for the providers and QoE fairness. Indeed, one of the ultimate goals in future multimedia networks is to provide a fair user-centric quality, so that the user QoE is properly handled for all users in a network. The reason is that it may happen that selfish applications maximize its own QoE, potentially worsening QoE levels of users of different applications. According to the general fairness metric defined in [132], which satisfies QoE-relevant properties, a system is defined absolutely QoE fair when all users receive the same QoE value.

The main contributions of this work and its importance in comparison with related works are:

- We discuss the technical and economic reasons which push ISPs and OTTs to collaborate as well as their different views on QoE management.
- We propose a general reference architecture for collaboration and information exchange between ISPs and OTTs. The peering agreements and CDNs are considered for content delivery whereas the concept of InterCloud has been considered to allow the providers to exchange information with common infrastructures (cloud) and APIs.

- We propose three different collaboration approaches, namely joint-venture, CLV-based (Customer Lifetime Value - based) and QoE-fairness-based. The first, proposed in [3], aims to maximize the revenue by providing better QoE to customers paying more. The second aims to maximize the profit by providing better QoE to Most Profitable Customers (MPCs). The third aims to maximize QoE and QoE fairness among all customers. We provide the pseudo code algorithm for each of these collaboration approaches.
- We conduct simulations to compare the three proposed approaches in terms of QoE provided to the users, profit generated for the providers and QoE fairness.

The chapter is organized as follows. In Section 8.2, we present the background regarding current collaboration agreements for service delivery and economic aspects of QoE management as well as the practical obstacles to achieving the collaboration. Section 8.3 presents the proposed reference architecture for collaboration and information exchange among ISP and OTTs. Section 8.4 presents the three proposed collaboration approaches based on revenue/profit maximization (joint-venture and CLV-based approach) and QoE fairness maximization (QoE-fairness-based approach). In Section 8.5 the results of the conducted simulations are discussed whereas Section 8.6 concludes the chapter.

8.2 Background

In this Section, we present the background regarding current collaboration agreements for service delivery and economic aspects of QoE management which are used in the following treatment. Finally, we discuss the practical obstacles to achieving the collaboration.

8.2.1 Collaboration agreements for service delivery

Although a few collaboration approaches between network and application providers have been proposed in the literature for specific services such as ALTO (Application-Layer Traffic Optimization) [93] and CINA (Collaboration Interface between Network and Application) [94] for P2P networks, and SAND (Server And Network Assisted DASH) [133] for Dynamic Adaptive Streaming over HTTP (DASH). General collaboration approaches that can be used to jointly manage different types of multimedia services are still missing in the literature.

In the typical Internet service delivery chain, different key players are involved including ISPs, OTTs, CDNs and Internet eXchange Points (IXPs). The role of each player is different and the keys for better service delivery are identified as interconnection (or peering) agreements and CDNs. Peering agreements regard the interconnection between two separate networks, which allow exchanging traffic between the users of each network, whereas CDNs replicate contents on replicas of the original servers, referred to as *surrogate servers* (SSs), to lower content retrieval latency [25]. On the basis of peering policies, peering agreements can be classified as Settlement-Free-Interconnection (SFI) and Paid-peering. The former is the mostly used peering strategy and assumes that the traffic is balanced between the parties, while the latter has a cost that is typically lower than transit cost but may be considered as a violation of Network Neutrality as it may be interpreted as prioritized service. Technically, the peering can be categorized as Direct or Private Network Interconnection (PNI) and Public Peering Interconnection (PPI) through the IXPs. Peering agreements established by some of the most popular OTT services, i.e., Google, Skype (Microsoft) and Netflix, are discussed in the following. YouTube video streaming is mostly delivered by Google[1] CDNs, which are connected to the Google core data centers. Google provides both PNIs and IXPs based peering to ISPs based on SFI peering policy with two different Google Autonomous Systems (ASs), which provide a complete set of Google services (common peering option) and a subset of Google's most popular content (available at a small number of locations). Similarly, for Skype, Microsoft[2] is providing SFI based PNIs and IXPs peering connections: the peer should declare all the routing paths and Microsoft carries traffic optimization for the traffic generated by Skype on the basis of network latency. In the case of Netflix, there are several SFI based open peering PNIs and IXPs connections for the ISPs; however, the ISP ready for the peering agreement needs to join the Netflix Open Connect Program, whose purpose is not only to provide better connection between Netflix's CDN and ISP but also to move the most popular contents closer to the end users inside the host's network with Netflix's SSs called Embedded Open Connect Appliances (OCAs)[3]. In Table 8.1, some of the most relevant OTTs that have direct peering with various ISPs in the U.S. are summarized [134].

Currently, most of OTT services use different CDNs for the decentralization of the contents to lower the content retrieval latency. Indeed, CDNs replicate contents on different sites and possibly on different independent ISP networks so that Web requests can be re-directed to one of the SSs according to some selection rule, which are typically based on the usage trends which are known to the OTT [25]. The SSs provide mutual benefits for the two providers: by decreasing the transit requests, the ISP decreases transit costs as well as the network overload in the backbone network; furthermore, it allows the OTT to deliver the contents with less latency than just a simple peering. Most of the collaborations based on SSs are SFI. An important case is the one of Google that is providing Google Global Cache

[1] https://peering.google.com/

[2] https://www.microsoft.com/Peering

[3] https://openconnect.netflix.com

Table 8.1 Some of the most relevant OTTs that have direct peering with various ISPs in the U.S.

OTT	AT&T	Comcast	Verizon	CenturyLink	Sprint
Google	X	X	X	X	X
Amazon	X	X		X	
Facebook	X	X	X		X
Microsoft	X	X	X	X	
Netflix		X	X	X	
Apple	X		X		

(GGC) edge nodes to ISPs with some Google's applications (e.g., YouTube, Google Search, Google Maps), which are capable of treating from 60% to 80% of the traffic generated by Google's applications. Another is the case of Microsoft that is providing SSs to ISPs for Skype services, whereas Netflix is providing the collaborating ISPs with OCAs SSs.

8.2.2 Economic aspects of QoE management

It is a matter of fact that the proper management of QoE brings to direct economic advantages; indeed, if a customer does not receive adequate QoE, it is more likely to become a churner. A survey conducted by Nokia showed that around 82% of customer defections are due to frustration over the service and inability of the operator to deal with this effectively [109]. Moreover, frustrated customers can influence other customers, resulting in negative publicity for the provider. It is worth mentioning that most of the customers do not complain before defecting but they simply leave once they become unsatisfied. Therefore, it stands critical for providers investigate user's expectations and constantly monitor delivered quality to prevent the user churn. From another survey conducted by Accenture, 20% of respondents reported that they would immediately leave a company because of poor service experience [110]. But most importantly, it resulted that the number of customers who leave because of poor customer experience is significantly higher than the number of those who leave a business because they found a lower price elsewhere: 68% versus 53%. Also, a survey of 2,000 Apple iPhone users in the UK and the U.S. has revealed that the 39% of respondents would pay more for a better mobile video-streaming service [135]. Indeed, over half (59%) of subscribers in both countries will abandon streaming a mobile video if they have to wait longer than 15 s, whereas nearly a fifth (19%) will abandon a video after only a five-second wait. Therefore, service quality has a great importance in telecommunications and multimedia services and should be strongly considered to contrast the user churn and increase the revenue.

Usually, the most used parameter for revenue measurement is the Average Revenue Per User (ARPU), which is calculated as the ratio of the total revenue to the total number of subscribers. However, this metric has several issues: i) it does not explicitly represent the impact of a single user on the total revenue; ii) it makes no consideration of the service profitability and does not consider the cost of managing the customer service within the network; iii) it was defined as a voice-based index while with the advent of smartphones and tablets mobile data services are becoming prevalent and new metrics are needed to consider data and voice services separately. Indeed, it is important for the providers to understand the *value* of each customer in terms of profit, i.e., the revenue (price paid by the customer for the services received) minus the costs (costs for providing the services and managing the customer). Indeed, for the providers, some customers have more value than others and the identification of these customers allows the operator to drive retention actions to the most profitable customers (MPCs). We then consider the Customer Lifetime Value (CLV) as the metric for customer selection because it is defined as *"the total value of direct contributions and indirect contributions to overhead and profit of an individual customer during the entire customer life cycle, from the start of the relationship to its projected ending"* [111].

The CLV can be computed as a function of several parameters such as the total customer revenue, the company profit margin and the number of subscriber's loyal years. However, several modified versions of CLV formula are provided in the literature. For example, in [112] the CLV is a function of all the transactions a customer i will make for q products the company is selling. Then, for the horizon h from the period t, the CLV is defined as (note that for presentation convenience h and q are not omitted in the left part of the equation)

$$CLV_{i,t} = \sum_{k=1}^{h} \sum_{j=1}^{q} \frac{CF_{i,j,t+k}}{(1+r)^k} = \sum_{k=1}^{h} \sum_{j=1}^{q} \frac{\pi_j x_{i,j,t+k}}{(1+r)^k}, \tag{8.1}$$

where $CF_{i,j,t}$ is the net cash flow generated by a product j sold to a customer i during period t as a function of the product usage $x_{i,j,t}$. π_j and r are respectively the marginal profit by a unit of product usage and the discount rate. This is a marketing equation for a general product but it may be used, for example, for computing the CLV of the customers of call services, for which the profit depends on the service usage. However, this equation cannot be applied in the case of flat rates because of the pricing structure that charges a single fixed fee for a service, regardless of usage. Since flat rates are common in telecommunication and multimedia services, we provide a CLV equation for this pricing strategy as follows

$$CLV_{i,t} = \sum_{k=1}^{h} \sum_{l=1}^{L} \frac{\pi_l y^s_{i,l,t+k}}{(1+r)^k}, \tag{8.2}$$

where l is the level at which a service s is provided and the marginal profit π_l depends on the level l. $y^s_{i,l,t} = 1$ if the customer i uses the service s at the level l. Otherwise $y^s_{i,l,t} = 0$. Indeed, generally, more than one fixed fee can be chosen by the customers, which depends on the level of usage of the service they need. If this level is overtaken, additional fees are applied.

One of the major objectives of ISP and OTT is to maximize their profit and to do this we focus on minimizing the churners. Specifically, for the CLV-based collaboration approach, the aim is to avoid that the MPCs become churners, because the cost of winning a new customer is much higher than retaining an existing one [136].

8.2.3 Practical obstacles to achieving the collaboration

It is important to highlight some practical obstacles to achieving the collaboration between OTTs and ISPs, such as Network Neutrality (NN) and privacy concerns. The respect of the NN should be considered while implementing solutions for network resource allocation among different applications to preserve the openness of the Internet. Indeed, based on the NN principle, the end users should have equal access to all the content on the Internet and the ISP should be prohibited from discriminating/blocking the content from any of the application providers [90]. Therefore, the network should deliver traffic in the best effort manner; however, lower levels of neutrality violation may be accepted as intrinsic prioritization, load management and blocking of illegal content. Accordingly, network management approaches which result in treating differently the flows to/from collaborating OTTs should be carefully treated in this respect.

With regard to the privacy concerns related to disclosing information, the users should be involved and informed by the providers to request their consensus in the sharing of some specific information related to her activity and preferences, even if that would mean to receive higher service quality. As an incentive, the users may be rewarded with discounts or service promotions. For example, in [27], a survey is conducted to investigate users' preferences and privacy concerns to allow the provider to install monitoring probes in the user terminals. It resulted that the 28% of the users do not want to install any probe under any condition, whereas 72% agreed on the probe installation in exchange for better quality (25%), discounts on the services (10%), extra bandwidth/data usage (11%) and a combination of these 3 benefits (26%).

