A Novel Monitoring Architecture for Media Services Adaptation Based on Network QoS to Perceived QoS Mapping

M. Sidibé[1], H. Koumaras[2], I. Kofler[3], A. Mehaoua[1, 4], A. Kourtis[2], C. Timmerer[3]

[1]*University of Versailles, PRiSM Lab. 45, Av. des Etats Unis, 78035 Versailles, France*

Phone : +33 1 39 25 43 27

Fax : + 33 9 59 10 42 27

Email : {mas, mea}@prism.uvsq.fr

[2]*National Center for Scientific Research 'Demokritos', Institute of Informatics & Telecommunications, 15310 Aghia Paraskevi Attikis, POB 60228, Athens, Greece*

Phone : + 30 210 7253 783

Fax : +30 210 65 32 175

Email : {koumaras, kourtis}@iit.demokritos.gr

[3]*Klagenfurt University, Universitätsstrasse 65-67, 9020 Klagenfurt, Austria*

Phone : + 43 463 2700 3613

Fax : + 43 463 2700 3699

Email : {ingo.kofler, christian.timmerer} @itec.uni-klu.ac.at

[4]*Université Paris Descartes, CRIP5 lab., 45 rue des saints pères 75006 Paris France*

Phone : +33 1 44 55 35 45

Fax : +33 1 44 55 35 35

Email : mea@math-info.univ-paris5.fr

Abstract— One of the future visions of multimedia networking is the provision of multimedia content at a variety of quality and price levels. Of the many approaches to this issue, one of the most predominant techniques is the concept of Perceived Quality of Service (PQoS), which extends the traditional engineering-based QoS concept to the perceptual satisfaction that the user receives from the reception of multimedia content. In this context, PQoS monitoring is becoming crucial to media service providers (SPs) for providing not only quantified PQoS-based services, but also service assurance based on multimedia content adaptation across heterogeneous networks. This work proposes a novel cross-layer monitoring architecture that utilizes a new Network QoS (NQoS) to PQoS mapping framework at the application level. The resulting QoS monitoring should allow the content delivery system to take sophisticated actions for real time media content adaptation, and aims to provide perceived service performance verification with respect to the QoS guarantees that have been specified in contractual agreements between providers and end-users. A subsequent performance evaluation of the proposed model conducted using a real test-bed environment demonstrates both the accuracy and feasibility of the network level measurements, the NQoS to PQoS mapping and the overall feasibility of the proposed end-to-end monitoring solution.

Keywords*: Perceived quality monitoring; End-to-end monitoring; service adaptation; NQoS/PQoS; SLA/SLS; integrated management.*

1 Introduction

Both the rapidly increasing amount of multimedia content that is offered on the Internet and the heterogeneity of the underlying networking technologies demand the development of new QoS-enabled mechanisms and architectures to efficiently control, manage and monitor the respective network resources. In this context, the end-to-end management of services, underlying networks, contents and terminals is required. Such end-to-end management needs the knowledge of dynamic continuous network conditions information as input for the management entities that are to take the appropriate actions for service assurance (adaptation).

Providing this dynamic network conditions information demands a monitoring solution that covers network-wide and service end-to-end scope. This paper describes and evaluates a service-oriented monitoring system designed for deployment in a multi-domain, heterogeneous networking environment, supporting real time cross-layer media service adaptation. More specifically, the proposed system should:

a. Assist service/network providers in verifying whether the QoS performance-guarantees committed in Service Level Agreements (SLAs) are actually satisfied. In the case of service degradation, the monitoring system provides information for taking remedial actions and for performing an end-to-end cross layer adaptation;

b. Assist network providers in optimizing the usage of the available network resources and avoiding undesirable conditions.

In this paper, we assume that the performance and traffic requirements of a requested service are described by an SLA and consequently its SLS (Service Level Specification) [1]. Both the SLA and SLS are basic elements in the operation of our proposed QoS-based monitoring system. SLAs provide the means to officially contract service level negotiations conducted between a customer (or Content Consumer, CC) and a service provider for a specific class of service (e.g., Golden, Silver or Bronze classes defined in this work). An SLS is a subset of an SLA that denotes the technical characteristics of a service offered. These service-related technical characteristics refer to the provisioning aspects of the service, such as request, activation and delivery aspects from the network perspective. In this section, two types of SLSs (and consequently of SLAs) are examined: customer-to-provider SLSs (cSLSs), and provider-to-provider SLSs (pSLSs) (and cSLAs and pSLAs respectively) [2]. Also, in this work, a cSLA/cSLS is established between a particular CC and a given service provider. A pSLS on the other hand is established between the service and network providers or between network providers [3] and is an agreement between providers for exchanging traffic with the purpose of expanding the geographical span of their offered services. Additionally, pSLSs are meant to support aggregate traffic (i.e. serving many customers), and it is assumed that they are already in place prior to any cSLS agreements with end-customers (Long Range Dependent Agreement). On the other hand cSLSs can differ significantly depending on the type of services offered (Short Range Dependent Agreement) and consequently can have different QoS requirements.

The monitoring architecture presented here was developed for the purpose of the IST ENTHRONE II project[1] that is a service-oriented project with the aim of constructing a complete architectural solution for multimedia content provision, assuring end-to-end QoS management in terms of the performance targets of the user, the application, the terminal and the network. The ENTHRONE II framework [4], [5] provides QoS-enabled media access services by creating, offering, transporting and delivering the underlying content. These access services are based on cooperation of several business actors (Content Providers - CPs, Service Providers - SPs, Network Providers – NPs and CCs) over heterogeneous multi-domain networking environment. A central concept is the ENTHRONE Integrated Management Supervisor (EIMS) system [6] which provides a number of functional components to each actor for managing the end-to-end service delivery. The EIMS consists of a set of functionality including - among others – an improved dynamic service management (policy based), MPEG-21-based cross-layer QoS adaptation, metadata management and an enhanced monitoring system. In principle, a specific sub-system – called manager – is responsible for each functionality. The EIMS Service Manager (EIMS-SM) located at the SP (EIMS-SM@SP) deal with the customer subscriptions (cSLAs), contracts with NPs through pSLSs, the services provided by the SP and the chosen access to the service. In contrast, the EIMS-SM at the NPs (EIMS-SM@NP) deals with pSLSs. In the event of service disruption, the proposed monitoring system provides input to cross-layer QoS Adaptation Manager (EIMS-AM) for content adaptation.

[1] This work is partially funded by European Commission (ENTHRONE EU project IST 507637). See www.enthrone.org. The authors would like to thank ENTHRONE EU project partners for their inputs and valuable discussions.

Following this introductory section, the rest of the paper is organized as follows: Section 2 discusses related work that can be found in the literature. Section 3 presents the new application level PQoS-aware cross-layer monitoring framework architecture. The NQoS to PQoS mapping approach is detailed in Section 4. Section 5 discusses the most relevant service level monitoring aspects. The cross-layer adaptation architecture is presented in Section 6. Section 7 then describes the experimental test-bed environment used to evaluate and validate our monitoring solution. Finally, conclusions are provided in Section 8.

2 Related Work

The Internet Engineering Task Force (IETF) has a number of working groups related to measurements and monitoring such as Remote MONitoring (RMON), IP Performance Metrics (IPPM), Real-Time Flow Measurement (RTFM), IP Flow Information Export (IPFIX), and Packet Sampling (PSAMP). These working groups are in the process of defining metrics, developing a common IP traffic flow measurement technology and specifying a standard set of capabilities for sampling packets through statistical and other methods respectively. In addition, numerous monitoring tools, such as the RIPE Test Traffic Measurement (TTM), NetFlow, SFlow, NIMI (National Internet Measurement Infrastructure) [7], Network Analysis Infrastructure (NAI), cflowd, RTG high-performance SNMP statistics monitoring system, Sskitter, NeTraMet, CoralReef, and Beluga of CAIDA (Cooperative Association for Internet Data Analysis), have been created. Sources [8] and [9] provide more detailed references for these activities. These measurement tools and systems collect, analyze and visualize forms of Internet or Intranet traffic data such as network topology, traffic load, performance, and routing.

