A New Adaptive Multimedia Streaming System for All-IP Multi-Service Networks

Gabriel-Miro Muntean, *Student Member, IEEE*, Philip Perry, *Member, IEEE*, and Liam Murphy, *Member, IEEE*

***Abstract*—A significant challenge in all-IP multi-service networks is to balance the goal of providing high-quality services to the end-users with the desire to maximize the number of end-users that can be simultaneously served. This paper presents a solution to this challenge by using the Quality-Oriented Adaptation Scheme (QOAS) for delivering multimedia streams. This adaptive mechanism uses feedback from clients regarding the quality of delivery to assist the server in making dynamic adjustments to the transmitted streams. Experimental objective and subjective test results illustrate the significant performance improvements achieved by QOAS, both in terms of number of simultaneous viewers served and of end-user perceived quality.**

***Index Terms*—Adaptive multimedia streaming, all-IP multi-service networks, dynamic feedback, grading scheme, statistical multiplexing.**

I. Introduction

INFORMATION in different forms (text, data, voice, video, etc.) is widely accessible through the Internet, wireless devices and, more recently, interactive TV systems. But a major change in the way information and entertainment are delivered to consumers is still to come in the form of on-demand-based access to rich media and full-motion very high quality video. Many cable operators have already upgraded their networks by introducing fiber into their systems allowing them to provide high-quality Video On Demand (VOD) services, while high-speed data service providers have constantly increased their share of the market. Some new services that require increased resources have already been launched, while others, such as interactive content and VOD, are waiting for large-scale deployment.

The success or failure of all these new services depends on widespread market acceptance, which, in turn, is heavily reliant on the price the end-user must pay. The companies involved in this area have pursued several avenues to reduce infrastructure costs and maximize the number of customers that can be serviced from a finite infrastructure. It now appears that the existing Hybrid Fiber Coax (HFC) networks will evolve toward an all-IP architecture [1] that would allow the use of popular IP-based applications and low cost hardware. The resulting reduction in operational costs will enable high market penetration with substantial revenues. Perhaps the greatest technical challenge is to devise systems that can increase the number of users per unit bandwidth, while maintaining a good level for the quality of service.

The multimedia streaming solution proposed here, denoted Quality-Oriented Adaptation Scheme (QOAS), makes use of a dynamic feedback-based adaptive mechanism that, in conjunction with the classic statistical multiplexing approach, allows for a significant increase in the number of simultaneous clients supported by a given delivery network in comparison to the nonadaptive case. This is done while maintaining a high end-user perceived quality for all the customers. The basic adaptive scheme, introduced in [2], takes into consideration server, network, and client-related problems that may negatively affect multimedia streaming by causing Periods of Unpredictable Delay and Loss (PUDLs). The QOAS reacts to PUDLs by adapting the transmitted quantity of data—and hence the quality of the multimedia stream—to the delivery conditions in order to maximize the viewers' perceived quality. This is enabled by the fact that the end-user perceived quality is regularly monitored and considered as an active factor in the QOAS-based adaptation mechanism, increasing its effectiveness.

The architecture of an all-IP multi-service delivery network suitable for the deployment of the QOAS-based adaptive multimedia streaming mechanism is presented next. Some existing adaptive solutions are then described and their limitations indicated. QOAS is described in detail in a separate section, followed by the results of both objective and subjective experimental tests performed in normal and loaded conditions. These results are presented in order to demonstrate the performance and the benefits of the QOAS-based adaptive system. QOAS applicability considerations and performance analysis are presented next, before conclusions are drawn and future work directions are indicated.

II. All-IP Broadband Multi-Service Delivery Systems

Different architectures for distribution of information in different forms to customers have been proposed in the past. Solutions for delivering multimedia-based services were first proposed for cable-TV or telephone infrastructures [3], [4] and lately were revised and extended in order to address broadband connectivity and target broadband all-IP networks [5], [6].

Totally centralized and pure distributed architectures have significant advantages and disadvantages that are in general balanced by hybrid solutions [7]. Such hybrid solutions balance the pressure placed on the IP-backbone and the high complexity of the multi-service distribution server in centralized approaches with the cost of maintenance for multiple less

Manuscript received December 13, 2002; revised January 15, 2004. This work was supported by the Research Innovation Fund of Enterprise Ireland.