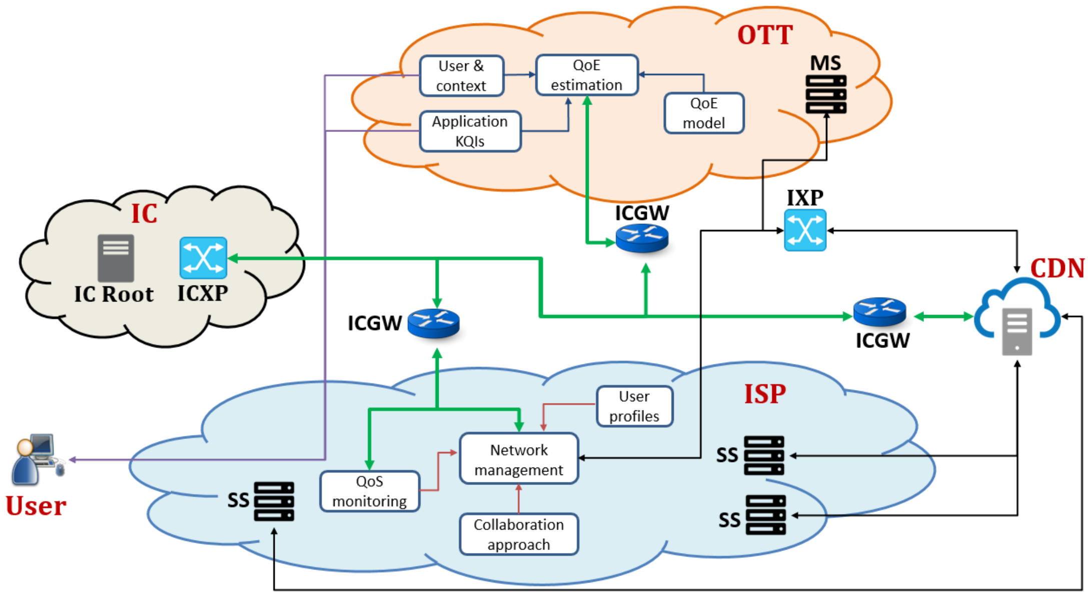

Fig. 8.1 Reference architecture for collaboration among ISP and OTTs.

8.3 Reference Architecture for Collaboration

In this Section, we present our proposed reference architecture for collaboration among ISP and OTTs, which is illustrated in Figure 8.1. This figure highlights the modules, information and network elements involved among ISP and OTTs to drive the joint QoE management of network and applications. Although just one OTT is shown in Figure 8.1, the collaboration approach can consider multiple OTTs involved in the collaboration approach with the ISP that provides the transport services to the end-users. As discussed in Section 1.2, OTT and ISP have different perspectives with regard to QoE management, and with the proposed collaboration we identify the role of the OTT as the QoE monitoring whereas the ISP monitors the QoS and manages network resources.

To accomplish the OTT's main task of QoE monitoring, specific QoE measurements techniques for the delivered services are needed, which would allow for a real-time prediction of delivered QoE, as a function of network KPIs and application KQIs, typically based on parametric models, deep learning and No-Reference methods [128–130, 79]. However, to provide a more reliable estimation of the QoE, such QoE-measurement solutions may need accurate information about the network state, that is not easily accessible by the OTT. The measured QoE is then used by the OTT to drive application level quality optimizations at the media server side and/or distribution of the contents with CDNs to lower the distance between the contents and the end users. However, in case of no collaboration, the SSs are typically placed at the edge, but outside, of the ISP network; instead, with the proposed collaboration, the ISP places the SSs inside its local network so that the content retrieval time is minimized.

The role of the ISP is mainly QoS monitoring and optimization of network resource allocation. Additionally, with the proposed collaboration the ISP provides settlement-free peering interconnection and also hosts the SSs of the collaborating OTTs in its network. With such a collaboration, the ISP reduces the transit costs; indeed, if contents are provided by direct peering and SSs, the ISP does not have to pay transit providers. Moreover, the provision of better service's quality to the end users may decrease the user churn for both the providers. The Network management module of the ISP is the core module that decides how the network resources should be managed as a function of the considered collaboration approach, which could be one of those discussed in Section 8.4. These collaboration approaches require some information to be shared between the ISP and the OTTs, so that the QoE can be accurately predicted and monitored to drive the network management. Different types and amounts of information may be shared, depending on the service provided as well as on the collaboration agreement: however, even the sharing of a minimal, but critical, amount of information may aid the ISP in making QoE-aware network management decisions, as opposed to the case of no collaboration.

Contents and data are exchanged at the IXPs, which are neutral physical infrastructures where the ASs (ISP, OTTs and CDN in this case) can exchange (mostly for free) Internet traffic with the others ASs with which they have a peering agreement. On the other hand, to allow the sharing of information and services among the ASs, we aim to base on the concept of the InterCloud (IC), motivated by the current *Softwarization* trend that has brought most of the biggest OTTs (e.g., Microsoft, Amazon, Google, Facebook) to implement their own cloud systems to provide their services all over the world. Even the future of ISPs is on the software with the shift towards network virtualization technologies, such as Software Defined Networking (SDN) and Network Function Virtualization (NFV).

The IC paradigm is supported by the definition of specific protocols and format that allow to create a network of clouds where content, storage, and computing functionalities are ubiquitous and interoperable [137]. The IC is composed of three main elements [138], which are shown in Figure 8.1: i) the *IC Root* is the core of the IC, which hosts the root servers containing all common mechanisms used by clouds participating to the IC, such as Naming Authority, Trust Authority, and Directory services. It acts as a broker in the IC overlay network and hosts a global cloud resource catalog which can be explored in order to discover desired cloud services; ii) the *IC eXchange* Points (ICXPs) are Neutral Access Points where clouds can interoperate and IC traffic can be properly handled (similarly to IXPs for network traffic). ICXPs provide negotiation and collaboration capabilities among heterogeneous and autonomous cloud environments. Each ICXP is affiliated with a particular IC Root element and hosts second-tier services; iii) the *IC Gateways* (ICGWs) are Internet routers representing interfaces between a particular cloud and the IC network. ICGWs translate IC requests and responses to the individual and customized protocol used by providers internally. Therefore, the IC addresses different problems related to cloud interoperability and specifically enables connectivity among disparate cloud environments avoiding the point-to-point collaboration (n^2 problem).

Being part of the IC, the ASs would be able to interoperate with each other for a combined QoE-aware management of network and services. For QoE monitoring, the OTT can make use of a cloud computing service to estimate in real-time the QoE of a specific service using QoE models (e.g., parametric models) specifically defined for the monitored service. A survey on parametric QoE estimation models for popular services is provided in [79]: such models are typically a function of both network KPIs and application KQIs. Being part of the IC, and establishing a collaboration agreement with the OTT, the ISP would be allowed to participate to the cloud computing service dedicated to QoE estimation to provide accurate information regarding the network KPIs, that the OTT alone could not be able to measure. The estimated QoE will be then the input of the Network management module run by the ISP, which in turn is a cloud computing service that takes decision on network management actions to target specific objectives depending on the collaboration approach between the providers.

To exchange information among the providers participating to the IC, we propose to use the XMPP protocol. Indeed, XMPP enables the near-real-time exchange of small pieces of structured data which are called *XML stanzas* [138]. XMPP specifies three types of XML stanzas: i) *message*, used to exchange messages among two entities over the XMPP overlay network; ii) *presence*, used to expose the availability of an entity in the XMPP overlay network; iii) *Info/Query (IQ)*, which allows for exchange request and response XML stanza. Besides stanza specific attributes, the XML root elements must also have addressing attributes, such as a *from* and a *to* attribute, whose values must contain a valid Jabber Identifier (JID) that in turn identifies the address of an XMPP entity. As an example, with reference to the aforementioned collaboration between OTT and ISP for QoE measurement, these would be the steps to be followed: i) the OTT uses the IQ stanza type `get` to explore the capabilities of the ISP's methods for QoS monitoring; ii) the ISP uses the IQ stanza type `result` to describe all QoS monitoring methods and actions that can perform; iii) the OTT can use common types defined for ISP's methods, i.e., `GET`, `POST`, `PUT`, and `DELETE`, which, together to the `request` and `response` elements (which describe the input and output representation data to be transferred), allow the OTT to make use of the ISP's needed methods for QoS monitoring.

Also, the message stanza can be used to exchange information between the providers, such as the measured QoE, the type of service provided by the OTT, and/or particular user and context information, which may help the ISP to drive QoE-aware network management algorithms.

An important point is what information is shared and how frequently. What information is shared depends on the collaboration agreement between the ISP and the OTT, so we refer to Section 8.4 where the collaboration approaches are presented. Since there are different types of information, we consider different updating frequencies for different types of information. We distinguish between static and dynamic information: *static information* does not change very frequently (weekly, monthly or even annually) and mainly regards some user information such as the user profile (end device, service subscription); this information can be updated only in case there are some changes. *Dynamic information* should be updated more frequently, especially when the user is active, because it regards service information such as application and context parameters as well as user's perceived QoE. Indeed, this information should be shared with the ISP to manage the network on the basis of service requirements, such as provide higher QoS in case the user's perceived QoE is too low. In order not to continuously exchange dynamic information, event-based techniques can be used so that information is exchanged only when specific events occur, such as quality degradation below specific thresholds [27], new services activated on the network, network congestion.

8.4 Collaboration strategies

In this section, we propose three collaboration approaches based on revenue/profit maximization and QoE fairness maximization. The first, (joint-venture, proposed in [3]), aims to maximize the revenue by providing better QoE to customers paying more. The second (CLV-based), aims to maximize the profit by providing better QoE to the most profitable customers (MPCs). The third (QoE-fairness-based), aims to maximize the QoE and the QoE fairness among all customers.

For each approach, we consider I OTT services (indexed by $i = 1,2,\dots,I$) that establish a collaboration agreement with an ISP to provide their services in J classes of service (CoS, indexed by $j = 1,2,\dots,J$). Also, we consider a total number N of customers that are in common among one (or more) of the OTTs and the ISP, i.e., the total number of users of the ISP's network. Table 8.2 lists the variables used in the proposed approaches.

Table 8.2 Variables used in the article.

Variable	Description
$i = 1, 2, \ldots, I$	Index of the OTT services
$j = 1, 2, \ldots, J$	Index of the classes of service (CoS)
$n = 1, 2, \ldots, N$	Index of the common customers between the ISP and all the OTTs
$l = 1, 2, \ldots, L$	Index of the location zones (LZs) controlled by the ISP
R	Revenue for the provider
P	Price paid by the customer to subscribe to the service
$U(QoE)$	Customer churn utility function
ζ	Number of new customers who subscribe with marketing
α, β	Weights to determine provider's value for the CVL-based approach
F	Fairness index
ρ	Relevance factor to indicate the importance of fairness

8.4.1 Joint-venture approach

The joint-venture (JV) approach refers to the JV collaboration between OTTs and ISP proposed in our past work. With such a collaboration, OTTs and ISP offer their services jointly into different classes of service: the higher the price to subscribe to a CoS, the higher is the quality of the service provided to the subscribers of that CoS. In this section, we briefly describe the proposed approach but more details are found in [3].

The objective of the JV collaboration is to maximize the total revenue for both ISP and OTTs generated at time t_x (with $x = 0, 1, 2, \ldots, X$), namely R^x, which is defined as

$$R^x = \sum_{i=1}^{I} R_i^x = \sum_{i=1}^{I} \sum_{j=1}^{J} N_{i,j}^x P_{i,j} = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(U_{i,j} N_{i,j}^{x-1} + \zeta_{i,j} \right) P_{i,j}, \tag{8.3}$$

where R_i^x is the combined revenue between the ISP and the i-th OTT, $N_{i,j}^x$ is the total number of customers subscribed to the j-th CoS of the i-th OTT at time t_x, $P_{i,j}$ is the price paid by the customers to subscribe to the j-th CoS of the i-th OTT, $U_{i,j}$ is the customer churn utility function for the j-th CoS of the i-th OTT, and $\zeta_{i,j}$ accounts for the customers subscribing to the j-th CoS of the i-th OTT with marketing and advertisement. The customer churn utility function computes the percentage of customers leaving/keeping the service as a function of the customer's perceived QoE

$$U_{i,j}(QoE_{i,j}) = \frac{1}{1 + e^{-z(QoE_{i,j} - QoE_{i,j}^m)}}, \tag{8.4}$$

where $QoE_{i,j}$ is the quality perceived by the customers subscribed to the j-th CoS of the i-th OTT, $QoE^{m}_{i,j}$ is the quality level at which half of the customers subscribed to the j-th CoS of the i-th OTT leave the service, and z is the sensitivity of the customers with respect to the paid price: customers who pay more expect higher quality [95].