There are also several other European research projects active in the field. For example, the objective of the IST-INTERMON project has been to develop an integrated inter-domain QoS monitoring, analysis and modelling system to be used in multi-domain Internet infrastructure for the purpose of planning, operational control and optimization [10]. The proposed solution assumes that a centralized manager negotiates monitoring operations with each domain along the service delivery path. Unfortunately, this results in a scalability problem for the INTERMON system as the inter-domain network expands. Thus, the focus of the IST-MoMe project [10] has been the enhancement of inter-domain real-time QoS architectures with integrated monitoring and measurement capabilities. In contrast, the objective of the IST-SCAMPI project was to develop an open and extensible network monitoring architecture for the Internet including a passive monitoring adapter at 10 Gbps speeds, and other measurement tools to be used for denial-of-service detection, SLS auditing, quality-of-service, traffic engineering, traffic analysis, billing and accounting [10]. IST-LOBSTER, its follow-up project, aimed at deploying an advanced pilot European Internet Traffic Monitoring Infrastructure based on passive monitoring sensors at speeds starting from 2.5Gbps and possibly up to 10Gbps [10]. Finally, the IST-AQUILA project is developing inter-domain QoS-metrics measurement mechanisms, based on the BGRP proposal, to enable measurement based admission control (MBAC) in large-scale IP environments [8], [10].

Our work differs from the other mentioned related works in that:

a. Its end-to-end scope and business model encompasses Content Providers (CPs), Service Providers (SPs), Network Providers (NPs) and Content Customers (CCs);
b. End-to-end service monitoring is tackled using an overlay network of service-level monitoring components communicating in a cascading fashion;
c. Network-specific measurements are collected and translated into a standard-compliant XML-based format.
d. Quality Meters [11] (PQoS probes) at the user-side measure the perceived quality level (Delivered PQoS) of an audio-visual stream.
e. Perceived quality (Derived PQoS) is assessed from measured network performances (Measured NQoS) in access/core networks. This NQoS to PQoS mapping is achieved at the application layer by using a dedicated monitoring component. Note that a core network is a backbone network that provides any-to-any connections among devices on the network. It may consist of several autonomous networking domains (Autonomous Systems, AS) managed by the NPs, while access networks (e.g., wireless local area networks) are used by the CCs to physically connect to a NP for consuming services provided by a SP.

Regarding the mapping of the network QoS sensitive parameters (delay, packet loss etc.) to perceived video quality, S. Kanumuri et al constructed a very analytical statistical model of the visual impact of packet-loss on the quality of decoded MPEG-2 video sequences [12], specifying the various factors that affect perceived video quality and visibility (e.g. Maximum number of frames affected by packet loss, which frame types are subject to the packet loss etc). Similarly, in [13] another transmission/distortion model for real-time video streaming over error-prone wireless

networks is presented. In this work, we choose to model impulse transmission distortion (i.e. the visual fading behavior of the transmission errors), as the previous models are very codec and content specific, while also not providing any end-to-end video quality estimation, namely of the degradation during the encoding process and the transmission/streaming procedure.

Our work proposes a generic model for mapping NQoS Sensitive Parameters, such as packet loss, to PQoS, ensuring consistency in the level of service offered by SPs.

3 End-to-End Monitoring System Architecture

In this section, we present and describe the proposed layered end-to-end QoS monitoring architecture, consisting of four distinct monitoring components and two signaling protocols. These monitoring components are located on different levels and are therefore defined as the Node, Network, Application and Service level monitors, respectively. EQoS-RM and EMon [14], [15] are the signaling protocols for exchanging monitoring data at both the inter- and intra-domain level. This new enhanced monitoring architecture is depicted in Figure 1. . Consequently, for efficiency and scalability reasons, the monitoring management architecture is structured into four levels: service-, application-, network- and node-monitoring levels, which are analytically described in the following subsections.

3.1 Monitoring at Node Level

Monitoring agents at node level are referred to as *Node level Monitors (NodeMons)*. In typical systems, two types of low-level measurements may be performed: *active* and *passive*. The active measurement processes inject synthetic traffic into networks based on scheduled sampling in order to observe network performance between two measurement end-points (clock synchronization is required), while passive measurements are used to simply observe actual data traffic transmitted through the network. In our work, NodeMons are used to perform active traffic measurements between any two edge nodes of an AS and to collect passive measurement information. They are configured with information about the NQoS metrics to be monitored, as well as the sampling and summarisation periods.

NodeMons include probes for effective measurements tasks. Thus, a distinction must be made between NQoS Probes in access/core networks and the PQoS Probes located at terminals. In fact, PQoS Probes perform per-flow measurements, providing effective viewer-perceived quality (Delivered PQoS) while NQoS Probes perform per-aggregate measures, providing input for PQoS assessment in core/access network (Derived PQoS). The PQoS assessment is presented more in detail in Section 4.

3.2 Monitoring at Network Level

The *Network level Monitor (NetMon)* is responsible for intra-domain monitoring that utilizes network-wide traffic measurements collected by all underlying NodeMons in order to build a physical and logical network view (i.e. the view of the routes that have been established over the network). At the network level, all network domain monitoring specifics (e.g., monitoring agent topology, measurements frequency, and so on) are abstracted, so that only those relevant for service agreement monitoring NQoS metrics (loss, delay, jitter, etc.) are reported back to the monitoring component at the application level.

3.3 Monitoring at Application Level

For scalability reasons, NQoS (e.g. packet loss rate) is measured for a specific NQoS class and at an aggregated level. This class of service monitoring is performed in order to anticipate service disruption in the end-to-end delivery chain. However, to improve service disruption anticipation, it might be necessary to also estimate the PQoS experienced by a number of application streams (i.e. cSLSs related flows) that are collectively delivered by a specific NQoS-class. Thus, the new enhanced monitoring architecture, as depicted in Figure 1. , introduces the NQoS to PQoS mapping layer at the application level. In order to support NQoS to PQoS mapping, a component called the *Application level Monitor (AppMon)* is introduced and service level monitoring components are extended with new functionality.

