Modelling an Adaptive-Rate Video-Streaming Service Using Markov-Rewards Models[1]

I.V. Martín F., Juan J. Alins-Delgado, Mónica Aguilar-Igartua, Jorge Mata-Díaz

Telematics Engineering Department, Technical University of Catalonia (UPC),
Jordi Girona 1-3, 08034, Campus Nord, Barcelona, Spain.
{isabelm, juanjo, maguilar, jmata}@ entel.upc.es

Abstract. Nowadays dynamic service management frameworks are proposed to ensure end-to-end QoS. To achieve this goal, it is necessary to manage Service Level Agreements (SLAs) which specify quality parameters of the services operation such as availability and performance. This work is focused on video-on-demand (VoD) services to investigate the goodness of performability techniques in end-to-end QoS scenarios. Based on a straightforward Markov Chain, Markov-Reward Chain (MRC) models are developed in order to obtain various QoS measures of an adaptive VoD service. The MRC model has a clear understanding with the design and operation of the VoD system. In this way, several design options can be compared. To compute performability measures of the MRC model, the randomization method is employed. Predicted model results fits well with the ones taken from a real video-streaming testbed.

1 Introduction

During the last years, video-on-demand (VoD) applications for the transmission and distribution of video have experienced a growing development and acceptance from the users. Video-streaming systems have a special relevance in wire and wireless networks. In these systems, the video is distributed for its reproduction in real-time [1]. The video server of a video-streaming system stores a set of movies that can be requested by any client. If the connection request is accepted, a session is initiated; then a multimedia stream flows through a set of heterogeneous networks from the video server to the client terminal. With the aim of reducing the high amount of information generated by the video source, loss compression techniques are applied. The most common coding techniques are H.26x and MPEG standards [2]. The price to pay for a high compression level is a degradation level in the image quality. Thus, the final Quality of Service (QoS) provided to the user of these streaming services depends on the available network resources and on the time-varying channel in wireless access networks. In end-to-end QoS scenarios, QoS measures such as packet loss, packet delay and jitter are guaranteed when the connection is admitted. These real-time guarantees required for VoD system could be achieved using QoS differentiation between traffic classes over heterogeneous networks. A proposal on a DiffServ model was

[1] This work has been financed by the Spanish investigation projects DISQET (CICYT TIC2002-00818), CREDO (CICYT TIC2002-00249) and ARPA (CICYT TIC2003-08184).

published recently in [3]. A more general approach for heterogeneous network with a well suggested solution is presented in [4].

Constant image quality MPEG video flows present a high variability in its transmission rate. This variability is produced by the coder algorithm and the complexity of the video sequences. In this scenario, it has a growing importance to design video sources able to adapt their output bit rate to the time-varying network resources. Some proposals of design and evaluation of adaptive VoD systems are presented in [5, 6, 7, 8, and 9]. However, most of these proposals use simulation models or real platforms to carry out the performance evaluation of VoD systems. These evaluation techniques difficult the system analysis and, also, the study of several design options. In addition, some analytical proposals do not regard the interaction between the different video sources sharing the network resources. On the other hand, works focused on characterizing and modelling a single video flow [10, 11, 12] are not enough to evaluate an adaptive VoD session because they do not take into consideration the dynamic of the video quality changes along the stream transmission.

Therefore, currently there is a lack of suitable tools or methods to accomplish the design and dimensioning of these adaptive VoD systems. Both the services providers and the customers are indeed interested in having tools that allow quantifying the system performance from their points of view. Analytical tools are the most appropriate mechanism to facilitate the evaluation required. Moreover, these tools must provide feasibility to incorporate modifications into the system in an easy way and admit a computational evaluation. This kind of analytic tools help to manage some of the main typical required objectives: maximizing the use of network resources, maximizing the QoS offered to the users and defining billing metrics. Likewise, these tools could compute diverse parameters in order to specify, manage and control the fulfillment of the Service Level Agreements (SLAs). The management of SLAs is a current challenge into the multimedia services area. Further, it has a great commercial interest. There are diverse recent proposals about SLA management (e.g. [13, 14]), although none of them specifies how to quantitatively evaluate the user-level SLA parameters.