Algorithm 1 describes the steps to be followed for the JV approach. The role of the OTT is to measure the average QoE perceived by the customers for each CoS ($\overline{QoE_{i,j}}$); then, this information is shared with the ISP, which has to maximize the revenue as a function of the QoE and the price paid by the customers so that customers who pay more receive higher QoE resulting in higher revenue for the providers. The average revenue is maximized during a reference period $t_{X_1} - t_{X_2}$ with the following equation

$$\bar{R}^{*} = \max_{\overline{QoE_{i,j}}, P_{i,j}} \left[\sum_{x=X_1}^{X_2} \left(\frac{\sum_{i=1}^{I} (\sum_{j=1}^{J} (U_{i,j} N_{i,j}^{x-1} + \zeta_{i,j})) P_{i,j}}{t_{X_2} - t_{X_1}} \right) \right]. \tag{8.5}$$

Therefore, the ISP manages its network in order to provide better QoS to the customers subscribed to higher-price CoS with the objective to maximize the QoE perceived by these customers and then reduce the customer churn for these CoS. Network traffic management techniques such as DiffServ, traffic prioritization, best routing path, are used by the ISP in this case. Although how the OTT and the ISP decide to divide the revenue is out of the scope of this article, a possibility may be to share the revenue proportionally to the amount of investment made by the providers over the total amount of investment for the service delivery. However, the criterion to be used in this scenario will be mostly based on political discussions among the different parties involved.

8.4.2 CLV-based approach

The JV approach presented in Section 8.4.1 is based on revenue maximization. However, as discussed in Section 8.2.2, the revenue just depends on the price paid by the customers to subscribe to the services, but it makes no consideration of the service profitability as well as of the cost of managing the customer service within the network. Therefore, it is important for the providers to understand the *value* of each customer in terms of profit, i.e., the revenue minus the costs. We identified the Customer Lifetime Value (CLV) as the metric for selection of the most profitable customers (MPCs) with the collaboration, i.e., those customers that for the providers have more value than others. Customers that are not MPCs are defined as standard customers (SCs). The CLV may be a function of several parameters, such as the revenue provided by the customer, the number of loyal years and the profit margin provided to the company. Also customer's information may be taken into account to determine the CLV,

ALGORITHM 1: Joint-venture approach

On the OTT side
for $\forall OTT_i,\ i \in I$ **do**
  for $\forall CoS_j,\ j \in J$ **do**
    OTT_i computes the average $QoE_{i,j}$, $\overline{QoE_{i,j}}$, perceived by the customers of the j-th CoS
    OTT_i computes the customer churn $U_{i,j}$ as a function of $\overline{QoE_{i,j}}$
    OTT_i takes into account the number of new customers $\zeta_{i,j}$
    OTT_i computes the current number of customers $N^x_{i,j}$
  end
  OTT_i computes the revenue R^x_i
  OTT_i shares with the ISP the $\overline{QoE_{i,j}}$ and R^x_i values
end
On the ISP side
for $\forall OTT_i,\ i \in I$ **do**
  for $\forall CoS_j,\ j \in J$ **do**
    ISP computes the total revenue R^x with Eq. (8.3)
    ISP maximizes $\bar{R}^*$ over $\overline{QoE_{i,j}}$ and $P_{i,j}$ with Eq. (8.5)
  end
end

such as laziness to churn (sleepers customers) or criticality (customers who often change subscription): in this case, low CLV may be assigned to sleepers whereas high CLV may be assigned to critical customers. Therefore, how the CLV is computed is strictly related to the needs of the providers and does not affect the validity of the proposed approach whose objective is to provide high quality to those customers that are the most valuable for the providers to avoid their churn. Customers with the highest values of CLV are considered as the MPCs and should be subjected to retention actions because the cost of winning a new customer is typically much larger than the cost of retaining an existing one [136], and the retention of the MPCs is even more convenient. As demonstrated in [3], the higher the QoE the lower the customer churn; therefore, the best way to decrease the churn of the MPCs is to provide them the best possible QoE.

We assume that ISP and OTTs have their own methods to compute the CLV (it may vary from company to company) but they just provide normalized CLV data not to disclose the real absolute values to the external world still allowing the collaboration to work properly. By considering the total number of common customers N, let CLV^{ISP}_n, $CLV_{n,i}$, and CLV^{OTT}_n be respectively the CLV of the n-th customer computed by the ISP, the i-th OTT and all the OTTs (if the customer is subscribed to more than one OTT). Then, we express the total CLV of the n-th customer as

$$CLV_n^{TOT} = \alpha CLV_n^{ISP} + (1-\alpha) CLV_n^{OTT} = \alpha CLV_n^{ISP} + (1-\alpha) \sum_{i=1}^{I} \beta_i CLV_{n,i}, \quad (8.6)$$

where α is a weight that determines the value of ISP and OTTs' customers for the total CLV computation, while β_i is a weight that determines the value of the OTTs in the collaboration agreement with the ISP: such weights are decided by ISP and OTTs on the basis of their agreement conditions. The last term takes into account the case of the same customer subscribed to more than one OTT: $\beta_i > 0$ when subscribed to the i-th OTT, otherwise $\beta_i = 0$. The N customers are then sorted as a function of CLV_n^{TOT}: the higher the CLV_n^{TOT}, the most profitable is the customer for both the ISP and the OTTs.

The objective of the ISP-OTT collaboration is to implement a QoE-aware management of the network to minimize the churn of the customers with the highest CLV_n^{TOT} so as to maximize the profit of both the providers. Firstly, the N customers are classified into two classes depending on their CLV_n^{TOT}. Specifically, a threshold is chosen to distinguish between the MPCs and the SCs, which depends on the difference of CLV_n^{TOT} among the customers. However, since resources are limited and costly, the percentage of MPCs should not be too high, such as the 20% or 30% for example. Though, these decisions should be evaluated case by case. Secondly, within the whole territory controlled by the ISP are identified the location zones (LZs) where most of the MPCs are situated. Let $l = 1, 2, \ldots, L$ indexes the LZs controlled by the ISP: these LZs are ranked on the basis of the number of MPCs belonging to each of them. Following this ranking, more network and application resources are provided to the highest ranked LZs to optimize the QoE of these customers and lose as few MPCs as possible.

Since the focus is on multimedia services, for QoE optimization several actions may be possible, such as i) network management implemented by the ISP for traffic prioritization on the basis of the type of service to be delivered [8]; ii) the places where the SSs should be located must be as close as possible to the highest ranked LZs; iii) the contents replicated on the SSs must be the most viewed by the MPCs. Indeed, as discussed in [25], CDN basically deals with the issues of identifying where to place the SSs and identifying what content to replicate on the servers. With the proposed collaboration approach, these issues are solved with a major focus on profit maximization. In fact, if the SSs will be situated closer to the highest ranked LZs, the MPCs belonging to these LZs will perceive better QoE due to the improvement of service delivery.

Algorithm 2 describes the steps to be followed for the CLV-based approach.

ALGORITHM 2: CLV-based approach

On the OTT side
for $\forall OTT_i,\ i \in I$ **do**
  for $\forall$ *Customer* $n \in N$ **do**
    OTT_i computes $CLV_{n,i}$ and $QoE_{n,i}$
  end
  OTT_i shares the information of $CLV_{n,i}$ and $QoE_{n,i}$ with the ISP
end
On the ISP side
for $\forall$ *Customer* $n \in N$ **do**
  ISP computes CLV_n^{ISP}, CLV_n^{OTT} and CLV_n^{TOT}
  ISP sorts the N customers according to CLV_n^{TOT} to define the MPCs and SCs
  ISP sorts the L LZs according to the number of the MPCs
  for $\forall LZ_l$, $l \in L$ **do**
    ISP places the SSs in the LZs where there is the highest number of the MPCs
  end
end

8.4.3 QoE-fairness-based approach

The collaboration approaches proposed in Sections 8.4.1 and 8.4.2 are based respectively on revenue/profit maximization by providing better QoE to customers paying more or which are the most profitable for the providers. However, one of the ultimate goals in future multimedia networks is to provide a fair user-centric quality, so that the user QoE is maximized for all users in a network. The reason is that selfish applications may maximize its own QoE, potentially worsening QoE levels of users of different applications. According to the general fairness metric defined in [132], which satisfies QoE-relevant properties, a system is defined absolutely QoE fair when all users receive the same QoE value. Therefore, the objective of the proposed collaboration scenario is to maximize the QoE and the QoE fairness of all the N customers in common among the ISP and the OTTs. Specifically, we consider maximization of QoE and QoE fairness within each CoS $j \in J$ because customers who pay more expect higher quality [95].

With regard to the QoE fairness index, F, we refer to that defined in [132], as follows

$$F = 1 - \frac{\sigma}{\sigma_{max}}, \tag{8.7}$$

where σ and σ_{max} are respectively the standard deviation and the maximum possible standard deviation of the QoE levels which may be maximally perceived by the users. Then, the higher the variance among the QoE values perceived by different users, the lower the fairness.

Also, the fairness value depends on the considered quality scale because $\sigma_{max} = \frac{H-L}{2}$, where H and L are respectively the maximum and minimum values of the QoE evaluation scale. For example, considering the ITU 5-level quality scale so that $L = 1$ and $H = 5$, Eq. (8.7) becomes

$$F = 1 - \frac{2\sigma}{5-1} = 1 - \frac{\sigma}{2}. \tag{8.8}$$

Therefore, considering the 5-level quality scale, the QoE fairness index computed among all users of the j-th CoS of the i-th OTT, $F_{i,j}$, is

$$F_{i,j} = 1 - \frac{\sigma_{i,j}}{2} = 1 - \frac{1}{2}\sqrt{\frac{\sum_{n=1}^{N_{i,j}} (QoE_{n,i,j} - \overline{QoE_{i,j}})^2}{N_{i,j}-1}}, \tag{8.9}$$

where $\sigma_{i,j}$ is the standard deviation of the QoE levels perceived by the customers subscribed to the j-th CoS of the i-th OTT, $QoE_{n,i,j}$ is the QoE perceived by the n-th customer subscribed to the j-th CoS of the i-th OTT, and $\overline{QoE_{i,j}}$ is the average QoE perceived by all $N_{i,j}$ customers subscribed the j-th CoS of the i-th OTT.

On the other hand, the QoE fairness of the network, computed among all the N_j customers of the ISP's network subscribed to the j-th CoS of all the OTTs, F_j^{ISP}, is defined as

$$F_j^{ISP} = 1 - \frac{\sigma_j^{ISP}}{2} = 1 - \frac{1}{2}\sqrt{\frac{\sum_{i=1}^{I}\sum_{n=1}^{N_{i,j}} (QoE_{n,i,j} - \overline{QoE_j})^2}{N_j-1}}, \tag{8.10}$$

where σ_j^{ISP} is the standard deviation of the QoE levels perceived by all customers subscribed to the j-th CoS of all the OTTs, $QoE_{n,i,j}$ is the QoE perceived by the n-th customer subscribed to the j-th CoS of the i-th OTT, and $\overline{QoE_j}$ is the average QoE perceived by all N_j customers subscribed to the j-th CoS of all the OTTs.