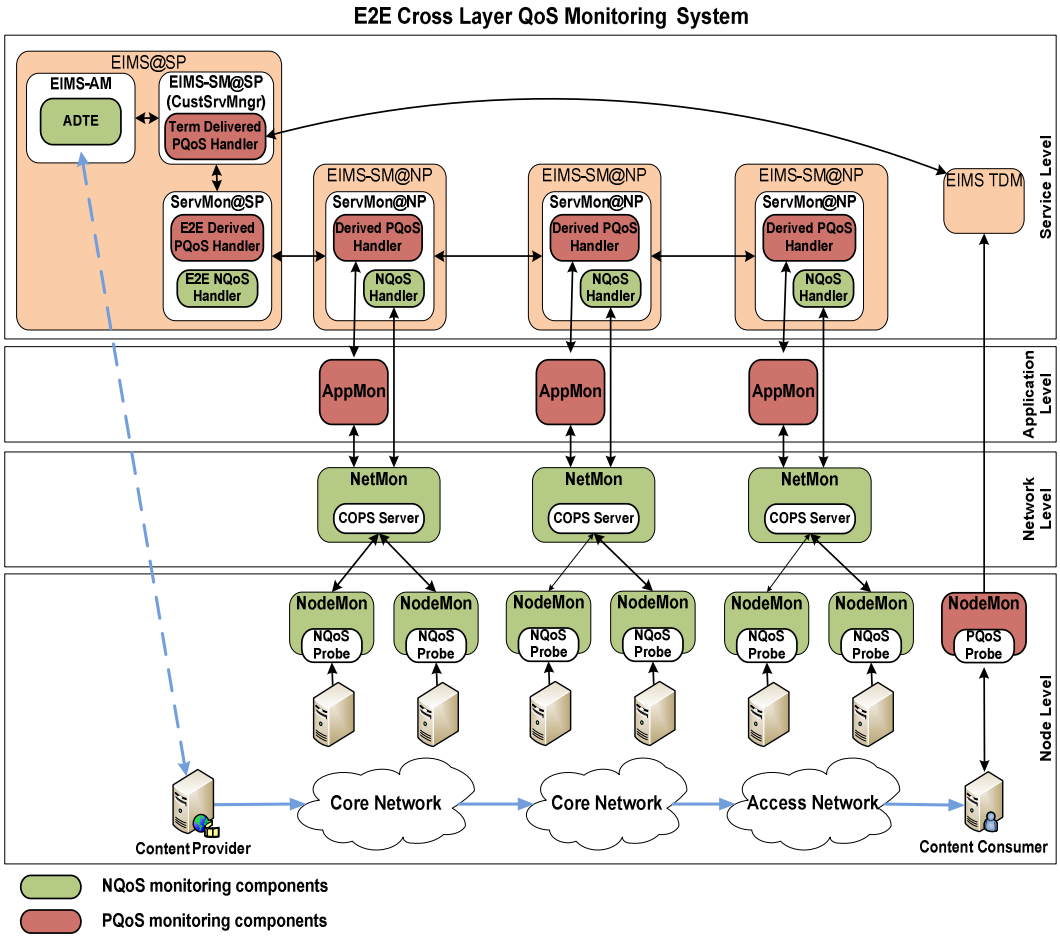

Figure 1. Overall monitoring system architecture

3.4 Monitoring at Service Level

At the service level, the Service level Monitors (ServMons) are dedicated to performing customer/provider-related service level monitoring, auditing and reporting. Thus, they provide in-service verification of value-added services (i.e. Digital Item consumption at several service levels), verifying whether the QoS performance guarantees committed to in the SLSs are being met.

The system presented here also allows for extensive cooperation between different providers while maintaining the authority, confidentiality and full control of each provider over its underlying resources. To achieve this, the monitoring subsystems are deployed as follows. Each NP has its own ServMon (ServMon@NP), AppMon, NetMon and NodeMons as specific provider-related monitoring subsystems while the SP includes a special ServMon (ServMon@SP) and the Customer Service Manager (CustSrvMngr) subsystems. The CustSrvMngr is the customer related service management part of the EIMS-SM@SP. The ServMon@NP is in charge of (1) partial NP related pSLS monitoring; (2) inter-domain QoS reporting on aggregated streams using XML-based measurement statistics; (3) processing partial NP related mapping feedback. On the other hand, the ServMon@SP is in charge of (1) end-to-end pSLS monitoring; (2) coordinating the pSLS related service level monitoring procedures and processing the information provided by the other ServMon entities of the networks involved in the end-to-end chain of QoS delivery; (3) providing end-to-end mapping feedback as input to the EIMS-AM (through the CusSrvMngr) for triggering content adaptation.

4 NQoS To PQoS Mapping

This section discusses the NQoS to PQoS mapping framework that the proposed monitoring module utilizes for providing advanced adaptation of media services.

4.1 Definitions, assumptions and model principles

The perceptual degradation of the initial video content across the content generation, distribution and consumption chain sets novel research perspectives. More specifically, the focus of research has been placed on the impact of each layer involved in the provision process (i.e. Service, Application and Network Layer) on the perceptual quality level of the delivered video service by defining and correlating the various QoS-related metrics of each layer. In this context, this section discusses and describes the theoretical basis of a generic model for the cross-layer mapping of Network QoS Sensitive Parameters, such as packet loss, to the Application Layer and finally to the Service Layer, providing subjectively validated video quality assessment of video services.

At the Service layer the critical metric is the user satisfaction (i.e. Perceived QoS - PQoS), which refers to the QoS that a user perceives during the provision of a multimedia service. For video service, this is directly related to the perceived video quality. The perceptual level of a multimedia service (i.e. the PQoS) is evaluated using specific metrics and methods, which are dependent on the type of the service. Therefore, for the case of video services, the generic concept of PQoS is concerned only with the assessment of video quality. In this framework, the PQoS evaluation gives the service provider and network operator the capability to minimize storage and network resources by allocating only the resources that are necessary to preserve a specific level of user satisfaction. The estimation of the PQoS can be performed either by subjective methods, which involve an audience of people with the task of evaluating the perceived quality of the video service being displayed, or by objective methods, which exploit mathematical models for emulating the respective subjective results.

At the Application layer, given that during the encoding/compression process of the initial video content the quality is degraded by the appearance of specific artifacts such as blockiness, blur, jerkiness etc., the values of the Application QoS (AppQoS) parameters (i.e. bit rate, resolution) determine the ultimate PQoS. Thus, the various encoding parameters must be considered as AppQoS, since they have a straightforward impact on the deduced PQoS level. If additional transmission problems are considered due to limited available bandwidth, network congestion etc., which result in packet loss during the service streaming/transmission, then the packet loss percentage and the expected decodable frame ratio should also be considered as metrics at the AppQoS layer.

Finally, at the Network layer, the Network QoS (NQoS) related metrics (i.e. Packet Loss Ratio, Packet Loss scheme and Packet Size) are used in an objective aspect. Although it is obvious that other NQoS-related phenomena may be present over a transmission network, such as jitter and delay, all these parameters can be expressed in terms of packet losses, since this is the final outcome at the video packet receiver. Otherwise, if no packet loss occurred due to these phenomena, then sophisticated buffer techniques might be able to eliminate their impact and therefore making it unnecessary to take it into consideration.

4.2 Theoretical Layered Mapping Approach

Once the content has been prepared for delivery at the requested PQoS level, the transmission phase follows. Thus, it is necessary to develop mapping rules between the Network Layer parameters (i.e. Packet Loss, Packet Size, Packet Loss Scheme) and the Application Layer (e.g. decodable frame rate). More specifically, with regards to the unidirectional NQoS to AppQoS mapping, this paper considers the correlation of both network packet loss ratio and packet size with the theoretically expected Decodable Frame Rate (Q). Q is an application-level metric, with values ranging from 0 to 1.0. Clearly, yhe larger the value of Q, the higher the successful decoding rate at the end user. Q is defined as the fraction of the decodable frame rate, which is the number of theoretically expected decodable frames from the total number of frames sent by a video source:

$$Q = \frac{N_{dec}}{(N_{total\text{-}I} + N_{total\text{-}P} + N_{total\text{-}B})}$$

Where Ndec is the sum of number of theoretically successfully decoded I, P, B frames, i.e. Ndec-I, Ndec-P, and Ndec-B [16].

Due to the fact that the frames in a MPEG-based encoded video sequence are interdependent, considering a packet loss, spatial error propagation will take place, affecting all the frames that are dependent on the specific frame that in which packet loss occurred. Thus, the impulse transmission of the distortion must be taken into consideration. Due to the very specific structure of an MPEG

stream (i.e. GOP type), which is specified by successive I, P and B frames, given that the loss of one packet deterministically results in the corresponding frame degradation and/or loss (i.e. DT=1.0), then a video frame may be considered theoretically expected undecodable directly or indirectly depending on the position of the packet that was lost during the transmission. That is, the frame of which the lost packet was a part of, is considered as directly theoretically expected undecodable, while the frames that are dependent on the successful decoding of the specific frame are considered as indirectly undecodable. More specifically, given a GOP structure and taking the decoding inter-dependencies between the three frame types into consideration, then the impact of the packet loss ratio can be mathematically and deterministically formulated [17], [16].