One of the main objectives addressed in the present work is to compute *a priori* the QoS offered to the user of a video-streaming application. A constant image quality is attempted to be provided to the customer of the designed VoD service. This can be achieved with the successful transmission of the flow coded with the image quality selected by the user. The bit rate variability of the flow raises renegotiations with the network in order to modify the allocated resources along the session. These renegotiations are performed at the temporal-scale of the scenes in a video sequence. In this way, the amount of network resources reserved along the session are reduced substantially, and a better exploitation of these resources is reached [1, 8, and 12]. Therefore, the number of concurrent streaming sessions in the system is incremented. However, the image quality will be reduced in some congestion moments when the flow with the selected quality cannot be transmitted. In these congestion situations, the service adapts the transmission bit rate to the available network resources applying a higher compression level or managing the enhanced layers when scalability techniques are employed. Then, the user-level QoS can be measured by means of various parameters, such as the image quality, the reserved resources, or the effectively used resources. In

order to compute these end-to-end QoS measures, a generic method to construct analytic models for VoD systems is proposed in this paper.

Some methodologies classically used in Performability analysis have been investigated to develop the proposed model and also to carry out computations of diverse system measures. Performability integrates the aspects of Performance and Dependability [15]. These methodologies have been previously employed, e.g., for buffer dimensioning [16, 17] and for evaluating connection admission control algorithms [18].

In the present work, performability analysis is proposed as a new way to design and evaluate adaptive multimedia services architectures. The proposed methodology considers the interaction between multiplexed connections sharing the network resources and it could include end-to-end time-varying channels. Moreover, our proposal allows that any system modification could be incorporated in the analytic model in an easy way. The applicability of this proposal is based on the characterization of the coded multimedia flows and the channel behavior. This characterization requires suitable markovian models of these elements.

The rest of the work is organized as follows. Section 2 describes the evaluated VoD system. In section 3, the preliminaries of the model are pointed out. In section 4 we propose a generic method to derive adaptive VoD system models. An example applying this method is presented too. Next, in section 5 some numerical results evaluating the exemplified VoD model are shown. These results are compared with experimental measurements obtained from the SSADE project (*http://ssade.upc.es*) implemented by the Telematics Services Research Group of the Polytechnic University of Catalonia. Finally, conclusions and future works are presented in section 6.

2 The System Description

Figure 1 depicts the VoD system analyzed in this work. A set of VBR MPEG-II flows for each sequence have been stored in the video server. Each available flow offers a different image quality according to the quantization step (Q) set on the MPEG codification. Different network resources are required during the whole transmission of a sequence. These required resources depend on the variation of the sequence complexity. The resources requirements are related with scenes or segments of the video sequence. The system uses the RSVP (Resource reSerVation Protocol) as signaling protocol to manage resources reservations requests. The stream transmitted to each accepted session can alternate between the different available coded flows of the sequence. The transmitted available flow is chosen according to the image quality selected by the user and to the result of the reservation request produced by the end-to-end admission control of the RSVP-based system. The available network resources change due to the interaction between the multiplexed connections. For each available flow, the *Statistical Planner* block (see Fig. 1) has previously calculated and stored *Traffic Specification* parameters (TSpec) of each scene and the resources renegotiation moments in each sequence. When scene changes or variations of available resources happen, the *Regulator/Negotiator* block decides what flow (Q_i) will be transmitted. To guarantee a minimum video quality, a minimum reservation should always be assured to transmit the lower image quality flow. The *Traffic Shaper* block extracts

the variability introduced by the frame coding modes (Intra, Predicted and Bidirectional-Predicted) of the MPEG algorithm. In this way, the bit rate is smoothed and it is maintained constant (r_{GoP}) for a GoP (Group of Pictures) interval.

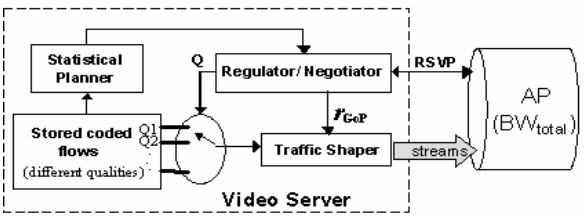

Fig. 1. The system model for the VoD service

3 Preliminary

It is not objective of this work to carry out a detailed study of the diverse proposals about resource renegotiation, neither to determine how the TSpec parameters for each flow are obtained. In this work, the resources required to transmit the scenes of each available flow are considered known and previously calculated. With this information, a markovian model for each available flow can be derived. This markovian model statistically characterizes the resources required by the transmission of the flow at scene level. Related models are proposed in works such as [19] and [8]. In these scenic markovian models, the states identify classes of scenes. In this way, scenes with similar activity or complexity levels are associated to the same state. On the sake of the simplicity, we will refer to each class of scenes as an activity level.

Straightforward scenic markovian models represent the scene changes by means of the transitions between states. For a set of video flow models of the same video sequence, the scene changes occur at the same time. Therefore, the transitions are time-tracked at the same activity level in a coordinated multiflow model. Hereafter, we will indistinctly refer to a change of the activity level of the sequence or a change of anyone of its coded flows.