The objective is to maximize the fairness index, F_j^{ISP}, as well as the average QoE of all customers, $\overline{QoE_j}$, within each CoS. To this, we refer to the maximum utility function proposed in [139]

$$U_j = (1-\rho)\overline{QoE_j} + \rho F_j^{ISP}, \tag{8.11}$$

where ρ is the relevance factor that indicates the importance of fairness in the management decision. On the one hand, the OTTs predict the QoE perceived by their customers using QoE metrics that may take into account application, human and context influence factors, and try to maximize this QoE by making the best possible use of the available network resources. On the other hand, OTTs do not have any control over the network infrastructure, whose performance has a strong influence on QoE and is shared among multiple and different OTT services. Therefore, the collaboration with the OTTs, that share the computed QoE of the

ALGORITHM 3: QoE-fairness-based approach

On the OTT side
for $\forall\ OTT_i,\ i \in I$ **do**
  for $\forall\ CoS_j,\ j \in J$ **do**
    for $\forall\ Customer\ n \in N$ **do**
      OTT_i computes the QoE perceived by the $N_{i,j}$ customers subscribed to the j-th CoS, $QoE_{n,i,j}$
    end
    OTT_i shares with the ISP the information of $QoE_{n,i,j}$
  end
end
On the ISP side

– ISP computes the average QoE perceived by all the N_j customers of the j-th CoS, $\overline{QoE_j}$
– ISP manages the network for the combined maximization of $\overline{QoE_j}$ and F_j^{ISP} using Eq. 8.11

customers, allows the ISP to globally manage the network for the combined maximization of the overall QoE and QoE fairness. The value of the relevance factor ρ is decided by ISP and OTTs depending on their agreements.

Algorithm 3 describes the steps to be followed for the QoE-fairness-based approach.

8.4.4 Limits and assumptions

Each of the proposed three models is based on some assumptions and has some limits. The JV approach assumes that OTT and ISP jointly provide the service to the customers, which have to paid with a single subscription. The main limit is to define how OTT and ISP should divide this revenue because it is not easy to translate the information and effort provided by the providers in terms of money. From the technical point of view, the network should be able to give priority to the traffic of the priority classes, which however is a reasonable assumption considering the currently available QoS technologies. The CLV-based approach assumes that the providers measure the value of each customer as a function of some specific parameters. The main limit is that a customer that is profitable for the ISP (OTT) may not be profitable for the OTT (ISP or some other OTTs). Therefore, how all customers should be ranked in terms of total profitability should be accurately defined by the providers not to fail to meet some provider's collaboration expectations (in terms of profit). Finally, the QoE-fairness-based approach is based on the provision of fair QoE levels to customers of different services. This approach has the limit that the utilization of network resources may violate the network neutrality because network QoS is not linearly related to the user's perceived QoE. As an example, the amount of network resources needed to provide a high quality video streaming service are typically more than those needed to provide a high quality VoIP or web browsing service. It is however important to note that also the other two techniques are violating somehow the network neutrality principle.

8.5 Simulations-based analysis

The simulations are performed to analyze the effectiveness of the aforementioned collaboration strategies in terms of the QoE delivered to the users, QoE fairness among the users and providers' profit. Section 8.5.1 provides the details of the simulation setup whereas in Section 8.5.2 simulation results are shown and discussed.

8.5.1 Simulation setup

The simulations are conducted on the MATLAB platform. We considered two different OTT services: VoIP and video streaming. The aforementioned applications are selected according to the studies conducted in [15, 16], where video streaming and VoIP are considered as the most sensitive multimedia applications with reference to network resources utilization. The setting of the major parameters have been done as follows. Table 8.3 summarizes simulation parameters and settings.

CoS, prices and costs: For each of the proposed approaches, two different CoS are considered, on the basis of the users' willingness to pay: 1) Premium (PR) service and; 2) Standard (ST) service. The price to subscribe to the PR service is higher than that for the ST service, because the PR service should guarantee a higher quality experience to the users. As a consequence, the providers may incur higher costs for the provision of the PR service than for the ST service, which is provided through the best effort approach. Prices and costs are normalized in the $0-1$ scale. We set the price to be paid by PR and ST users respectively to 0.6 and 0.3, on the basis of our previous studies, where maximum revenue was generated on these prices [3]. Additionally, we consider the cost associated to PR and ST users to range randomly between 0.1 and 0.5, and between 0.1 and 0.2, respectively, so that a minimum profit of 0.1 is always assured to the providers. We consider the cost to vary from user to user because many causes may infer service costs, such as the distance of the user from the server. Note that for the CLV-based approach, both PR or ST users may be considered as MPCs or SCs, because they are further classified on the basis of their CLV value.

Table 8.3 Simulation parameters and settings.

Parameters	**Settings**
Initial total No. of VoIP and video streaming users	100
Initial No. of PR and ST users in VoIP and video streaming	50
Price paid by PR users	0.6
Price paid by ST users	0.3
Cost associated with PR users	$0.1-0.5$[1]
Cost associate with ST users	$0.1-0.2$[1]
No. of loyal years	$1-5$[1]
Sensitivity of the user with the price, z	3.5
Quality level at which half of ST users leave the service, $QoE^{m}_{i,ST}$	2.5
Quality level at which half of the PR users leave the service, $QoE^{m}_{i,PR}$	3.5
Weights for the CLV-based approach, α, β	0.5

Initial population and churn: We consider a starting number of 100 users for both VoIP and video streaming applications, with 50% of the users belonging to the PR service and 50% to the ST service. Each user is assigned with the following characteristics: 1) price paid by the user and cost associated to keep the user in the service, which depend on the CoS the user belongs to; 2) profit associated with the user, calculated as the price minus the cost and; 3) number of loyal years of the user in the service, which are assigned randomly in the range $1-5$ years. In order to make the fair comparison among the collaboration strategies, the same initial pole of users are considered for each strategy. The user churn is calculated using the Eq. (8.4), where $z = 3.5$ for each CoS while $QoE^{m}_{i,j}$ is selected to be 2.5 and 3.5 for ST and PR users, respectively. Note that $QoE_{i,j}$ is computed as the average QoE of all users' QoE predictions per month. On the basis of the computed user churn, the number of the users in each CoS and type of application are updated on the monthly basis. We also consider the number of users joining the service each month $\zeta^{x}_{i,j}$, which is computed as a random number from the Poisson distribution with the mean equal to the 5% of users belonging to each CoS of an application. A random cost is assigned to each new user, depending on the CoS the user belongs to. Furthermore, the fairness index in each CoS is computed on the monthly basis using Eq. (8.10).

Quality models: We consider the E-model for the VoIP application, which is a planning parametric model defined by the ITU for VoIP applications, which measures the quality in terms of the R-factor ranging from 0 to 100, where 100 is the best quality [104]. In our simu-

[1]These parameters are randomly assigned with uniform distribution.

lations, we use the simplified version of the E-model, proposed in [106], which emphasizes the effect of sources of quality degradation observed over data networks, namely one-way delay, packet loss ratio, and coding scheme. Furthermore, we considered [106] for converting the R-factor with values between 1 and 5 as the MOS. For the video streaming service, the QoE is predicted by using the model proposed in [81, 125] because it is particularly designed for the non adaptive HTTP video streaming

$$MOS = \alpha \cdot e^{-\beta(L) \cdot N} + \gamma, \tag{8.12}$$

where L and N are the stalling duration and the total number of stalling events during the video session, respectively. α, γ and $\beta(L)$ are some formula parameters that have been set as follows: $\alpha = 3.5$, $\gamma = 1.5$ and $\beta(L) = 0.15 \cdot L + 0.19$. Moreover, for a 30-second-long video session in [125], L and N are defined as a function of the buffer size d of the client player, the frame arrival rate λ and the video frame rate μ, such that $L = \frac{d}{\mu \cdot a}$ and $N = \frac{1-a}{d^*}$, whereas $a = \frac{\lambda}{\mu}$ and $d^* = \frac{d}{\mu}$ are the load factor and the normalized buffer size respectively. We consider the buffer size equal to 4 s as recommended in [125] for the video streaming case. In the simulations, network throughput and video bit rate are considered respectively as λ and μ (instead of frame arrival rate and video frame rate), as also considered in the model in [125].

Quality as a function of the number of customers: To understand the relationship between the QoE and the number of users we have used the network emulator Mininet. It is important to highlight that this analysis has been conducted offline with respect to the major MATLAB simulation framework. Indeed, we specifically needed the curves of quality versus the number of customers for the considered network. In the emulator, we considered the ISP's network to be composed of different virtual slices with a fixed channel capacity for each different CoS. For the VoIP service encoded with G.729a, the number of UDP flows with the encoding bit rate (equal to 31.2 kbps) are varied from 50 to 100 over a fixed capacity link of 1.95 Mbps. The mean playout delay and jitter are observed by varying the number of flows for both PR and ST VoIP use cases. The packet loss probability is computed from the cumulative density function of the observed mean playout delay and jitter as highlighted in [140]. The measured playout delay and packet loss probability are used for the computation of the R-factor [104] and delivered QoE in terms of MOS [106]. With regard to video streaming, the videos were encoded with the H.264 codec. The PR users stream 1080p HD in 1920×1080 resolution while the ST users stream 480p SD videos in 854×480 resolution; the frame rate for both HD and SD videos is 30 fps. The encoding bit rate for HD and SD videos varies from 3 Mbps to 6 Mbps and 500 Kbps to 2 Mbps, respectively.[4] For the video streaming, the traffic composed of TCP flows is generated where the number of flows is varied from 50 to 100 over fixed capacity links of 375 Mbps and 125 Mbps for PR and ST CoS respectively. The observed change in the throughput with the increase in number flows/users is used to measure the delivered QoE from Eq. (8.12). With this setting we wanted to stress the network with also sending traffic quite exceeding the network capacity. We would like to highlight that this setting an extensive modelling of quality versus the number of customers is not the main focus of the chapter but it was important to conduct the analysis of the proposed algorithms performance.

Performed simulations: With the described setting, we ran the simulations that implemented the three proposed algorithms in MATLAB and we observed the evolution of the system, i.e., number of customers per class and service and the quality provided, over time providing results as the average over 100 runs. Each approach manages the network resources differently: the JV approach provides more network resources to PR users, who are the users paying more, to maximize the revenue (see Section 8.4.1); the CLV approach provides more network resources to the Most Profitable Customers (MPC) to maximize the profit (see Section 8.4.2); the QoE-fairness-based approach provides fair network resources to maximize QoE and QoE fairness among users of the same CoS (see Section 8.4.3). Again, to provide a fair comparison of the collaboration approaches, we assume that the considered ISP's network portion, where we apply the CLV-based algorithm, is the LZ where there is the highest number of MPCs. So we are not considering multiple LZs in the simulations.

8.5.2 Simulations results

This section provides the simulation results regarding the comparison of the three collaboration strategies, presented in Section 8.4, in terms of quality provided to the users, QoE fairness among the users and profit generated for the providers.

Figure 8.2 shows the average QoE delivered in each CoS of both video streaming and VoIP applications over a period of 2 years for each collaboration strategy. In case of the PR VoIP and video streaming, the CLV-based approach provides better quality as compared to JV and QoE-fairness-based approaches, i.e., 4.38 and 4.74 for PR VoIP and PR video streaming, respectively. The reason is that the number of users classified as MPCs (and are treated with PR CoS) on the basis of CLV (30%) is less that those of the JV and QoE-fairness-based approaches (50% are PR users). Hence, better quality is delivered to MPCs in case of CLV-based approach because more resources are available for less users. However, in case of ST VoIP and video streaming, the CLV-based approach provides lower quality, i.e., 3.23

[4] https://support.google.com/youtube/answer/2853702?hl=en

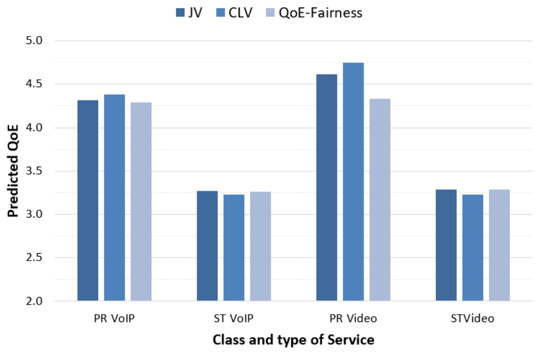

Fig. 8.2 Comparison in terms of delivered QoE.