Based on the objective metrics that were exploited at the application layer, the extension of the theoretical framework to the service layer requires a mapping between the dropped frames ratio (AppQoS) and its impact on the delivered video quality (PQoS). Based on the relevant literature [18], a video signal may suffer several forms of degradation at any stage of the transmission chain, resulting in severe motion discontinuities in video streaming that the end-user may perceive as jerky motion and instantaneous fluidity breaks. Packet loss (or delay jitter as another aspect of the packet loss effect) in the transmitted networks is the main cause of this perceived jerkiness/break since they cause the discarding of sporadic frames during the decoding process because of the limited buffering time. In [18] the term temporal discontinuity is used as a perceptual synonym of a dropped picture burst. More specifically, the estimated quality function in relation to a single burst of dropped frames for different durations has been estimated for video sequences of 10 sec duration, a period long enough in order to avoid the forgiveness effect and the spatiotemporal variation of the content. Furthermore, this short duration minimizes the probability of multiple burst packet losses within this period, making the study of single packet loss bursts schemes statistically quite accurate and satisfactory.

Based on the model, which has been proposed, presented, tested and evaluated in [18] together with subjectively evaluation processes, the mapping of the dropped frames to a perceptual quality level with regard to mean opinion scores (MOS) over various spatiotemporal content of 10 sec duration, CIF Resolution and 25fps is analytically described by the following expression:

$$PQoS\ Level = 85.8 - \frac{53.03}{1 + (562/x)^{1.01}}$$

Where x is the discontinuity duration computed from all content in msec. This equation maps the AppQoS metric of the discontinuity durations to the respective PQoS level degradation as a percentage based on subjectively collected MOS assessments. Due to the subjective nature of the above equation, for zero percentage of dropped frames, the formula provides a 85.8 score out of 100, which corresponds to an 'Excellent' evaluation description according to the DSCQS (Double Stimulus using a Continuous Quality Scale) method presented in [19] and [20], given that due to the statistical nature of the MOS metric, the absolute excellent value of 100 is never achieved. Considering that the duration of all the tested signals in [18] is 10 sec, then the variable x can be also expressed as a percentage of the overall signal length. By describing the variable x as such, x can be mapped to the percentage of dropped frames from the total number of frames in the 10 sec signal. Moreover, this mapping is 1-to-1 and does not require any further sophisticated adoption. So, the variable x can be substituted with the percentage of the dropped frames (i.e. the complimentary of Q), which has been mathematically modelled in the previous section, allowing the above equation to be formulated as:

$$PQoS\ Level = 85.8 - \frac{53.03}{1 + \left(\frac{562}{(1-Q)10^4}\right)^{1.01}}$$

Thus, the above equation between subjective MOS-based evaluation (i.e. PQoS layer) and discontinuity duration (i.e. AppQoS layer) provides the mapping between the AppQoS and PQoS layer.

5 Service Level Monitoring Approaches

Service level monitoring aims to track the level of end-to-end service provided to customers. To achieve this, ENTHRONE proposes two types of monitoring services: (1) the monitoring of aggregated streams (focus of this paper) within the core and access networks of [15] by the

ServMons; (2) the monitoring of a particular customer's stream by the CustSrvMngr using information collected by a PQoS Meter [11] located at the terminal.

Monitoring within the core and access networks is performed on a per-domain basis by the ServMon@NP and involves periodic active and/or passive measurements of pre-established pSLSs. During this scenario, pSLS QoS performance is continuously monitored, and the retrieved results are made available to the ServMon@SP. Following the successful setup of monitoring components, continuous service monitoring and its measurement reporting are NodeMon-driven using COPS_RM (see [14] and [15]) in push mode as depicted in Figure 2. . Using the specified measurement frequency, the NodeMons regularly send back their measurement reports to the NetMon. The NetMon aggregates the different received measurement reports and forwards them to the upper layer AppMon for PQoS assessment. Then, the results are forwarded to ServMon@NP. This last entity evaluates the degree of satisfaction of service agreements crossing their domain and sends an EQoS monitoring report (EQoS_RM Report) along the return path to the ServMon@SP that initiated the current continuous monitoring procedure.

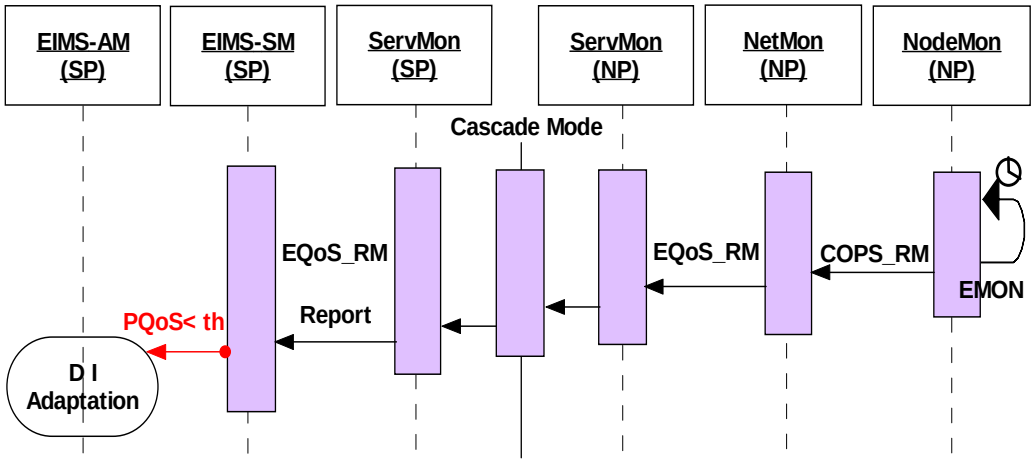

Figure 2. Continuous service monitoring procedure

Within the ServMons, PQoS Handler/E2E PQoS Handler modules (see Figure 1.) support the application level NQoS in its PQoS mapping functions [21]. In fact, a cSLS violation is detected if a derived PQoS falls below a certain threshold, as defined within the related SLA. Following this, a PQoS alert containing network conditions information (related NQoS, such as loss, delay, jitter and bandwidth) is sent to the MPEG-21 based EIMS-AM to initiate the adaptation of the content (MPEG-21 based, see the next section). Also, ServMons periodically provide network conditions information to the EIMS-SM for dynamic service management support, allowing improved dynamic behaviour of Admission Control (AC) algorithms leading to the better utilisation of the network resources.

In order to avoid the failure of data reporting in the above mentioned *continuous* monitoring scenario, the monitoring components can be configured with a timer threshold value, allowing the terminated of a requested continuous monitoring process if no data can be received or if a report cannot be generated. At an SP, this failure triggers the *on-demand* (request-response based) monitoring operation with the aim of locating the domain(s) that is/are the source(s) of end-to-end QoS degradation along service delivery path. This procedure facilitates the verification of the conformance status of each involved AS (network conditions) involved in the service provisioning process, thus allowing the ServMon@SP to periodically retrieve a performance response and then provide input to the EIMS-AM/EIMS-SM.

6 MPEG-21-Based Content Adaptation

The ultimate goal of monitoring the network conditions of video streaming and mapping the conditions to an estimation of a perceived quality is to detect service degradations that mainly result from packet loss over congested network links. In such cases the goal of the service provider is to react to the conditions appropriately and to maintain the provisioning of the audiovisual services for content consumers. Since the service degradation is due to over utilized links, an obvious solution for preventing packet loss is to reduce the required bandwidth by adapting the streamed content.

Adaptation of audiovisual content can be performed in many dimensions which are briefly discussed in the following.

The video part of the content can be adapted in several ways. Spatial adaptation means to change the spatial resolution of each video frame. It is possible to achieve a decrease of spatial resolution by cropping frames to a region of interest (ROI) or by applying spatial resampling techniques. Temporal video adaptation typically reduces the number of frames per second. A decrease in the frame rate usually causes a loss of motion information and, depending on the content, results in jerky motion. However, most video codecs support frame dropping in the compressed domain, allowing frames to be removed from the encoded content which is in general computationally cheap. Another popular type of adaptation is frame quantization which is a codec specific adaptation step in the Signal-to-Noise Ratio (SNR) domain. By modifying the quantization parameters that are used for steering the compression, the bitrate of a video can be adjusted within a certain range. Since most of the video codecs that are in use are based on a Discrete-Cosine Transformation (DCT) that operates on a block of samples, this kind of adaptation can lead to artefacts at the border of each block. As a consequence, the observer perceives a certain degree of blockiness or blurriness in the video frames which increases with the level of quantization. In addition to that of the video, the bitrate of the audio stream can also be adjusted by applying adaptation. Typical methods of audio adaptation include a reduction of the sampling rate, the number of bits per sample, the number of audio channels or a combination thereof. In general the focus of multimedia adaptation is on video since audio typically makes up only a small part of the total bitrate of the content.