The generic methodology proposed in this work can be developed starting from these scenic markovian models. With the characterization of the available constant-quality flows, for each sequence, an analytical model of the adaptive VoD systems will be constructed.

4 The Analytic Model

This section describes the proposal to construct analytic models for VoD systems. These models characterize the behavior of a set of VoD connections interacting and sharing network resources with QoS guarantees. Modelling generic methodology consists on the following steps:

1. The establishment of a markovian model that characterizes the resources required by each available flow, coded with a fixed image quality.
2. The integration of the models obtained from step 1 reflecting the dynamic introduced by the renegotiation mechanism when a scene change happens or the assigned network resources are modified. Incorporating the transitions between the available flows this integration is done. The defined transitions identify the behavior of the video stream electing the suitable coded flow. Thus, the model of the delivered stream for a single adaptive connection in the system is obtained, where the transitions capture the designed dynamic scheme.
3. The model for the system with N connections is derived as follows. The system state is defined by the number of connections being in each one of the states from the model in step 2. The transitions are easily defined fromthe markovian behavior of the connections. In this way, a Continuous-Time Markov Chain (CTMC) is derived. This CTMC models the aggregated traffic generated by N connections accepted in the VoD. Then, a VoD system model is accomplished applying step 5 to this CTMC. This model allows evaluating several measures of the VoD operation; but it does not allow obtaining measures for a particular user. To achieve this, let's continue to step 4.
4. The system model that allows evaluating a single connection in interaction with other N accepted connections is reached. To reach this, the integration of the models established in steps 2 and 3 is done. Step 2 provides the state of the analyzed connection and step 3 provides the state of the other N connections. In this way, a CTMC that identifies the state of a connection interacting with the other N connections is attained. This markovian process evolves according to the interaction between all the connections.
5. A reward rate is associated to each CTMC state. This reward rate represents a measure of the steady-state performance gained by the system per time unit in the corresponding state. In this way, a Markov-Reward Chain (MRC) is achieved. This chain models the behavior of the VoD system. Let $M(t)$ be the random variable (r.v.) of the accumulated reward during an observation time t. Several statistical measures of $M(t)$ can be computed. These measures provide different performance values of interest for the VoD system.

In order to compute solutions, the so called randomization or uniformization method is applied to the MRC achieved after applying step 5. This method involves transforming a continuous-time chain into an equivalent discrete-time chain. An extensive description of this methodology can be found in [20]. Both the reward markovian models and the randomization method are generally used to accomplish transient solutions of many performability measures.

Then, varying the rewards structure and applying the randomization technique to solve the MRC, different statistical measures can be computed. Some of these measures are:

- **Expected value accumulated at time t, $E[M(t)]$.** It is the expected value of a measure $M(t)$ accumulated by time t. For example, it allows obtaining the mean number of bits transmitted to a user for their whole session. To reach this, the reward is the transmission bit rate associated to each state. The length time of the session is t. Equation 3.10 in [20] provides the numeric evaluation of $E[M(t)]$.

- **Mean Failed Time.** A failure in our system can be defined as the non-fulfillment of some agreement of the service contract. Hence, measures related to the time of failure are achieved assigning a reward structure of such that: if the system state is in failure, the assigned reward is 1; otherwise, the assigned reward is 0. The expected failed time is calculated with $E[M(t)]$ applying this reward structure.
- **Probability distribution function (PDF) of the Failed Time.** This measure can be interpreted as the probability of non-fulfillment time of the service contract. Let $P\{M(t) < u\}$ be the probability that the total time in non-compliant SLA situations be smaller than u units of time. In this case the same reward structure presented for the *mean failed time* should be applied. An efficient solution to numerically evaluate this measure can be found in [21].

Concluding, several design options of VoD systems can be modelled and solved using this generic methodology. Moreover, different user-level QoS measures can be computed. Next, we develop the analytic model of the VoD system described in section 2, where some design options have been particularized. Only a summary is presented due to limited space reasons.

4.1 An example to obtain a VoD model using the proposed methodology

Following, the steps involved in the model construction methodology are explained.

Step 1: Modelling the available flows. In this step, the model for each available fixed-quality video flow in the VoD system is obtained. As the starting point, we have considered a simple markovian model which is derived in a similar way than in [8]. This model is composed of three states that define three activity levels. The transition rates between states are inferred from the statistical analysis of the duration of the scenes classified into the discretized set of activity levels. Therefore, scenes with similar, but different, required resources are classified in the same activity level. Then, in order to characterize the amount of required resources in each activity level, a set of states must be established. The number of states can vary from one to the number of scenes classified in the activity level.