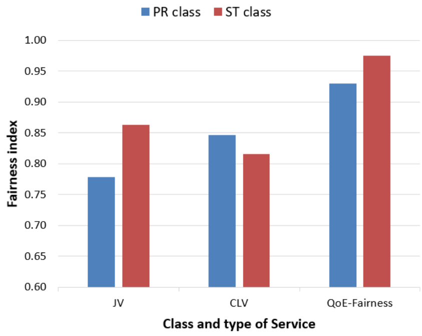

Fig. 8.3 Comparison in terms of QoE fairness.

and 3.22 as compared to JV and QoE-fairness-based approaches because more users are classified as SCs (which are treated as ST users) on the basis of their CLV. Hence, lower quality is perceived by the SCs for VoIP and video services. Moreover, the JV approach appears to deliver better quality for PR video as compared to QoE-fairness-based approach.

Figure 8.3 provides the comparison in terms of the QoE fairness among the different collaboration strategies. The QoE fairness is computed within the two CoS, PR and ST, considering the QoE measured for the customers of both the video streaming and VoIP services. As expected, the QoE-fairness-based approach provides higher degree of fairness in terms of QoE among the same CoS. In case of PR VoIP and PR video streaming, the CLV-based approach provides more fairness at application level as compared to JV because the number of MPCs is lower than that of PR users of JV, hence, higher (and fair) quality is provided to the MPCs. However, in case of ST VoIP and ST video streaming, JV provides better fairness in comparison with the CLV-based approach for the opposite reason. Indeed, there are less ST users in JV than SCs in the CLV-based approach.

Figures 8.4-8.6 show the evolution in the number of users for the different approaches by the end of every month over a period of 2 years. The 95% confidence interval is shown, which is computed for the 100 runs executed. With regard to the JV approach, the number of ST and PR users increases with the time. However, the number of PR users has an important increase whereas the number of ST users has a slight increase reaching a saturation level. This is due to the lower quality perceived by the ST users (which pay less than the PR users but do not have a good guaranteed quality) that increases the number of churners. The number of users for the CLV-based evolves differently than that for the JV approach. Indeed, while for the JV approach 50% of the users paying less are considered ST users (50 users), for the CLV-based approach the users are further classified on the basis of CLV and mostly (70%, i.e., 70 users) are considered as SCs. Then, at the initial time most of network resources

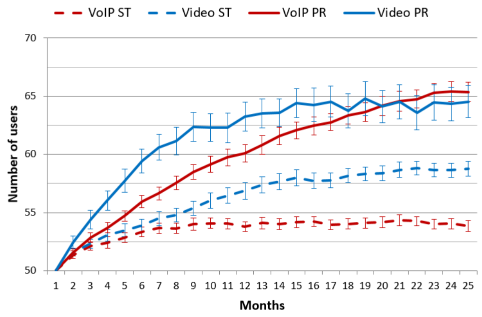

Fig. 8.4 Evolution in the number of users for the Joint Venture (JV) approach.

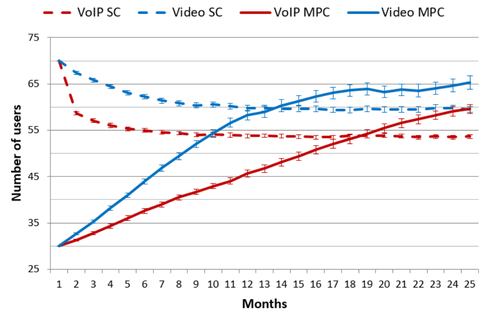

Fig. 8.5 Evolution in the number of users for the CLV-based approach.

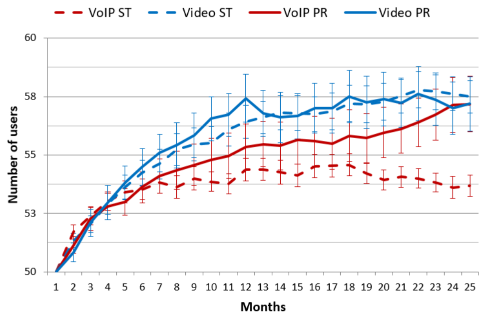

Fig. 8.6 Evolution in the number of users for the QoE-fairness-based approach.

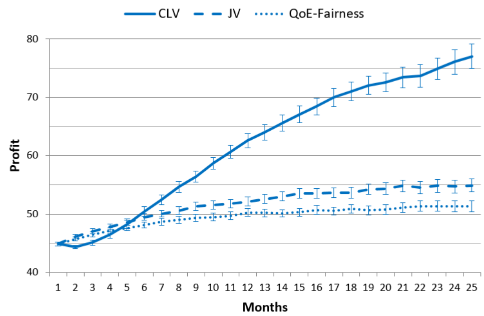

Fig. 8.7 Comparison in terms of the profit.

are already allocated for the SCs, who receive low quality services and leave the service. However, after some months a saturation level is reached. The opposite situation concerns the MPCs, which are lower than the SCs at the initial time, but have an important grow over time due to the greater amount of network resources available which allow them to receive good service quality. Finally, with regard to the QoE-fairness-based approach, after an initial time in which the number of users for each CoS and service increases, the evolution of the number of users becomes more unpredictable and indented. This may be due to the fairness algorithm that, trying to provide fair QoE to each user, may provide insufficient network resources that in turn transform into insufficient QoE to the users, which become churners. Then, more network resources are available that allow new users to join. Again, the action of the algorithm may decrease the QoE causing new churners and so on. We remind that the value of ρ set for these simulations in Eq. (8.11) was $\rho = 0.9$ so as to apply a fairness-aware solution.

Finally, Figure 8.7 compares the collaborative approaches on the basis of the total profit generated by the providers over a period of two years. The profit per user is computed as the difference between the price paid by the user to subscribe to the CoS and the cost associated to keep the user into that CoS. The total profit is calculated as the sum of the profit of each user of both PR and ST CoS and both VoIP and video streaming applications. As expected, the CLV-based approach provides the highest profit, followed by the JV approach and the QoE-fairness-based approach. Differently from the JV approach, that provides better quality to users just paying more, the CLV-based approach considers the profit provided by each customer, by taking into account the cost associated to each user in addition to the price paid. Therefore, the CLV-based approach is able to provide more resources and then higher quality to the users who are the most profitable for the providers and then the total profit is higher with respect to the other two approaches. For example, it may happen that some users classified as PR by the JV approach are actually not profitable because their associated cost is high and they are customers just since one year (for example, $Price = 0.6$, $Cost = 0.5$, $Profit = 0.1$, $years = 1$, $CLV = 0.6 \times 0.1 \times 1 = 0.06$). On the contrary, some users classified as ST may be more profitable because their associated cost is low and they are customers from 5 years (for example, $Price = 0.3$, $Cost = 0.1$, $Profit = 0.2$, $years = 5$, $CLV = 0.3 \times 0.2 \times 5 = 0.3$). In this way, higher quality is provided not only to the most profitable customers but also to the loyal customers, and maybe even better results may be obtained by this approach if a model which takes into account also user's loyalty would be used (we based on the user churn model which defines if the user is satisfied just on the basis of the perceived quality and the price paid, loyalty is not modelled). Finally, as also discussed before, the QoE-fairness-based approach is mostly based on providing fair QoE to the users. For this reason, sometimes insufficient QoE level is provided to the users which become churners decreasing the profit of the providers.

8.6 Conclusions and Future Work

In this chapter, we considered the collaboration among ISPs and OTTs for a joint QoE-aware service management in terms of technical and economic aspects. Accordingly, we proposed a reference architecture for a possible collaboration and information exchange among them. We also proposed three different collaboration strategies, namely: Joint-venture, CLV-based and QoE-fairness-based approaches. Then, we conducted simulations to compare the performance of these approaches in terms of quality delivered to the users, QoE fairness among users and total profit generated by providers. From simulation results, the CLV-based and JV approaches provide better QoE to PR and ST users, respectively. As expected, the QoE-fairness-based approach provides higher degree of QoE fairness among the users of the same CoS as well as on the network level. With regard to the profit, the CLV-based approach provides the best results, followed by the JV approach and the QoE-fairness-based approach.

In future works, we aim to improve the simulations of our proposed collaboration approaches by considering virtualized networks, such as Software-Defining Networks (SDN), to test the performance of the approaches on a virtual network environment. Indeed, with SDN it would be possible to run the algorithm on the SDN controller whereas on the network plane management actions can allocate network resources on the basis of the algorithm's decisions. Also, further studies are required to evaluate the impact of frequency of exchange of information among the OTTs and ISP on the network load as well as on the amount of data and the cost of data storage. Moreover, QoE models valid for longer time periods need to be studied in combination with a more accurate user churn function, which are required to drive the collaborative management of the services. Furthermore, more accurate models are required to correlate the change in the QoE and QoS with the change in the number of users/flows.

Chapter 9

Conclusion and Future Work

In this thesis, the concepts of the Quality of Experience regarding collaborative QoE monitoring and management by OTT and ISP are studied where novel approaches as compared to the state-of-the-art work are proposed.

Chapter 2 provides the discussion of challenges that ISPs need to face when implementing QoE monitoring systems for future networks. One of these challenges stems from the fact that ISPs and OTTs will generally move from a network-centric to a more user-centric perspective. Traditional SLAs are not applicable for negotiating and communicating quality to customers, and it can be imagined that QoE will take a more prominent role here — also in the network optimization goals for ISPs.Another one of the most important issues is introduced by the increased amount of application-level encryption. How can ISPs monitor application-level data when most OTT traffic is going to be end-to-end encrypted? Here, current works explore machine learning approaches which allow ISPs to predict QoE from lower-level network measurements. Another proposed solution would be a collaboration between OTTs and ISPs. To be more precise, a side-channel could allow OTTs can report certain KPIs and KQIs to ISPs. However, net neutrality concerns play a large role in such a scenario: although net neutrality regulations mostly affect how traffic shaping is being implemented by an ISP, the use of reliable QoE models to quantify the user experience is a precondition for that. We conclude that ISP–OTT collaboration is a highly political topic, but there are technological opportunities that offer ways for both parties to create better services for their users.

In Chapter 3, a multilayered QoE monitor solution at the user terminal is proposed for measuring the quality of multimedia applications. The proposed approach considers the perspective of ISPs for QoE monitoring. The proposed approach provides a cost-effective and scalable solution for measuring multimedia application quality which is not only flexible but also easier to deploy with the current network infrastructure without bringing any change into the existing network architecture. In this chapter, we also provide the novel concept of the discrete time QoE monitoring at user terminal where we also present the investigation of the impact of the multilayer user end probe

Chapter 4, a multilayer passive monitoring solution is proposed where the activation of the probe is based on the quality degradation event. From the results of a survey regarding user's privacy and preferences, we motivate the user's willingness for the installation of the probe in the terminal. The proposed approach is not only cost-effect to be incorporated into the current infrastructure of ISPs' network but also scalable. The experimental results highlight the novelty of the approach in terms of an on-device resource such as CPU, RAM and battery level. The proposed approach outperforms the traditional constant frequency based monitoring approach while utilizing less on-device resources. The future research will be focused on the investigation of cloud service utilization w.r.t to the activation of probes as well as consideration of different layers in the monitoring solution.