Since these adaptation techniques offer a variety of adaptation possibilities there is a need for a control component that selects the most appropriate adaptation parameters for the prevailing network conditions. In the literature the task of finding an optimum parameter set for the adaptation is often referred to as Adaptation Decision Taking (see [22] and [23]) and the corresponding software component that performs this task is called Adaptation Decision Taking Engine (ADTE) [24].

Part 7 of the recently standardized MPEG-21 multimedia framework [25], known as Digital Item Adaptation (DIA), provides the means for tackling this problem [26], [27]. This part of the standard offers three different tools to enable a generic adaptation decision-taking. The notion of a tool within MPEG-21 refers to a description with normative semantics and syntax. As MPEG-21 relies heavily on XML, the syntax for the tools is defined using an XML schema. One of the most important tools is the Usage Environment Description (UED) which is used to represent the usage context of the content consumer. It is used to describe both network and terminal capabilities, preferences and impairments of the content consumer and the natural environment in which the multimedia content is actually consumed. For example, the available video and audio codecs on the terminal, the type of end-device (e.g. PC or PDA), or display and audio playback capabilities can be described. However, the most important part of the UED in the context of our work is the network-related part that is used for describing network capabilities and conditions like delay, jitter, and packet loss. In our approach the monitoring system permanently measures network conditions and generates the XML-based Usage Environment Description on-the-fly. This exchange of normative metadata instead of proprietary monitoring data allows for interoperability on the service level.

In addition to the UED, the standard also offers the Adaptation QoS tool (AQoS) which is used to specify how multimedia content can be adapted. By using an AQoS description one can describe the properties, parameters and qualities of a resource and the relationships between them. In this case, the properties, parameters etc. are described using IOPins. An IOPin is a concept introduced within MPEG-21 DIA and is comparable with a mathematical variable. It has a unique name and can hold exactly one value. Each IOPin has a certain domain, which is the set of values that can be assigned to the variable. Depending on the actual content and the use case, IOPins can represent the bitrate of a video, its spatial resolution, the frame rate and the resulting quality in terms of PSNR. Dependencies exist between these properties that have to be considered during decision-taking. For example, an increase in spatial resolution will also increase the video's bitrate and the PSNR value. The Adaptation QoS tool offers the concept of a module to model these dependencies. A module can be thought of as a mathematical function which can be characterised by its input values (the function's arguments) and the output (the function's value). Three different types of modules are defined within the standard. The look-up table and the utility function can be used to describe discrete functions by explicitly listing all possible combinations of input values and the corresponding output. Additionally, the interrelation of continuous IOPins can be described by stack-functions that can be seen as algebraic expressions in Reverse Polish notation (RPN). In the context of our approach, the Adaptation QoS description is used to build a Cross-Layer Model (XLM) that covers aspects of the Service Layer (quality and end-user centric), Application Layer (parameter selection for the video and audio encoding) and the Network Layer (monitoring data).

The missing link between the Adaptation QoS tool, that describes the content's adaptation possibilities, and the UED tool, that specifies the actual usage context, is the Universal Constraint Description (UCD) tool. It specifies two types of constraints that can be used to link the descriptions mentioned above. Limitation constraints are used to restrict the value of a certain IOPin (=variable). In most cases these limitation constraints are used in combination with references to the usage environment description to limit certain properties of the content based on the usage context. For example, streaming a video with a resolution that cannot be displayed on a terminal should not be regarded as a feasible adaptation decision. The constraints are therefore limiting the possible adaptation space in terms of feasible adaptation parameters. Finally, the concept of optimization constraints can be used to steer the selection of the final adaptation decision from the feasible adaptation space. The optimization constraint can be seen as an objective function that can either be maximized or minimized. Furthermore, the standard allows the definition of multiple optimization constraints.

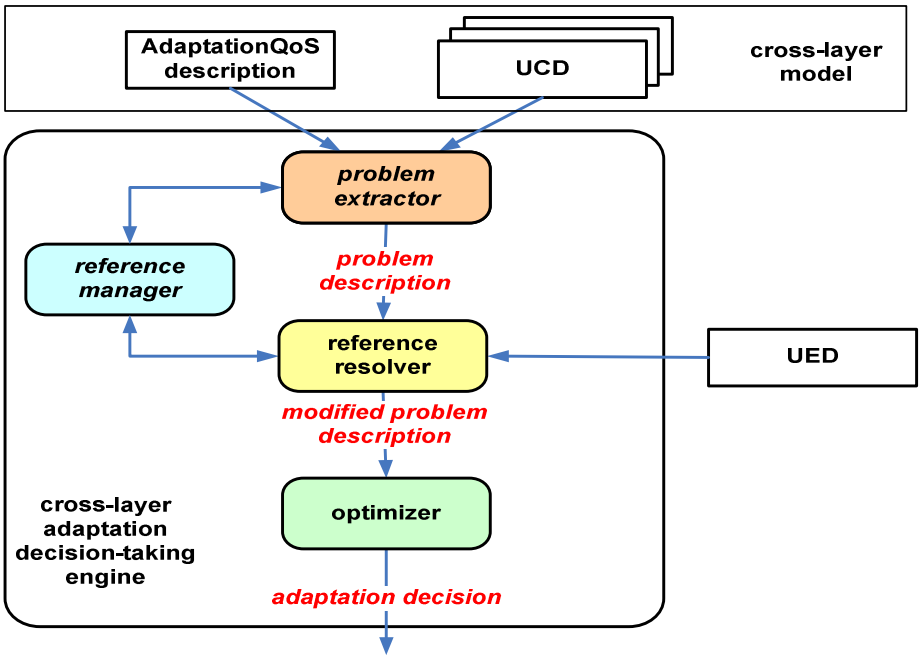

Figure 3. Architecture of MPEG-21 based Cross-Layer ADTE

The advantage of modelling the decision-taking as an optimization problem is that it can be then processed and solved by a generic software component. An architecture of such an MPEG-21 based Cross-Layer ADTE is given in Figure 3. . It comprised of a problem extractor component that parses the cross-layer model that is represented by MPEG-21 Adaptation QoS and UCD descriptions and extracts a description of the mathematical optimization problem. The problem is refined by resolving references to the usage context description (UED) which includes the monitoring data of the network, effectively replacing the values referenced in the constraints e.g. the vertical screen resolution or the available bandwidth are substituted with the values taken from the UED. This task is carried out by the reference resolver component which is also responsible for parsing the UED, which is in turn controlled by the reference manager which keeps track of all references that are encountered in the cross-layer model. The output of the reference resolver is therefore a modified version of the initial problem description. Finally, this optimization problem is fed into an optimizer component which applies a mathematical optimization algorithm. The goal is to find a optimal assignment of IOPins values for the limitation constraints with respect to the optimization constraints. This value assignment is considered as the solution of the optimization problem and conveys the parameters for the actual content adaptation.

Within the ENTHRONE project the XL ADTE is encapsulated in a dedicated component of the EIMS system called the Adaptation Manager. This is used to act as the control entity for a content processing entity that is referred to as the television and multimedia processor (TVM) within the project. Thus there is a sharp distinction between the actual component that performs the adaptation (the TVM) and the component that is controlling it.