Simplicity has been chosen instead of accuracy when developing the model proposed in this work. Then, a single state is defined for each activity level. Figure 2 shows the flow models for "*The Graduate*" sequence coded with a quantification step Q equals to 4 and 8. For each flow, values showed in each state i correspond to the amount of resources required to being in activity level i. Applying a conservative approach, this value is calculated averaging the *Peak Rate* (bits/GoP) value of all the scenes classified into level I, weighted by the scene length. The transition rate $l_{i,j}$ from state i to state j is equal to the inverse of the sojourn time in state i multiplied by the probability that the connection visit the state j given it is in state i.

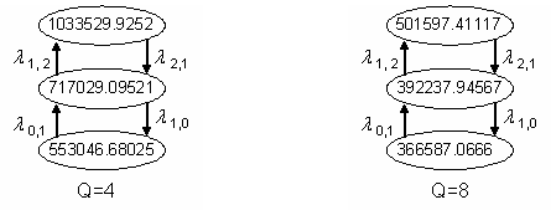

Fig. 2. Traffic models for the sequence of "*The Graduate*" movie coded with Q = 4 and Q = 8.

Step 2: Connection modelling. Along an adaptive VoD session, the stream delivered to the network matches with one of the F flows in one of its three activity levels. Transitions between activity levels determine the resources renegotiation moments. Note that, the same transition rates will be considered in the models of the available coded flows since they belong to the same video sequence and, therefore, they are time-tracked. Let e_f^a be a connection state, where a flow of quality f is transmitted (1: worst quality… F: best quality) in the activity level a (0: regular, 1: medium, 2: high).

The CTMC shown in Fig. 3 is an example derived from this step. This chain models the behavior of a connection in the system. For clarity of the drawing only 3 different constant image-quality flows have been depicted. Each column corresponds with the model of each available flow for each video sequence, which has been obtained in step 1. The vertical axis corresponds with the activity levels. Transitions between states reflect the scene changes and renegotiation decisions that have been designed in the VoD. In this example, when an activity change in the sequence happens, the system chooses the higher quality flow that the already assigned resources allow. In addition, if the change increases the activity, the system maintains or diminishes the flow quality; if the change decreases the activity, the system maintains the flow quality or it improves to the next better quality. In both cases, the system releases the resources not needed. While the stream remains in the same activity level, the system periodically tries to improve sending request for the next better quality. This process will be called *polling of improvement*.

Figure 3 depicts all the potential transitions allowed by this designed system. However, these transitions will exist depending on the required resources and on the available resources. Let $y(e_f^a, e_g^b)$ be the transition rate from state e_f^a to state e_g^b for a connection. In order to obtain expressions for these transition rates, some binary factors are defined. These factors summarizes if these transitions exist or not. Moreover, these factors determine the election of the flow quality that will be provided. For example, the increase of the activity level in the video sequence is formulated by means of the equations:

$$y(e_f^a, e_f^{a+1}) = l_{a,a+1} \cdot b(e_f^a, e_f^{a+1}) \text{ for } f = 1,..., F. \quad (1)$$

$$y(e_f^a, e_g^{a+1}) = l_{a,a+1} \cdot b(e_f^a, e_g^{a+1}) \cdot \prod_{i=g+1}^{f} \overline{b}(e_f^a, e_i^{a+1}) \text{ for } f > 1 \text{ and } g = 1,2,..., f-1. \quad (2)$$

Where $l_{a,a+1}$, $0 \le a < 2$, are the transition rates from the activity level a to the activity level a+1; the $b(e_f^a, e_g^b)$ factor equals to 1 if the resources already assigned to the connection in state e_f^a are enough to visit the state e_g^b, and equals to 0 otherwise. Finally, the $\overline{b}(e_f^a, e_g^b)$ factor is equal to $\left(1 - b(e_f^a, e_g^b)\right)$.

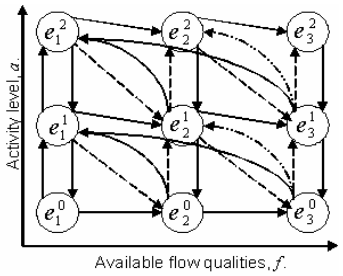

Fig. 3. Generic model of an accepted connection in the system

In order to fulfill the adaptive VoD system modelling when several connections are being served, it is mandatory to know the state of all the connections interacting to each other.

Step 3: Model to evaluate the system with *N* accepted sessions. In this example, we assume that the system serves all the users with the same QoS profile and, therefore, the same set of flows can be transmitted along their sessions. Thus, the VoD serves to a single class of users. In this case, each connection is characterized with the same values of the parameters of the generic connection model described in step 2. Then, the system state specification can be expressed through the joined formulation of the states of the *N* connections.