Chapter 5 proposes a novel approach for collaborative QoE management between OTTs and ISPs. The proposed model also takes into account important factors such as the user churn, pricing, and marketing, making it the novel. Also, with the consideration of the different QoE models for different applications, we investigated the flexibility and adaptability of our collaboration model which has proven to be robust, reliable and adaptable with respect to any change in QoE model. Furthermore, with simulations, we highlighted that if ISPs and OTTs adopt the proposed collaboration approach they will not only increase the revenue but will also provide better QoE to their users with relatively lower prices. Though the QoE based service delivery requires the collaboration among OTTs and ISPs, the research in this domain is suffering from key challenges. One of these is that no inter-operable interface exists to date which contributes towards scalability of the approach. Hence, it will not only require standardized interfaces among OTTs and ISPs to exchange QoE based information but it will also require standardized interfaces among ISPs like peering connections or exchange points to share QoE related information. Therefore, the future research should focus on the provision of QoE-centric interfaces between OTTs-ISPs to enable them QoE based service delivery. The Software Defined Networks (SDN) and Network Function Virtualization (NFV) can provide an opportunity in this regard because of their programmability and flexibility. However, scalability and security remain as an open issue in collaboration even if centralized SDN controller will be used for the QoE management. The computational complexity for the QoE measurements may appear to be another issue contributing to the scalability of the collaborative approach and complexity may increase with the increase in the number of OTT applications and customers. Moreover, there will be the requirement of storing data related to QoE, user churn and revenue generation which may increase the cost of network planning and operations as well. Additionally, although big efforts have been conducted in

QoE modeling, most of these are developed for providing an estimation of the perceived quality for very short periods of time. This aspect raises a practical issue when applying the resulting models to the considered scenario where the QoE affecting the churn should provide the level of experience quality resulting from long periods of service consumption. Hence, the development of robust and reliable QoE models valid for longer periods of time is essential for QoE based service delivery. Notwithstanding the importance of user churn, no model has been proposed which can correlate user churn with QoE which is important at the enterprise level. Nevertheless, the creation of user churn prediction model will be requiring the real customer data and analysis of that data over the significant periods of time. Moreover, the Network Neutrality and user privacy is also another future challenge for collaborative QoE based service delivery. All these challenges need to be taken into account in the future research so that QoE based service delivery can be possible.

In Chapter 6, a QoE-aware collaboration approach between OTT and ISP is proposed, based on the maximization of the profit. The objective is to improve the service delivery to the LZs with the highest number of Most Profitable Customers (MPCs) (classified in terms of the Customer Lifetime Value (CLV)) by distributing Surrogate Servers (SRs) closer to these LZs. The simulation results show that the proposed approach delivers better QoE as compared to the no collaboration approach.

Chapter 7 discusses the SDN based emulation platform called *Timber* for QoE management experimental research purposes. *Timber* is developed on the top of Mininet SDN emulator using Ryu controller. *Timber* has the capability to not only monitor the KQIs of video streaming session through the client side (user end probe) but also to perform QoE-aware traffic engineering tasks through SDN controller application. This work provides the architectural and implementation details of *Timber*, which can be used by the research community to evaluate future QoE management and monitoring strategies. Moreover, we provide experimental results to highlight the major functionalities of *Timber* by 4 different scenarios which include traffic shaping through DiffServ and dynamic resource allocation by queuing strategies. The future work will be focused on adding more multimedia services to the platform.

Chapter 8 provides the comparative analysis of the collaborative QoE management strategies namely: Joint-venture, CLV-based and QoE-fairness-based approaches are considered. Then, we conducted simulations to compare the performance of these approaches in terms of quality delivered to the users, QoE fairness among users and total profit generated by providers. From simulation results, the CLV-based and JV approaches provide better QoE to premium and standard users, respectively. As expected, the QoE-fairness-based approach provides the higher degree of QoE fairness among the users of the same CoS as well as on the network level. With regard to the profit, the CLV-based approach provides the best results, followed by the JV approach and the QoE-fairness-based approach.

In future works, we aim to improve the simulations of our proposed collaboration approaches by considering virtualized networks, such as Software-Defining Networks (SDN), to test the performance of the approaches on a virtual network environment. Indeed, with SDN it would be possible to run the algorithm on the SDN controller whereas on the network plane management actions can allocate network resources on the basis of the algorithm's decisions. Also, further studies are required to evaluate the impact of frequency of exchange of information among the OTTs and ISP on the network load as well as on the amount of data and the cost of data storage. Moreover, QoE models valid for longer time periods need to be studied in combination with a more accurate user churn function, which is required to drive the collaborative management of the services. Furthermore, more accurate models are required to correlate the change in the QoE and QoS with the change in the number of users/flows.

References

[1] A. Raake, J. Gustafsson, S. Argyropoulos, M. Garcia, D. Lindegren, G. Heikkila, M. Pettersson, P. List, B. Feiten *et al.*, "IP-based mobile and fixed network audiovisual media services," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 68–79, 2011.

[2] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2016-2021 - White paper," 2017. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf

[3] A. Ahmad, A. Floris, and L. Atzori, "QoE-centric service delivery: A collaborative approach among OTTs and ISPs," *Computer Networks*, vol. 110, pp. 168–179, 2016.

[4] P. Le Callet, S. Möller, A. Perkis *et al.*, "Qualinet white paper on definitions of quality of experience," *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)*, 2012.

[5] ITU-T Rec. P.10/G.100, *Vocabulary for performance and quality of service. Amendment 5: New definitions for inclusion in Recommendation ITU-T P.10/G.100*, ITU-T Recommendation, July 2016.

[6] W. Robitza, A. Ahmad, P. A. Kara, L. Atzori, M. G. Martini, A. Raake, and L. Sun, "Challenges of future multimedia qoe monitoring for internet service providers," *Multimedia Tools and Applications*, vol. 76, no. 21, pp. 22 243–22 266, 2017.

[7] T. Hossfeld, M. Fiedler, and T. Zinner, "The QoE provisioning-delivery-hysteresis and its importance for service provisioning in the Future Internet," in *7th EURO-NGI Conference on Next Generation Internet (NGI)*. IEEE, 2011, pp. 1–8.

[8] J. Seppanen, M. Varela, and A. Sgora, "An autonomous QoE-driven network management framework," *Journal of Visual Communication and Image Representation*, vol. 25, no. 3, pp. 565–577, 2014.

[9] S. Baraković and L. Skorin-Kapov, "Survey and challenges of QoE management issues in wireless networks," *Journal of Computer Networks and Communications*, vol. 2013, 2013.

[10] Y.-C. Chen, J.-W. Chang, and H.-Y. Wei, "A multi-level QoE framework for smartphone video streaming applications," in *Globecom Workshops (GC Wkshps), 2014*. IEEE, 2014, pp. 225–230.

[11] P. Rengaraju, C.-H. Lung, F. R. Yu, and A. Srinivasan, “On QoE monitoring and E2E service assurance in 4G wireless networks,” *Wireless Communications, IEEE*, vol. 19, no. 4, pp. 89–96, 2012.

[12] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, “An architecture for differentiated services,” 1998.

[13] S. Shenker, “Specification of guaranteed quality of service,” 1997.

[14] J. Zhang and N. Ansari, “On assuring end-to-end QoE in next generation networks: challenges and a possible solution,” *Communications Magazine, IEEE*, vol. 49, no. 7, pp. 185–191, 2011.

[15] G. Gómez, J. Lorca, R. García, and Q. Pérez, “Towards a QoE-driven resource control in LTE and LTE-A networks,” *Journal of Computer Networks and Communications*, vol. 2013, 2013.

[16] P. Oliver-Balsalobre, M. Toril, S. Luna-Ramírez, and J. M. R. Avilés, “Self-tuning of scheduling parameters for balancing the Quality of Experience among services in LTE,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, no. 1, pp. 1–12, 2016.

[17] E. Liotou, D. Tsolkas, N. Passas, and L. Merakos, “Quality of experience management in mobile cellular networks: key issues and design challenges,” *Communications Magazine, IEEE*, vol. 53, no. 7, pp. 145–153, 2015.

[18] V. Chellappan, K. M. Sivalingam, and K. Krithivasan, “A centrality entropy maximization problem in shortest path routing networks,” *Computer Networks*, vol. 104, pp. 1–15, 2016.

[19] A. Khan, L. Sun, E. Jammeh, and E. Ifeachor, “Quality of experience-driven adaptation scheme for video applications over wireless networks,” *IET Communications*, vol. 4, no. 11, pp. 1337–1347, 2010.

[20] O. Awobuluyi, J. Nightingale, Q. Wang, and J. M. Alcaraz-Calero, “Video quality in 5G networks: Context-aware QoE management in the SDN control plane,” in *2015 IEEE Int. Conf. on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM).* IEEE, 2015, pp. 1657–1662.

[21] (2000) A framework for integrated services operation over diffserv networks. NWG RFC 2998.

[22] M. Varela, P. Zwickl, P. Reichl, M. Xie, and H. Schulzrinne, “From service level agreements (sla) to experience level agreements (ela): The challenges of selling qoe to the user,” in *Communication Workshop (ICCW), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1741–1746.

[23] T. Hoßfeld, L. Skorin-Kapov, Y. Haddad, P. Pocta, V. A. Siris, A. Zgank, and H. Melvin, “Can context monitoring improve QoE? a case study of video flash crowds in the internet of services,” in *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM).* IEEE, 2015, pp. 1274–1277.

[24] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A Survey on Quality of Experience of HTTP Adaptive Streaming," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 469–492, 2015.

[25] A. Passarella, "A survey on content-centric technologies for the current internet: Cdn and p2p solutions," *Computer Communications*, vol. 35, no. 1, pp. 1–32, 2012.

[26] A. Ahmad, L. Atzori, and M. G. Martini, "Qualia: A multilayer solution for QoE passive monitoring at the user terminal," in *IEEE Int. Conf. on Communications (ICC), 2017*. IEEE, 2017, pp. 1–6.

[27] A. Ahmad, A. Floris, and L. Atzori, "Towards qoe monitoring at user terminal: A monitoring approach based on quality degradation," in *Proc. of the IEEE Int. Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), 2017*. IEEE, 2017.

[28] ——, "OTT-ISP joint service management: A customer lifetime value based approach," in *Accepted for IFIP/IEEE International Symposium on Integrated Network Management*. IEEE, 2017.

[29] ——, "Timber: An sdn based emulation platform for qoe management experimental research," in *Quality of Multimedia Experience (QoMEX), 2018 Tenth International Conference on*. IEEE, 2018.

[30] A. Floris, A. Ahmad, and L. Atzori, "Qoe-aware ott-isp collaboration in service management: Architecture and approaches," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 2s, p. 36, 2018.

[31] P. Le Callet, S. Möller, and A. Perkis, "Qualinet White Paper on Definitions of Quality of Experience," European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Tech. Rep. March, 2013.

[32] Y. Chen, W. Kaishun, and Z. Qian, "From QoS to QoE: A survey and tutorial on state of art, evolution and future directions of video quality analysis," *IEEE Communications Surveys and Tutorials*, 2014.

[33] M. Alreshoodi and J. Woods, "Survey on QoE/QoS correlation models for multimedia services," *International Journal of Distributed and Parallel Systems*, vol. 4, no. 3, 2013.

[34] V. Bajpai and J. Schönwalder, "A survey on internet performance measurement platforms and related standardization efforts," *Communications Surveys & Tutorials, IEEE*, vol. 17, no. 3, pp. 1313–1341, 2015.

[35] R. Hofstede, P. Celeda, B. Trammell, I. Drago, R. Sadre, A. Sperotto, and A. Pras, "Flow monitoring explained: From packet capture to data analysis with NetFlow and IPFIX," *Communications Surveys & Tutorials, IEEE*, vol. 16, no. 4, pp. 2037–2064, 2014.