7 Experimental Results

7.1 Test-bed Configuration

We set up a test-bed comprised of two autonomous domains, AS1 and AS2 (as can be seen in Figure 4. , which effectively represents the node level of Figure 1.), to simulate two NPs configured with egde-to-edge domain RTTs of 50 ms and 60 ms respectively. Each NP domain uses the "NIST Net[1]" software to emulate the NP network-wide (WAN) behaviour. The test-bed was deployed to validate the end-to-end monitoring system, to evaluate its response time and accuracy and to show NQoS to PQoS mapping results. In this work, the one-way loss is obtained from the IP Performance Metrics (IPPM) QoS metric (see [28] and [29]) that is used as the main input for the AppMon. We use Differentiated Services traffic classes, termed Expedited Forwarding (EF) for Golden services, Assured Forwarding (AF) for Silver services, and Best-Effort (BE) for the Bronze services [29].

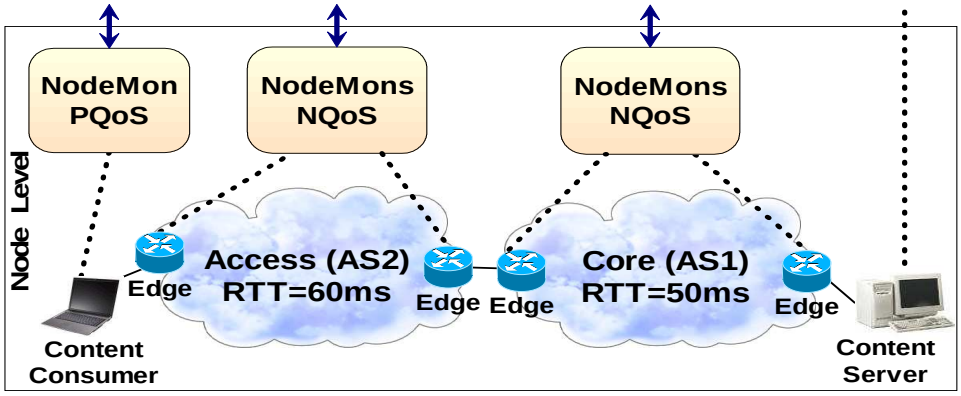

Figure 4. Node level part of monitoring system deployement test-bed

7.2 Monitoring System Response time Analysis

Figure 5. shows the response time of the monitoring system when the network load is gradually increased in steps of 4% of the total capacity of links between two edge routers. Here, the response time is effectively the time elapsed between the monitoring order issuance at the EIMS CustSrvMngr and the time when the monitoring results are received. We assume that the EF traffic has a fixed bandwidth share that allows the traffic to be serviced even during congestion periods. The signalling traffic, marked as EF traffic, uses a dedicated part of the total bandwidth and is subsequently not affected by network conditions. Figure 5. clearly reveals that the measured values of the response time for each service class are rather stable over the time. The oscillations are due to the natural behaviour of TCP/SCTP as explained above. Since all signalling traffic was marked as EF traffic, fairly good response times were maintained (approximately 500 ms for all services classes). Hence, the network load dynamics affect only the user traffic and not the signalling traffic.

1 NIST Net is a network emulation package. http://www-x.antd.nist.gov/nistnet/

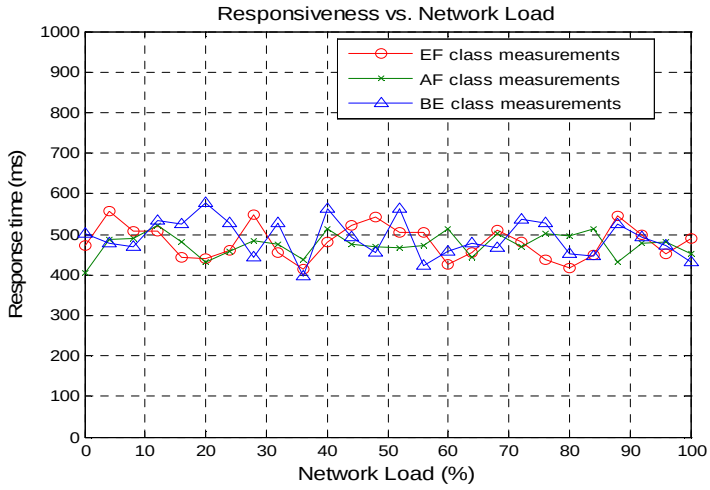

Figure 5. Monitoring system response time for different level of network load

7.3 Network Level Loss Measurement Accuracy Analysis

In order to characterize our monitoring system accurately, we explicitly introduced specific loss rate ratios of 0, 15 and 30%, for EF, AF and BE service classes respectively. Three monitoring jobs were created to measure the respective QoS metrics related to each traffic class (EF, AF, BE). Furthermore, the measurements were repeated 25 times to get more information about the "long-term" accuracy of our monitoring system and its ability to continuously perform measurements and produce measurement results.

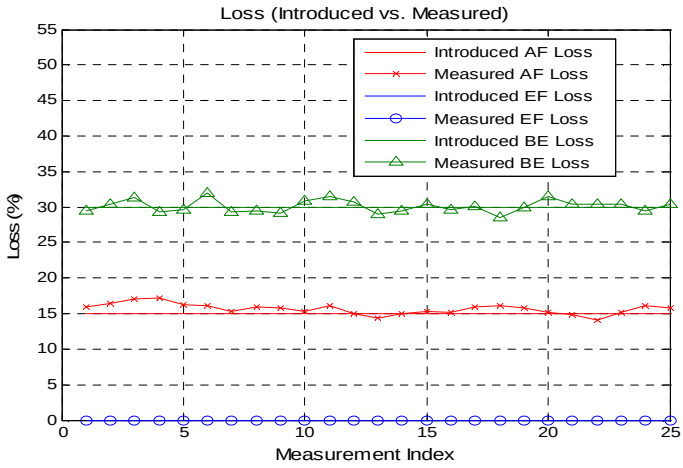

Figure 6. Packet loss ratio measured for EF, AF and BE traffic classes

As shown in Figure 6. , the loss rate values measured were very close to the target values introduced, even if the measurements fluctuated around the introduced mean loss rate for AF and BE traffic over time. This is due to the burst characteristic exhibited in the packet losses generated by NIST Net that uses the well-known Gilbert (good/bad) model to generate the packet loss patterns.

7.4 NQoS to PQoS Mapping Results

In order to avoid the stochastic nature of packet loss on PQoS degradation, the authors in [17] and [16] considered the uniform random packet loss scheme, which corresponds to the theoretically worst case scenario for a decoding threshold of 1.0, since it results in a uniform distribution of packet losses and the consequent minimization of the respective expected number of successfully decoded frames. For this reason, we investigated and developed a theoretical mathematical model in [17] and [16], which estimates the theoretically expected decodable frame rate for a given uniform packet loss distribution, without requiring any experimental procedure or measurement. For the completeness of the current paper, we summarize the important results and the proposed model as

presented [17] and [16]. The proposed NQoS to AppQoS model of the theoretically expected decodable frame ratio is presented in the following equation:

$$Q = \frac{N_{dec}}{(N_{total-I} + N_{total-P} + N_{total-B})} = \frac{N_{dec-I} + N_{dec-p} + N_{dec-B}}{(N_{total-I} + N_{total-P} + N_{total-B})} \Rightarrow$$

$$Q = \frac{(1-p)^{C_I} * N_{GOP} + (1-p)^{C_I} * \sum_{j=1}^{N_p} (1-p)^{jC_p} * N_{GOP} + \left[(1-p)^{C_I + NC_P} + \sum_{j=1}^{N_p} (1-p)^{jC_p} \right] * (M-1) * (1-p)^{C_I + C_B} * N_{GOP}}{(N_{total-I} + N_{total-P} + N_{total-B})}$$

Where CI CP CB are the mean number of packets for transporting the data of each frame type, p is the packet loss rate, NGOP is the total number of GOPs in the video flow, Ndec is the total number of decodable frames in the video flow, Ndec-I Ndec-P Ndec-B are the number of decodable frames in each type and Ntotal-I Ntotal-P Ntotal-B are the total number of each type of frames.