Let $S = \{(n_1^0, n_1^1, n_1^2), (n_2^0, n_2^1, n_2^2), \ldots, (n_F^0, n_F^1, n_F^2)\}$ be the system state, where each component n_f^a is the number of connections transmitting a flow of quality f in activity level a. Therefore, we can establish $\sum_{\forall a} \sum_{\forall f} n_f^a = N$ where N is the number of accepted connections within the system. The maximum number of system states is equal to $(N+3F-1)!/(N!(3F-1)!)$. However, some system design restrictions will reduce the space states of the adaptive VoD system.

We consider a temporal space between sessions enough to assume uncorrelated transmission. Hence, given the markovian behavior of the connections, only one connection could change at the same time. Let $S_{+(g,b)}^{-(f,a)}$ be a system state with one more connection transmitting a flow of quality g in activity level b and one less connection transmitting a flow of quality f in activity level a, regarding to state S. Then, only transitions from state S to state $S_{+(g,b)}^{-(f,a)}$ could be produced. The transition rate

$\Psi\left(S, S^{-(f,a)}_{+(g,b)}\right)$ is directly calculated as the sum of the transition rates of all the connections being in state e^a_f. So, when a connection diminishes or increases its activity level, these rates are expressed as:

$$\Psi(S, S^{-(f,a)}_{+(g,b)}) = n^a_f \cdot \gamma\ (e^a_f, e^b_g), \tag{3}$$

where $\gamma\ (e^a_f, e^b_g)$ are the rates defined in step 2.

Step 4: Model to evaluate a session in the system with *N* other accepted sessions. In order to evaluate an individual connection within the system, it is required to identify the evolution of this connection. This is carried out integrating the models obtained in steps 2 and 3.

Let $\left\{e^a_f, S\right\}$ be a system state, where e^a_f describes the state of the connection under evaluation, and where *S* describes the state of the rest of the already accepted connections (*N*) in the system. Transitions from state $\left\{e^a_f, S\right\}$ to state $\left\{e^b_g, S\right\}$ are due to a transition of the observed user. Transitions from state $\left\{e^a_f, S\right\}$ to state $\left\{e^a_f, S^*\right\}$ are due to a transition of some of the other *N* connections. The rates of these transitions are the same obtained in step 2 and 3, respectively.

Step 5: Assigning rewards. The statistical analysis of the flows allows determining any steady-state performance measure associated to each state for each flow.

From the CTMC obtained in step 3, measures of the global system operation can be performed. Note that, the reward assigned to each state *S* depends on the state of all the *N* sessions. For example, assigning a reward equals to the sum of the mean bit rate used by each one of the *N* sessions in state *S*, the total bit rate transmitted in the system can be computed.

From the CTMC obtained in step 4, the QoS of an individual user can be evaluated. In this case, the reward assigned to each state of the CTMC only depends on the state of the connection under evaluation. For example, to measure the delivered PSNR to the user, a reward equals to the PSNR associated to flow *f* in activity level *a* should be assigned to all the states $\left\{e^a_f, S\right\}$ of the chain.

5 Numerical Results

In this section numeric examples are presented. They are computed from the expected mean value of the QoS offered to a user of the VoD system. The system model as example in section 4 has been evaluated. Where the *E*[*M(t)/t*] for the PSNR, the transmitted BW and the reserved BW were computed. These results have been contrasted with experimental values taken from a real platform testbed. This testbed is the

SSADE video distribution system that provides an adaptive video-on-demand service. The configuration parameters of the VoD system used in these examples are:

- All the customers have the same QoS profile. They can reserve all the available BW and they must release any no-needy BW.
- The renegotiation decisions are the ones defined in step 2 of section 4.1.
- All the clients request the movie "*The Graduate*" in uncorrelated moments.
- For this sequence, the VoD has available a set of flows coded with Q equals to 4, 8 and 16. The code algorithm is the MPEG-II, with an IBPBPB frame mode scheme, 25 frames per second and with a format of 325x288 pels (an 8 bits pixel).

Table 1 summarizes the values of each available flow from the model described in section 4. For simplicity, a conservative admission control has been applied, based on the *Peak Rate* of the TSpec signaled in each scene. Transition rates between activity levels are shown in Table 1a. The total length of this sequence is 25325 GoPs (1 GoP=240 ms.) grouped in 140 scenes. Hence, the observation time t is 6078 seconds. The *polling* rate of improvement is 0.04 polls per GoP. Subject to the measured value, the reward rate associated to each state is the PSNR, the exactly transmitted BW (bits/GoP) or the reserved BW (bits/GoP).