[36] H. Feng-Hui, Z. Wen-An, and D. Yu, "QoE Issues of OTT Services over 5G Network," in *2014 Ninth International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA)*. IEEE, 2014, pp. 267–273.

[37] M. Shahid, A. Rossholm, B. Lövström, and H.-J. Zepernick, “No-reference image and video quality assessment: a classification and review of recent approaches,” *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, p. 40, 2014. [Online]. Available: http://jivp.eurasipjournals.com/content/2014/1/40/abstract

[38] A. Aaron, Z. Li, M. Manohara, J. Y. Lin, E. C.-H. Wu, and C.-C. J. Kuo, “Challenges in cloud based ingest and encoding for high quality streaming media,” in *2015 IEEE International Conference on Image Processing (ICIP)*, vol. 2015-Decem. IEEE, sep 2015, pp. 1732–1736. [Online]. Available: http://ieeexplore.ieee.org/document/7351097/

[39] U. Reiter, K. Brunnström, K. De Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank, “Factors influencing Quality of Experience,” in *Quality of Experience: Advanced Concepts, Applications and Methods*. Springer, 2014, pp. 55—-72.

[40] L. Skorin-Kapov and M. Varela, “A multi-dimensional view of QoE: the ARCU model,” in *MIPRO, 2012 Proceedings of the 35th International Convention*. IEEE, 2012, pp. 662–666.

[41] V. Aggarwal, E. Halepovic, J. Pang, S. Venkataraman, and H. Yan, “Prometheus: Toward quality-of-experience estimation for mobile apps from passive network measurements,” in *Proceedings of the 15th Workshop on Mobile Computing Systems and Applications*. ACM, 2014, p. 18.

[42] A. Ahmad, L. Atzori, and M. G. Martini, “Qualia: A multilayer solution for QoE passive monitoring at the user terminal,” in *IEEE ICC 2017 Communications Software, Services, and Multimedia Applications Symposium (ICC’17 CSSMA)*. Paris, France: IEEE, May 2017.

[43] P. Casas, M. Seufert, and R. Schatz, “YOUQMON: A system for on-line monitoring of youtube QoE in operational 3G networks,” *ACM SIGMETRICS Performance Evaluation Review*, vol. 41, no. 2, pp. 44–46, 2013.

[44] Q. A. Chen, H. Luo, S. Rosen, Z. M. Mao, K. Iyer, J. Hui, K. Sontineni, and K. Lau, “QoE doctor: Diagnosing mobile app QoE with automated ui control and cross-layer analysis,” in *Proceedings of the 2014 Conference on Internet Measurement Conference*. ACM, 2014, pp. 151–164.

[45] F. Wamser, M. Seufert, P. Casas, R. Irmer, P. Tran-Gia, and R. Schatz, “Yomoapp: A tool for analyzing QoE of youtube HTTP adaptive streaming in mobile networks,” in *Networks and Communications (EuCNC), 2015 European Conference on*. IEEE, 2015, pp. 239–243.

[46] A. Perkis, P. Reichl, and S. Beker, *Quality of Experience - Advanced Concepts, Applications and Methods*. Springer, 2014, ch. Business Perspectives on Quality of Experience, pp. 97–108.

[47] R. Schatz, T. Hoßfeld, and P. Casas, “Passive YouTube QoE Monitoring for ISPs,” in *2012 Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)*. IEEE, 2012, pp. 358–364.

[48] I. Orsolic, D. Pevec, M. Suznjevic, and L. Skorin-Kapov, "YouTube QoE Estimation Based on the Analysis of Encrypted Network Traffic Using Machine Learning," in *Fifth IEEE International Workshop on Quality of Experience for Multimedia Communications (QoEMC)*, Washington, 2016.

[49] G. Dimopoulos, I. Leontiadis, P. Barlet-Ros, and K. Papagiannaki, "Measuring Video QoE from Encrypted Traffic," in *Proceedings of the 2016 ACM on Internet Measurement Conference - IMC '16*. New York, New York, USA: ACM Press, 2016, pp. 513–526. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2987443.2987459

[50] C.-L. Yeh, "Conceptualized framework for regulation of OTT video services: A new battlefield of interconnection and peering," in *20th ITS Biennial Conference, Rio de Janeiro 2014: The Net and the Internet-Emerging Markets and Policies*. International Telecommunications Society (ITS), 2014.

[51] A. Passarella, "A survey on content-centric technologies for the current Internet: CDN and P2P solutions," *Computer Communications*, vol. 35, no. 1, pp. 1–32, 2012.

[52] Google. (2017) Google's edge network. [Online]. Available: https://peering.google.com/

[53] Netflix. (2017) Netflix open connect. [Online]. Available: https://openconnect.netflix.com

[54] B. Frank, I. Poese, Y. Lin, G. Smaragdakis, A. Feldmann, B. Maggs, J. Rake, S. Uhlig, and R. Weber, "Pushing CDN-ISP collaboration to the limit," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 3, p. 34, jul 2013. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2500098.2500103

[55] M. Wichtlhuber, R. Reinecke, and D. Hausheer, "An SDN-based CDN/ISP collaboration architecture for managing high-volume flows," *IEEE Transactions on Network and Service Management*, vol. 12, no. 1, pp. 48–60, 2015.

[56] "PeeringDB API," (Date last accessed 23-March-2017). [Online]. Available: https://peeringdb.com/apidocs/

[57] M. Kasbekar and V. Kanitkar, "CDNs, Network Services and Encrypted Traffic," in *Managing Radio Networks in an Encrypted World (MaRNEW) Workshop*. IAB, 2015.

[58] E. Thomas, M. van Deventer, T. Stockhammer, A. C. Begen, and J. Famaey, "Enhancing MPEG DASH performance via server and network assistance," 2015.

[59] H. Habibi Gharakheili, A. Vishwanath, and V. Sivaraman, "Perspectives on Net Neutrality and Internet Fast-Lanes," *ACM SIGCOMM Computer Communication Review*, vol. 46, no. 1, pp. 64–69, 2016.

[60] C. Marsden, *Network Neutrality: From Policy to Law to Regulation*. UK: Manchester University Press, 2017.

[61] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, "Network function virtualization: State-of-the-art and research challenges," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 236–262, 2016.

[62] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90–97, 2015.

[63] A. Moreira, J. Moreira, D. Sadok, A. Callado, M. Rodrigues, M. Neves, V. Souza, and P. P. Karlsson, "A case for virtualization of content delivery networks," in *Proceedings of the Winter Simulation Conference*, 2011, pp. 3183–3194.

[64] D. P. Francesco, G. G. Claudio, M. Salvatore, and S. Vincenzo, "Future internet architecture: Control-based perspectives related to quality of experience (qoe) management," in *34th Chinese Control Conference*. IEEE, 2016.

[65] T. Carey and R. Pattabhiraman, "Management of Virtual CPES," 2016, uS Patent 20160020962.

[66] A. Manzalini, R. Minerva, F. Callegati, W. Cerroni, and A. Campi, "Clouds of virtual machines in edge networks," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 63–70, 2013.

[67] D. Kreutz, F. M. Ramos, P. Esteves Verissimo, C. Esteve Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: A comprehensive survey," *Proceedings of the IEEE*, vol. 103, no. 1, pp. 14–76, 2015.

[68] R. Bruschi, P. Lago, G. Lamanna, C. Lombardo, and M. S., "Openvolcano: An open-source software platform for fog computing," in *2016 International Teletraffic Congress*. IEEE, 2016.

[69] Cisco, "VNI Forecast and Methodology, 2014 – 2019," Tech. Rep., 2015.

[70] B.-H. Chu, M.-S. Tsai, and C.-S. Ho, "Toward a hybrid data mining model for customer retention," *Knowledge-Based Systems*, vol. 20, no. 8, pp. 703–718, 2007.

[71] A. Floris, L. Atzori, and G. Ginesu, "Addressing un-interoperability issues in qoe models: Is a layered modelling effective?" in *Communications Workshops (ICC), 2014 IEEE International Conference on*. IEEE, 2014, pp. 563–568.

[72] A. Ahmad, A. Floris, and L. Atzori, "QoE-aware Service Delivery: A Joint-Venture Approach for Content and Network Providers," in *2016 Eighth International Workshop on Quality of Multimedia Experience (QoMEX)*, Lisbon, 2016.

[73] F. Wamser, M. Seufert, P. Casas, R. Irmer, P. Tran-Gia, and R. Schatz, "YoMoApp: A tool for analyzing QoE of YouTube HTTP adaptive streaming in mobile networks," *2015 European Conference on Networks and Communications, EuCNC 2015*, pp. 239–243, 2015.

[74] I. Brandusoiu and G. Toderean, "Churn prediction in the telecommunications sector using support vector machines," *Annals of the Oradea University*, no. 1, pp. 19–22, 2013. [Online]. Available: http://imtuoradea.ro/auo.fmte/files-2013-v1/BrandusoiuIonut1.pdf

[75] M. Butler, “Android: Changing the mobile landscape,” *IEEE Pervasive Computing*, vol. 10, no. 1, pp. 4–7, 2011.

[76] F. Wamser, P. Casas, M. Seufert, C. Moldovan, P. Tran-Gia, and T. Hossfeld, “Modeling the YouTube stack: From packets to quality of experience,” *Computer Networks*, 2016.

[77] R. K. Mok, E. W. Chan, and R. K. Chang, “Measuring the quality of experience of HTTP video streaming,” in *12th IFIP/IEEE International Symposium on Integrated Network Management (IM 2011) and Workshops*. IEEE, 2011, pp. 485–492.

[78] “Cisco visual networking index: Global mobile data traffic forecast update, 2013–2018,” *white paper*, 2014.

[79] D. Tsolkas, E. Liotou, N. Passas, and L. Merakos, “A survey on parametric QoE estimation for popular services,” *Journal of Network and Computer Applications*, vol. 77, no. 1, pp. 1–17, 2017.

[80] P. Ameigeiras, J. J. Ramos-Munoz, J. Navarro-Ortiz, P. Mogensen, and J. M. Lopez-Soler, “Qoe oriented cross-layer design of a resource allocation algorithm in beyond 3g systems,” *Computer Communications*, vol. 33, no. 5, pp. 571–582, 2010.

[81] T. Hoßfeld, R. Schatz, E. Biersack, and L. Plissonneau, “Internet Video Delivery in YouTube: From Traffic Measurements to Quality of Experience,” in *Data Traffic Monitoring and Analysis*, E. Biersack, C. Callegari, and M. Matijasevic, Eds. Springer Berlin Heidelberg, 2013, pp. 264–301.

[82] “Qualinet white paper on definitions of quality of experience,” 2013.

[83] M. Hassan, L. Atzori, and M. Krunz, “Video transport over wireless channels: a cycle-based approach for rate control,” in *International conference on Multimedia, 2004*. ACM, 2004, pp. 916–923.

[84] M. Varela, P. Zwickl, P. Reichl, M. Xie, and H. Schulzrinne, “From Service Level Agreements (SLA) to Experience Level Agreements (ELA): The challenges of selling QoE to the user,” in *2015 IEEE International Conference on Communication Workshop (ICCW)*. IEEE, 2015, pp. 1741–1746.

[85] W. Dai, J. W. Baek, and S. Jordan, “Modeling the impact of QoS pricing on ISP integrated services and OTT services,” in *11th Int. Conf. on Network and Service Management (CNSM)*. IEEE, 2015, pp. 85–91.

[86] G. Ibrahim and D. Kofman, “Improving content delivery through coalitions,” in *2014 Network Operations and Management Symposium (NOMS)*. IEEE, 2014, pp. 1–9.

[87] N. Shetty, G. Schwartz, and J. Walrand, “Internet QoS and regulations,” *IEEE/ACM Transactions on Networking*, vol. 18, no. 6, pp. 1725–1737, 2010.