The experimental results of [17] and [16] show that the dependency of the theoretically expected decodable frame rate on the packet loss rate for the random uniform packet loss scheme can be successfully described by the following equation: $Q = C_1 Ln(p) – C_2$, where C_1 and C_2 are constants depending to the packet size distribution and other appQoS parameters of the stream under examination, such as the content spatiotemporal dynamics. Different packet sizes have been examined with the observation that they produce a slight variation in the slope and offset of the curves (Figure 7.).

Table 1. Analytical approximations of Q

Packet Size	Analytical Expression	R-square
500 bytes	Q = 0,2815Ln(p) – 0.2227	0.9901
1000 bytes	Q = 0,3211Ln(p) – 0.4094	0.9971
1500 bytes	Q = 0,2697Ln(p) – 0.4573	0.9198

The generic form Q = C1Ln(p) – C2 is specialized for the common packet sizes of 500, 1000 and 1500 bytes as depicted in Table 1, along with the respective R-square values, denoting the similarity of the experimentally derived results and the proposed logarithmic expression.

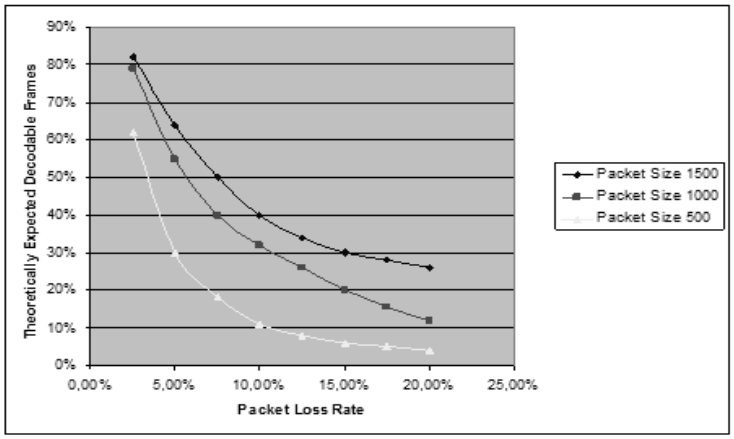

Figure 7. The Theoretically Expected Decodable Frames vs. the Packet Loss Rate for various packet size distributions

Going beyond the random uniform packet loss scheme, [16] examined the case of burst packet loss schemes, which are more suitable for emulating the real transmission conditions by experimentally specifying an offset multiplier on Q:

$$Offset = \begin{cases} \dfrac{1}{-3.9204p + 1.0315} + \dfrac{0.05}{(C_1 Ln(p) - C_2)}, & 0.01 < \dfrac{(C_1 Ln(p) - C_2)}{-3.9204p + 1.0315} < 0.5 \\ \dfrac{1}{-3.9204p + 1.0315}, & 0.5 < \dfrac{(C_1 Ln(p) - C_2)}{-3.9204p + 1.0315} < 0.1 \end{cases}$$

Therefore, depending on the packet loss ratio p and the packet size, the Theoretically Expected Percentage of Successfully Decodable frames (TEPSD) for burst and non-uniform packet loss schemes is provided by the following formula:

$$\text{TEPSD} = \textit{Offset Multipier} * Q$$

Using the NS-2 simulation environment, the proposed calibrated model was compared with experimentally measured values of successfully decoded frames over packet loss environment with burst schemes. The Q-Q plot of Figure 8. was subsequently derived from this comparison.

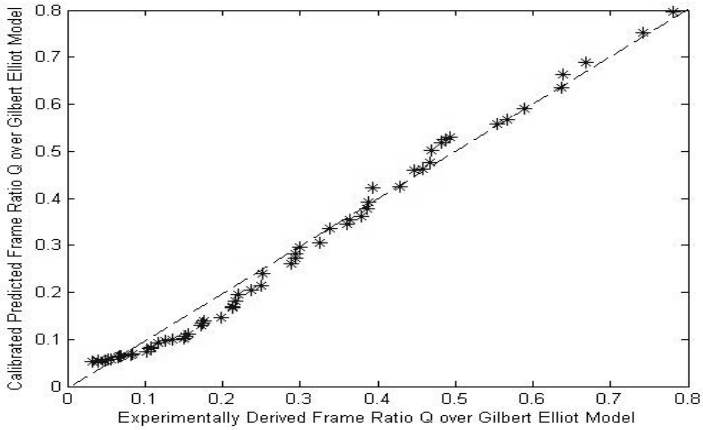

Figure 8. Comparison of the experimental Decodable Frame Rate Q for G-E packet loss scheme to the predicted Q of the calibrated model

Comparing the Q-Q plot of the calibrated model, it can be deduced that the proposed offset multiplier provides satisfactory mapping of the predicted number of successfully decodable frames to those derived throughout the experiment. Therefore, the proposed model has been successfully validated for the realistic case of burst packet loss schemes, providing satisfactory approximations.

Thus, returning again to the PQoS equations of Section 4, the variable x can be described as the duration percentage of the discontinuities over a period of ten seconds, which means that it can be further mapped to the percentage of the dropped frames from the total frames of a 10 sec signal. Moreover, this mapping is 1-to-1 without requiring any further sophisticated implementation. So, the variable x (measured in msec) can be substituted by the percentage of the dropped frames (i.e. the complimentary of the TEPSD multiplied by 10^4). Using this, the above equation can be further formulated as:

$$PQoS\ Level = 85.8 - \frac{53.03}{1 + \left(\frac{562}{(1 - TEPSD)10^4}\right)^{1.01}}$$

Based on the described layered approached to the NQoS to PQoS mapping ,a mapping tool was developed which receives the packet loss ratio of the transport ratio (i.e. the NQoS statistics) as input and estimates the respective degradation that is caused across all the layers described in the proposed approach (i.e. application and service layers). Therefore, the tool provides information on:

a. the estimated successfully decoded frames
b. the Theoretically Expected Percentage of Successfully Decodable frames (TEPSD) for burst and non-uniform packet loss schemes
c. the offset multiplier
d. the predicted MOS score
e. the MOS description that corresponds to the specific score

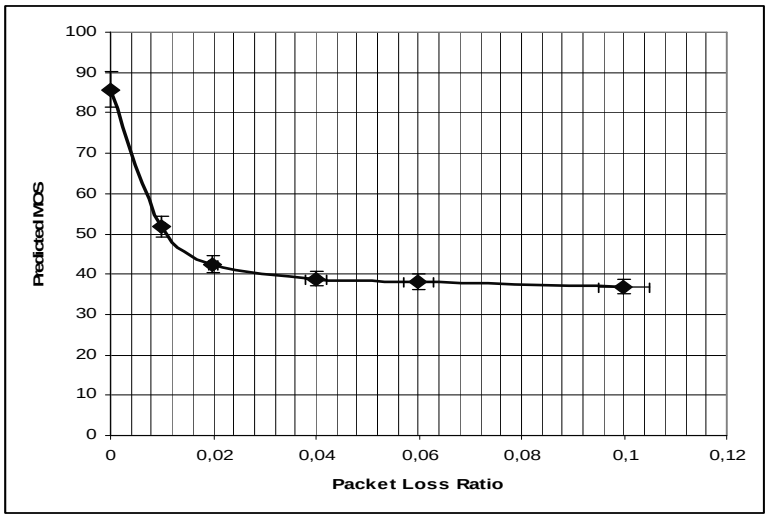

Figure 9. The experimentally estimated MOS value derived from the proposed NQoS-to-PQoS monitoring tool

In this framework, experimental measurements were taken from a reference tool that provides the mapping of the NQoS statistics to PQoS estimation for the case of packets of 1000 bytes. The experimental results are depicted in the Figure 9. , where it can be observed that the perceptual quality of a media service suffers high initial degradation for relatively small packet loss percentages, while for higher ratios that estimated quality the degradation follows a more gradual slope, showing that due to masking effects, the viewers do not perceive subsequent degradation so strongly.