Figures 4 and 5 depict the mean PSNR provided to a user of the VoD and the link utilization[2], respectively. Both figures are represented as a function of the total BW assigned to the VoD service and for a variable number (N) of accepted streaming sessions. The PSNR measure concerns to the customers, since it estimates the video quality perceived by them. The link utilization measure has high interest for the service providers because it quantifies the exploitation level of the allocated resources for the multimedia streaming service. The dark line values have been calculated using equation (3.10) of [20] with the reward values presented in Tables 1.b, 1.c and 1d. The grey line values have been obtained from the real transmissions carried out in the testbed. Discontinuities in the analytic curves are produced as a result of the discretization of the activity levels of the flows. In the markovian model of each flow, a single state was defined for each activity level, where each state has associated the mean value of the measure that concerns. This mean value is obtained from all scenes classified into the same activity level. Therefore, the performability results are discretized and they are softened when N is high. This happens because the connections are multiplexed and the values associated to the MRC states are softened when N is increased.

Figure 6 shows the evolution of the PSNR provided to a user as a function of the number of accepted connections (N) in the VoD. The BW assigned to the video service is equal to 6, 7, 12 or 20 Mbps. For a given BW, the QoS diminishes when N is increased since a higher number of customers are sharing the same network resources. In this figure, when the associated bar is nonexistent in any N value for a given BW, means that the system does not admit such quantity of users in this BW. Note that, the QoS provided by this VoD system never experiments a high degradation level. The reason is that each accepted connection can access to all the available resources. These resources are quickly occupied, so the connection soon reaches the maximum quality. On the other hand, the minimum resources required to accept a new connec-

[2] Link utilization = (mean total BW effectively used / total BW assigned to the video service).

tion are quite high. Therefore, the admission control designed in this VoD service provides suitable QoS guarantees but a strict access to the system. In this way, once the connection is accepted, the user will perceive a video quality close to the maximum selected.

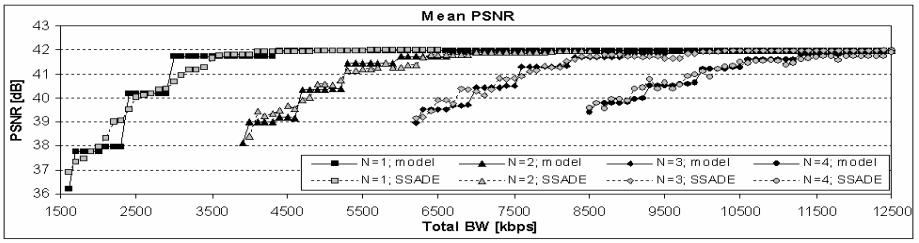

Fig. 4. Mean PSNR provided to a customer during the transmission of "*The Graduate*" movie

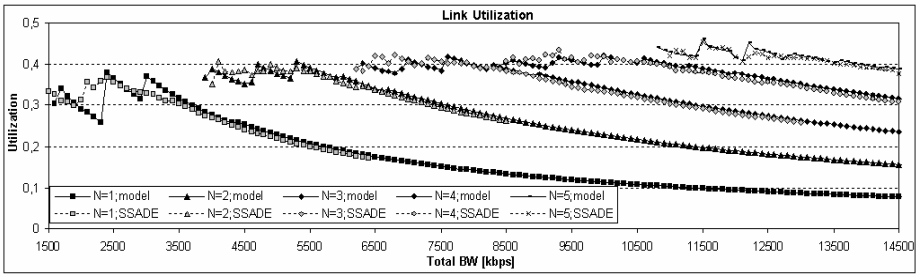

Fig. 5. Link utilization while the VoD is transmitting "*The Graduate*" movie to *N* customers

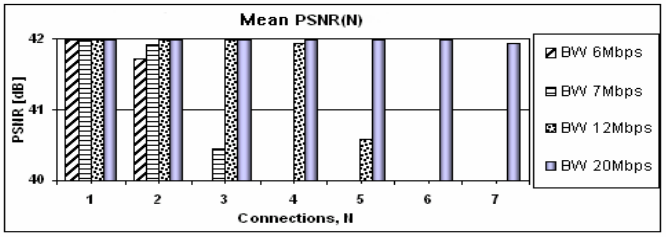

Fig. 6. Expected Mean PSNR computed from the analytic model

Table 1. Parameters[3] of the analytic model for each available coded flow of "*The graduate*" sequence with Q equals to 4, 8 and 16

1a). Transition rates, [1/GoP]

$l_{0,1}$	0.002881
$l_{1,2}$	0.002977
$l_{2,1}$	0.003850
$l_{1,0}$	0.008790

[3] These values have been computed from the statistical analysis of each flow.