[88] S. Wahlmueller, P. Zwickl, and P. Reichl, “Pricing and regulating Quality of Experience,” in *2012 8th EURO-NGI Conference on Next Generation Internet (NGI)*. IEEE, 2012, pp. 57–64.

[89] A. Odlyzko, “Paris metro pricing for the internet,” in *Proceedings of the 1st ACM conf. on Electronic commerce*. ACM, 1999, pp. 140–147.

[90] Ramneek, P. Hosein, W. Choi, and W. Seok, “Disruptive network applications and their impact on network neutrality,” in *2015 17th International Conference on Advanced Communication Technology (ICACT)*. IEEE, 2015, pp. 663–668.

[91] H. Habibi Gharakheili, A. Vishwanath, and V. Sivaraman, “Perspectives on net neutrality and internet fast-lanes,” *ACM SIGCOMM Computer Communication Review*, vol. 46, no. 1, pp. 64–69, 2016.

[92] S. Ellouze, B. Mathieu, T. Ahmed, and N. Schwan, “Collaboration between Networks and Applications in the Future Internet,” in *Media Networks - Architectures, Applications, and Standards*, H. Moustafa and S. Zeadally, Eds. CRC Press, 2012, pp. 121–140.

[93] “Application-layer traffic optimization (ALTO) protocol,” 2014, iETF RFC 7285.

[94] “Co-optimisation of overlay applications and underlying networks,” 2012, eNVISION FP7 Project.

[95] I. Abdeljaouad and A. Karmouch, “Monitoring IPTV Quality of Experience in overlay networks using utility functions,” *Journal of Network and Computer Applications*, vol. 54, pp. 1–10, 2015.

[96] M. Fiedler, T. Hossfeld, and P. Tran-Gia, “A generic quantitative relationship between Quality of Experience and Quality of Service,” *IEEE Network*, vol. 24, no. 2, pp. 36–41, 2010.

[97] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, “Smart Data Pricing (SDP): Economic solutions to network congestion,” in *Recent Advances in Networking*, H. Haddadi and O. Bonaventure, Eds., 2013, pp. 221–274.

[98] M. Leppaniemi and H. Karjaluoto, “Factors influencing consumers’ willingness to accept mobile advertising: a conceptual model,” *International Journal of Mobile Communications*, vol. 3, no. 3, pp. 197–213, 2005.

[99] P. Srinuan, C. Srinuan, and E. Bohlin, “An empirical analysis of multiple services and choices of consumer in the swedish telecommunications market,” *Telecommunications Policy*, vol. 38, no. 5, pp. 449–459, 2014.

[100] N. J. Nilsson, *Artificial intelligence: a new synthesis*. Morgan Kaufmann, 1998.

[101] B. Karlik and A. V. Olgac, “Performance analysis of various activation functions in generalized MLP architectures of neural networks,” *International Journal of Artificial Intelligence and Expert Systems*, vol. 1, no. 4, pp. 111–122, 2011.

[102] Q.-T. Nguyen-Vuong, Y. Ghamri-Doudane, and N. Agoulmine, “On utility models for access network selection in wireless heterogeneous networks,” in *NOMS 2008, IEEE Network Operations and Management Symposium*. IEEE, 2008, pp. 144–151.

[103] K. Yamagishi and T. Hayashi, "Parametric packet-layer model for monitoring video quality of IPTV services," in *ICC'08. IEEE International Conference on Communications, 2008*. IEEE, 2008, pp. 110–114.

[104] "The E-model: a computational model for use in transmission planning." Recommendation ITU-T G.107, 2015.

[105] J. Joskowicz and J. Carlo Lopez Ardao, "Enhancements to the opinion model for video-telephony applications," in *LANC '09 Proceedings of the 5th International Latin American Networking Conference*. ACM, 2009, pp. 87–94.

[106] R. G. Cole and J. H. Rosenbluth, "Voice over IP performance monitoring," *ACM SIGCOMM Computer Communication Review*, vol. 31, no. 2, pp. 9–24, 2001.

[107] "Voip - per call bandwidth," https://www.cisco.com/c/en/us/support/docs/voice/voice-quality/7934-bwidth-consume.html.

[108] "Youtube live encoder settings, bitrates and resolutions," https://support.google.com/youtube/answer/2853702?hl=en.

[109] Nokia. (2017) Nokia White Paper, Quality of Experience (QoE) of mobile services: Can it be measured and improved? [Online]. Available: http://afutt.org/Qostic/qostic1/MOB-GD-MGQ-NOKIA-040129-Nokia-whitepaper_qoe_net-final.pdf

[110] Accenture. (2017) Customer service, not price, remains top cause of customer churn, accenture study finds. [Online]. Available: https://newsroom.accenture.com/subjects/customer-relationship-management/customer-service-not-price-remains-top-cause-customer-churn-accenture-study-finds.htm

[111] J. C. Hoekstra and E. K. R. E. Huizingh, "The Lifetime Value Concept in Customer-Based Marketing," *Journal of Market-Focused Management*, vol. 3, no. 3, pp. 257–274, 1999.

[112] N. Glady, B. Baesens, and C. Croux, "Modeling churn using customer lifetime value," *European Journal of Operational Research*, vol. 197, no. 1, pp. 402–411, 2009.

[113] Microsoft. (2017) Microsoft peering. [Online]. Available: https://www.microsoft.com/Peering

[114] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2016-2021 - White paper," 2017.

[115] E. Liotou, G. Tseliou, K. Samdanis, D. Tsolkas, F. Adelantado, and C. Verikoukis, "An SDN QoE-service for dynamically enhancing the performance of OTT applications," in *Seventh Int. Workshop on Quality of Multimedia Experience (QoMEX), 2015*. IEEE, 2015.

[116] A. Bentaleb, A. C. Begen, R. Zimmermann, and S. Harous, "SDNHAS: An SDN-Enabled Architecture to Optimize QoE in HTTP Adaptive Streaming," *IEEE Transactions on Multimedia*, vol. 19, no. 10, pp. 2136–2151, 2017.

[117] A. Farshad, P. Georgopoulos, M. Broadbent, M. Mu, and N. Race, "Leveraging SDN to Provide an In-network QoE Measurement Framework," in *IEEE Conf. on Computer Communications Workshops (INFOCOM WKSHPS), 2015*. IEEE, 2015, pp. 239–244.

[118] S. Khairi, M. Bellafkih, and B. Raouyane, "QoS management SDN-based for LTE/EPC with QoE evaluation: IMS use case," in *Fourth Int. Conf. on Software Defined Systems (SDS), 2017*. IEEE, 2017, pp. 125–130.

[119] S. Ramakrishnan, X. Zhu, F. Chan, and K. Kambhatla, "SDN Based QoE Optimization for HTTP-Based Adaptive Video Streaming," in *IEEE Int. Symposium on Multimedia (ISM), 2015*. IEEE, 2015, pp. 120–123.

[120] H. Nam, K.-H. Kim, J. Y. Kim, and H. Schulzrinne, "Towards QoE-aware video streaming using SDN," in *IEEE Global Communications Conference (GLOBECOM), 2014*. IEEE, 2014, pp. 1317–1322.

[121] E. Grigoriou, A. A. Barakabitze, L. Atzori, L. Sun, and V. Pilloni, "An SDN-approach for QoE management of multimedia services using resource allocation," in *IEEE Int. Conf. on Communications (ICC), 2017*. IEEE, 2017.

[122] M. García-Pineda, S. Felici-Castell, and J. Segura-García, "Adaptive SDN-based architecture using QoE metrics in live video streaming on Cloud Mobile Media," in *Fourth Int. Conf. on Software Defined Systems (SDS), 2017*. IEEE, 2017, pp. 100–105.

[123] L. Fawcett, M. Mu, M. Broadbent, N. Hart, and N. Race, "SDQ: Enabling rapid QoE experimentation using Software Defined Networking," in *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, 2017, pp. 656–659.

[124] F. Wamser, M. Seufert, P. Casas, R. Irmer, P. Tran-Gia, and R. Schatz, "Yomoapp: A tool for analyzing qoe of youtube http adaptive streaming in mobile networks," in *2015 European Conference on Networks and Communications (EuCNC)*. IEEE, 2015, pp. 239–243.

[125] T. Hoßfeld, C. Moldovan, and C. Schwartz, "To each according to his needs: Dimensioning video buffer for specific user profiles and behavior," in *IFIP/IEEE Int. Symposium on Integrated Network Management (IM), 2015*. IEEE, 2015, pp. 1249–1254.

[126] Cisco. (2017) Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021 white paper. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[127] P. Le Callet, S. Möller, and A. Perkis. (2012) Qualinet White Paper on Definitions of Quality of Experience (2012). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, Version 1.2, March 2013.

[128] M. Torres Vega, D. C. Mocanu, J. Famaey, S. Stavrou, and A. Liotta, "Deep Learning for Quality Assessment in Live Video Streaming," *IEEE Signal Processing Letters*, vol. 24, no. 6, pp. 736–740, 2017.

[129] M. Torres Vega, D. C. Mocanu, S. Stavrou, and A. Liotta, "Predictive no-reference assessment of video quality," *Signal Processing: Image Communication*, vol. 52, pp. 20–32, 2017.

[130] M. Torres Vega, D. C. Mocanu, and A. Liotta, "Unsupervised deep learning for real-time assessment of video streaming services," *Multimedia Tools and Applications*, vol. 76, no. 21, pp. 22 303–22 327, 2017.

[131] S. Schwarzmann, T. Zinner, and O. Dobrijevic, "Towards a Framework for Comparing Application-Network Interaction Mechanisms," in *28th Int.l Teletraffic Congress (ITC 28), 2016*. IEEE, 2016, pp. 13–18.

[132] T. Hoßfeld, L. Skorin-Kapov, P. E. Heegaard, and M. Varela, "Definition of QoE Fairness in Shared Systems," *IEEE Communications Letters*, vol. 21, no. 1, pp. 184–187, 2017.

[133] D. I. Forum. (2016) DASH-IF Position Paper: Server and Network Assisted DASH (SAND).

[134] StreamingMediaBlog.com. (2017) Chart shows which content owners have direct interconnect deals with isps. [Online]. Available: http://blog.streamingmedia.com/2014/05/chart-shows-which-content-owners-have-direct-interconnect-deals-with-isps.html

[135] FierceWireless. (2017) iphone users cite video buffering as biggest annoyance. [Online]. Available: http://www.fiercewireless.com/europe/iphone-users-cite-video-buffering-as-biggest-annoyance

[136] P. Reichl, S. Egger, S. Möller, K. Kilkki, M. Fiedler, T. Hoßfeld, C. Tsiaras, and A. Asrese, "Towards a comprehensive framework for qoe and user behavior modelling," in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*. IEEE, 2015, pp. 1–6.

[137] D. Bernstein, E. Ludvigson, K. Sankar, S. Diamond, and M. Morrow, "Blueprint for the Intercloud - Protocols and Formats for Cloud Computing Interoperability," in *Proc. of the Fourth Int. Conf. on Internet and Web Applications and Services, 2009*. IEEE, 2009, pp. 328–336.

[138] A. Stanik and O. Kao, "A proposal for REST with XMPP as base protocol for intercloud communication," in *Proc. of the 7th Int. Conf. on Information, Intelligence, Systems & Applications (IISA), 2016*. IEEE, 2016.

[139] T. Hoßfeld, P. E. Heegaard, L. Skorin-Kapov, and M. Varela, "No silver bullet: QoE metrics, QoE fairness, and user diversity in the context of QoE management," in *Proc. of the Ninth Int. Conf. on Quality of Multimedia Experience (QoMEX), 2017*. IEEE, 2017.

[140] L. Sun and E. Ifeachor, "New models for perceived voice quality prediction and their applications in playout buffer optimization for VoIP networks," in *IEEE Int. Conf. on Communications, 2004*, vol. 3. IEEE, 2004, pp. 1478–1483.