The above results show that loss rate within traffic classes can be accurately and individually monitored and utilized to provide estimates of the expected PQoS. Thus, our monitoring system can provide real-time and accurate monitoring data as inputs to EIMS-AM for end users service assurance and network resource optimization.

8 Conclusion

This paper describes a quality aware end-to-end cross-layer QoS monitoring system for ensuring that an acceptable level of end-to-end service is provided to customers by service providers. The monitoring system provides sufficient information (network conditions information and perceived quality assessment) for appropriate remedial actions to be taken, e.g. service adaptation in case of link degradation or failure, or non-conformance with SLSs. To achieve this aim, following the presentation of monitoring components, this paper discusses how the errors and impairments of the transmission channel can be mapped to the various QoS-related layers of the video service. More specifically, it discussed the phenomena that occur in PQoS, AppQoS and NQoS layers as a result of transmission predicaments. In this context, the effect of the packet loss ratio on the theoretically expected ratio of decodable frames is discussed, describing how the interdependencies of the encoded frames create error propagation. Following this, the perceptual impact of the lost frames is mapped to the Service Layer, exploiting a subjectively validated mapping of frame loss to perceived video quality. Then, by combining the described mappings, a NQoS to PQoS mapping framework is proposed. Finally an MPEG-21-compliant cross-layer media content adaptation module is also presented, completing the presentation of a service level monitoring system that supports the dynamic management of the end-to-end services offered to customers.

References

[1] D. C. Verma, “Service Level Agreements on IP Networks”, *Proceedings of the IEEE*, vol. 92, no. 9, pp. 1382-1388, September 2004.

[2] P. Flegkas, ed., D1.1: “Specification of Business Models and a Functional Architecture for Inter-domain QoS Delivery”, May 2003. http://www.mescal.org/ Public Deliverables.

[3] TRT, ed., ENTHRONE I Deliverable D24F, “Specification of protocols, algorithm, and components, the architecture, and design of SLS Management”, July 2005.

[4] E. Le Doeuff, ed., ENTHRONE I Deliverable 01 “Overall system requirements and functional architecture specification”, 31 March 2004.

[5] P. Brétillon, ed., ENTHRONE II Deliverable D01, “Overall system architecture – version 2”, February 2007.

[6] C. Timmerer, et.al., “An Integrated Management Supervisor for End-to-End Management of Heterogeneous Contents, Networks, and Terminals enabling Quality of Service”, *Proceedings 2nd European Symposium on Mobile Media Delivery (EuMob) 2008*, Oulu, Finland, July 2008.

[7] V. Paxson, J. Mahdavi, A. Adams,, and M. Mathis “An Architecture for Large-Scale Internet Measurement” *IEEE Communications Magazine*, vol. 36 no. 8, pp. 48-54, August 1998. .

[8] A. Asgari, P. Trimintzios, G. Pavlou, R. Egan, “Scalable Monitoring Support for Resource Management and Service Assurance”, *IEEE Network Magazine*, Dec./Nov. 2004, Vol. 18, No. 6, pp. 6-18.

[9] RTG high-performance SNMP statistics monitoring system is available at sourceforge.net/projects/rtg/.

[10] IST, European IST research projects, for more information visit: www.cordis.lu/ist/. Specifically for IST-INTERMON visit: www.ist-intermon.org/, for IST-MoMe visit: www.ist-mome.org/, for IST-LOBOSTER visit: www.ist-lobster.org/, for IST-AQUILA visit: www-st.inf.tu-dresden.de/Aquila/, and for IST-SCAMPI visit: www.ist-scampi.org/.

[11] P. Brétillon, ed., ENTHRONE I Deliverable 25.2 “Perceived Quality Meters and Agents Prototypes”, August 2005.

[12] S. Kanumuri, P. C. Cosman, A.R. Reibman, V.A. Vaishampayan, “Modeling Packet-Loss Visibility in MPEG-2 Video”, *IEEE Transactions on Multimedia*, vol.8, no.2, pp. 341-355, April 2006.

[13] Z. He, H. Xong, “Transmission Distortion Analysis for Real-Time Video Encoding and Streaming over Wireless Networks”, *IEEE Transactions on Circuits and Systems for Video Technology*, vol.16, no.9, pp.1051-1062, September 2006.

[14] T. Ahmed, ed., ENTHRONE Deliverable D23f “End-to-end QoS Signalling & Policy-based Management Architectures,” August 2005.

[15] A. Mehaoua, T. Ahmed, H. Asgari, M. Sidibé, A. NAFAA, G. Cormentzas, and T. Kourtis, "Service-driven Inter-domain QoS Monitoring System for Large-scale IP and DVB Networks", *Computer Communication*, Special Issue on Monitoring and Measurements of IP Networks, vol. 29, no. 10, pp. 1687-1695, June 2006.

[16] C.-H.-Ke, C.-H.-Lin, C.-K. Shieh, and W.-S. Hwang, “A Novel Realistic Simulation Tool for Video Transmission over Wireless Network,” *IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC2006)*, Taiwan, 2006.

[17] H. Koumaras, A. Kourtis, C-H Lin, C-K Shieh, “A Theoretical Framework for End-to-End Video Quality Prediction of MPEG-based Sequences”, *Proceedings of International Conference on Networking and Services (ICNS07)*, Athens, Greece, June 2007.

[18] R. R. Pastrana-Vidal, J. C. Gicquel, C. Colomes, and C. Hocine, “Sporadic frame dropping impact on quality perception,” *Proceedings SPIE Electronic Imaging, Human Vision and Electronic Imaging IX*, pp. 182–193, 2004.

[19] H. Koumaras, A. Kourtis, D. Martakos, J. Lauterjung, "Quantified PQoS Assessment Based on Fast Estimation of the Spatial and Temporal Activity Level", *Multimedia Tools and Applications*, vol. 34, no. 3, pp. 355-374, Sep. 2007.

[20] H. Koumaras, F. Liberal, L. Sun, "PQoS Assessment Methods for Multimedia Services", in "Wireless Multimedia: Quality of Service and Solutions", Editors Dr. Nikki Cranley, Dr. Liam Murphy, IGI Global Pub. ISBN: 978-1-59904-820-8, July 2008.

[21] M. Sidibé, ed., ENTHRONE II Deliverable D06F, “Service Management and Monitoring”, February 2008.

[22] D. Jannach, K. Leopold, C. Timmerer, and H. Hellwagner, "A Knowledge-based Framework for Multimedia Adaptation", *Applied Intelligence*, vol. 24, no. 2, pp. 109-125, 2006.

[23] D. Mukherjee, E. Delfosse, J.-G. Kim, and Y. Wang, "Optimal Adaptation Decision-Taking for Terminal and Network Quality-of-Service", *IEEE Transactions on Multimedia*, vol. 7, no. 3, pp. 454-462, June 2005.

[24] I. Kofler, C. Timmerer, H. Hellwagner, A. Hutter, and F. Sanahuja,, "Efficient MPEG-21-based Adaptation Decision-Taking for Scalable Multimedia Content", *Proceedings of the 14th SPIE Annual Electronic Imaging Conference – Multimedia Computing and Networking (MMCN 2007)*, San Jose, CA, USA, January/February 2007.

[25] I. Burnett, R. Koenen, F. Pereira, and R. Van de Walle (eds.), *The MPEG-21 Book*, Wiley, 2006.

[26] A. Vetro, “MPEG-21 Digital Item Adaptation: Enabling Universal Multimedia Access”, *IEEE Multimedia*, vol. 11, no. 1, pp. 84-87, January-March 2004.

[27] A. Vetro and C. Timmerer: “Digital Item Adaptation: Overview of Standardization and Research Activities”, *IEEE Transactions on Multimedia*, vol. 7, no. 3, pp. 418-426, June 2005.

[28] V. Paxson et al., “Framework for IP Performance Metrics,” *IETF RFC-2330*, May 1998.

[29] IETF, For information on the various IETF working groups including IPPM and DiffServ, visit http://www.ietf.org.