1b). Mean required resources to transmit the sequence to a connection, [bits/GoP]

	Level 0	Level 1	Level 2
Vbr16	364476.06	364476.06	364476.06
Vbr8	366587.066639	392237.9456713	501597.4111675
Vbr4	553046.680253	717029.0952097	1033529.925223

1c). Mean PSNR provided to a connection, [Lineal]

	Level 0	Level 1	Level 2
Vbr16	0.00032386851	0.00049464229	0.00045332403
Vbr8	0.00013828712	0.00021552240	0.00022287058
Vbr4	5.56828935E-5	8.461717324E-5	9.752284877E-5

1d). Mean bit rate effectively transmitted to a connection, [bits/GoP]

	Level 0	Level 1	Level 2
Vbr16	76536.002774	114251.89567	161180.08402
Vbr8	119331.87367	202084.27605	275269.41624
Vbr4	227509.09722	393274.79948	615892.17413

6 Conclusions and Future Work

The end-to-end QoS provisioning in IP-based networks is a current challenge. Services providers of streaming multimedia applications in IP network are usually forced to over-dimension the allocated resources in order to ensure the level of service contracted by the customers. Moreover, the established agreements specifying the offered service level habitually only consider the specifications at the packet-transport layer. These specifications do not properly reflect the quality perceived by the user, especially in multimedia services.

In this work a new methodology has been presented to design and evaluate multimedia services in end-to-end QoS scenarios. The presented methodology facilitates the evaluation of these services by means of Performability analysis and markovian models. A continuous-time Markov-rewards model characterizing the session dynamic is constructed to analyse the system behaviour and to study the experience of an observed customer. Their evaluations are carried out applying the uniformization technique.

The methodology assigns an appropriate set of rewards to a continuous-time Markov chain according to the analyzed parameter. In this way, a wide range of parameters can be probabilistically quantified *a priori*. These parameters are related with the user's perception or the service provider's resources.

In general, the proposed methodology can be applied to any multimedia service that could be characterized through a Markovian model. With the aim to illustrate the proposal, the methodology has been developed over a video-streaming service. The considered VoD system has been implemented to manage sessions under an RSVP-enabled network. The design of this system tries to deliver a constant image quality to the user, maintaining the level of degradation introduced by the MPEG-VBR coder. To be efficient in the network resource exploitation, each session of the VoD renegotiates the amount of network resources at scene time-scale. This renegotiation increase the number of concurrent sessions allowed in the system. The improvement is obtained allocating a well fitted amount of network resources along the session. How-

ever, some renegotiations try to increment the current allocated network resources and, therefore, they can be rejected. These congestion situations are managed by the video server adapting the delivered bit stream through a higher compression level. This adaptation reduces the image quality perceived by the user.

The proposed model for VoD systems considers both intrinsic and extrinsic system variations. The intrinsic variations are due to the changes in the complexity of the video sequence. The extrinsic variations occur because of the system adaptations. These adaptations are performed as a consequence of the changing network conditions.

A Markov-Rewards Chain (MRC) has been proposed to compute the performance of the VoD system. This chain directly reflects the service operation. This important characteristic facilitates the incorporation and evaluation of possible modifications into the system design. Thus, the MRC model assists in the specification, design and dimensioning of adaptive video-streaming systems. Further, the model is straightforward, computationally evaluated and flexible enough to obtain *a priori* measures of the system performance.

The proposed analytic model allows easily computing and analyzing the user-level parameters of the Service Level Agreements (SLA). As an example, the mean PSNR provided to a customer and the utilization of shared resources have been presented. Both measures are represented as a function of the total BW assigned to the VoD service and for a variable number (N) of accepted streaming sessions. The PSNR measure concerns to the customers, since it estimates the video quality perceived by them. The link utilization measure has high interest for the service providers because it quantifies the exploitation level of the allocated resources for the multimedia streaming service.

The obtained results have been validated with the values measured in an experimental VoD testbed. The accuracy of the adjustment of these results allows verifying that the developed methodology gives a good estimation for adaptive multimedia systems in end-to-end QoS scenarios. These scenarios can include a heterogeneous group of networks, including wireless networks. The accuracy of the obtained results will depend on the suitable characterization of multimedia sources and the transmission channels with markovian models.

In the present work, the problem with homogenous customers has been evaluated as a start point. As future lines of research, we are studying efficient ways to develop the heterogeneous problem, where diverse QoS profiles and different available movies are considered. Expanding the methodology proposed in this work, the heterogeneous model corresponds to the integration of homogeneous connections models for each class of customer. But the computational cost dramatically can increase due to a high number of system states. Based on the proposed models, we are actually deriving still simpler analytical models to summarize the resources reserved by a group of users.

References

1. D. Wu, Y. T. Hou, W. Zhu, Y-Q. Zhang, J. M. Peha, "Streaming Video over Internet: Approaches and Directions", IEEE Trans. On Circuits and Systems for Video Technology, Vol. 11, Nº3: 282-300, 2001.

2. M. Ghanbari, "Video Coding: An Introduction to Standard Codecs (IEE Telecommunications Series 42)", IEE Publishing, December 1999.
3. M. Heusse, P. Starzetz, F. Rousseau, G. Berger-Sabbatel, A. Duda, "Bandwidth Allocation for DiffServ based Quality of Service over 802.11", Proceedings IEEE GLOBECOM'03, Vol. 2, pp. 992 - 997, December 2003.
4. H. de Meer, P. O'Hanlon, "Segmented Adaptation of Traffic Aggregates", Proceedings IWQoS, Vol. 2092, pp. 342-356, LNCS Springer-Verlag, Germany, June 2001.
5. G.-M. Muntean, L. Murphy, "A New Adaptive Multimedia Streaming System for All-IP Multi-Service Networks", IEEE Transactions on Broadcasting, Vol. 50, Nº1: 1-10, 2004.
6. A. Lombardo, G. Schembra, "Performance Evaluation of an Adaptive-Rate MPEG Encoder Matching IntServ Traffic Constraints", IEEE/ACM Transactions on Networking, Vol. 11, Nº1: 47-65, 2003.
7. C. Luna, L. Kondi, A. Katsaggelos. "Maximizing User Utility in Video Streaming Applications", IEEE Trans. on Circuits and Systems for Video Technology, Vol. 13, Nº2, 2003.
8. L. J. De la Cruz, J. Mata, "Performance of Dynamic Resource Allocation with QoS Guarantees for MPEG VBR Video Traffic Transmission over ATM Networks", Proceedings of the IEEE GLOBECOM'99, IEEE Communications Society, 1999.
9. R. S. Ramanujan, J. A. Newhouse, M.N. Kaddoura, A. Ahamad, E.R. Chartier, K.J. Thurber, "Adaptive streaming of MPEG video over IP networks", Proceedings 22nd Annual Conference on Local Computer Networks, IEEE, pp. 398-409, 1997.
10. A. Adas, "Traffic Models in Broadband Networks", IEEE Communications Magazine, pp. 82-89, July 1997.
11. L. J. De la Cruz, M. Fernández, J. Alins, J. Mata "Bidimensional Fluid Model for VBR MPEG Video Traffic", Fourth International Conference on Broadband Communications, IFIP, TC6/WG6.2, 538-549, 1998.
12. P. Manzoni, P. Cremonesi, G. Serazzi, "Workload Models of VBR Traffic and Their Use in Resource Allocation Policies", IEEE/ACM Transactions on Networking, Vol.7, Nº3: 387-397, 1999.
13. G. Cortese et. al, "CADENUS: Creation and Deployment of End-User Services in Premium IP Networks", IEEE Communications Magazine, ISSN: 0163-6804, pp.54-60, January 2003.
14. TAPAS IST Project (Trusted and QoS-Aware Provision of Application Services) *http://www.newcastle.research.ec.org/tapas/*
15. J.F. Meyer, "Performability Evaluation of Telecommunication Networks", in M. Bonatti, editor, Teletraffic Science for Cost-Effective Systems, Network and Services, ITC-12, pp. 1163-1172, Elesevier Science Publishers B. V. (North Holland), 1989.
16. E. de Souza e Silva, R. M. M. Leao, M. Diniz, "Transient Analysis Applied to Traffic Modeling". ACM SIGMETRICS Performance Evaluation Review, Vol. 28, I. 4, 2001.
17. B. Sericola, "Transient Analysis of Stochastic Fluid Models", Publication interne nº 1099, INRIA, Campus de Beaulieu, 35042 Rennes Cédex, France, April 1997.
18. J. F. Meyer, "Performability of an Algorithm for Connection Admission Control", IEEE Trans. on Computers, Vol.50, Nº 7: 724-733, 2001.
19. O. Rose, "Simple and efficient models for variable bit rate MPEG video traffic", Performance Evaluation, Vol. 30, Nº1-2: 69-85, 1997.
20. B. R. Haverkort, R. Marie, G. Rubino, K. Trivedi, "Performability Modelling. Techniques and Tools", John Wiley & Sons, ISBN: 047149195-0, 2001.
21. R. Vallejos, J. Martinez, "A Simple Method for Evaluating the Probability Distribution Function of Cumulative Operational Time in Repairable Systems", Fifth International Workshop on Performability Modeling of Computer and Communication Systems (PMCCS 5), pp. 61-66, Erlangen Germany, September 2001.