G.-M. Muntean and P. Perry are with the School of Electronic Engineering, Dublin City University, Dublin 9, Ireland (e-mail: munteang@eeng.dcu.ie; perryp@eeng.dcu.ie).

L. Murphy is with the Computer Science Department, University College Dublin, Dublin 4, Ireland (e-mail: Liam.Murphy@ucd.ie).

Digital Object Identifier 10.1109/TBC.2004.824745

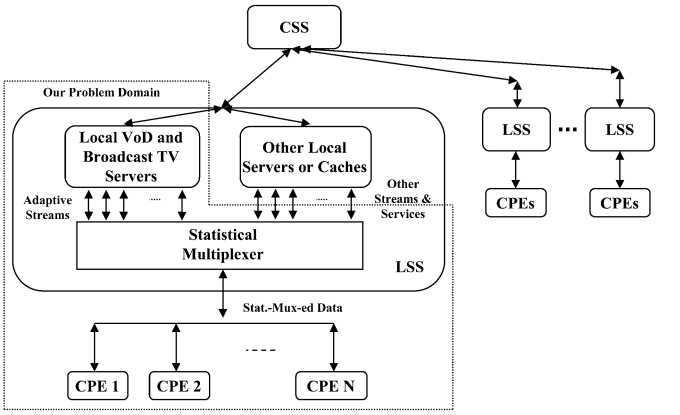

Fig. 1. A hybrid all-IP multi-service system including the adaptive multimedia streaming system.

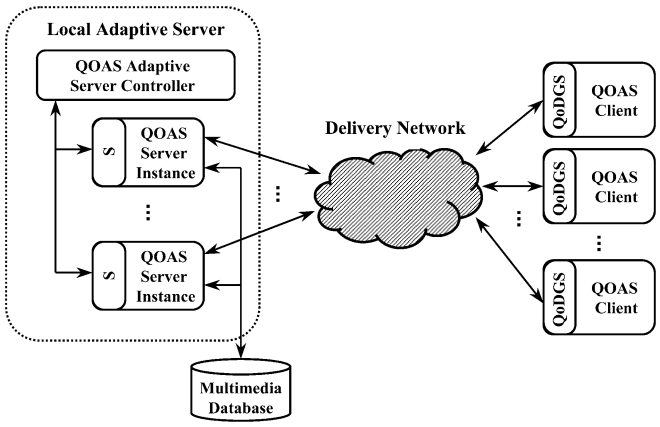

Fig. 2. The architecture of the QOAS-based multimedia delivery system.

complex servers and updates for their associated databases involved in distributed solutions.

A. A Hybrid All-IP Multimedia Delivery System

A hybrid all-IP multi-service delivery system, similar to the one presented in [5] for VOD only, was considered in this paper. It consists of a Centralized Service Server (CSS), Local Service Servers (LSS), a Distribution Network (DN) and the Customer Premises Equipment (CPEs). Data may be stored both in the CSS and in the LSS depending on the operators' cost and performance requirements. This is based on the assumption that an optimally designed distributed system can achieve lower cost than a centralized one [3]. The architecture of the all-IP multi-service delivery system considered and the location of the proposed adaptive mechanism for multimedia streaming are illustrated in Fig. 1.

The first LSS shows the deployment of the proposed adaptive multimedia delivery solution on a group of servers whereas the other servers that belong to the same LSS do not have any adaptation mechanism. The focus in this paper has been on studying the behavior of the adapted multimedia streams during a delivery process that includes statistical multiplexing. The influence of nonadaptive servers is therefore modeled as background traffic.

B. Existing Solutions

The main goal of this research is to increase the efficiency of multimedia delivery while maximizing the utilization of the network resources.

The work presented in [8], extending early work reported in [9], proposes a system that creates an adaptive mechanism that includes both the statistical multiplexer and the MPEG-2 encoders. The system uses either the information received directly from the multiplexer in a feedback-based solution or some statistical information saved by the encoders during a look-ahead phase. Based on this information the outputs of the encoders are matched to the statistically available bandwidth.

Existing commercial solutions for real-time optimal statistical multiplexing proposed by Cisco: Cisco 6920 RateMUX [10] and Harmonic Inc.: DiviTrackXE [11] use complex processing for input streams rate adaptation (re-quantization, re-encoding [10] or look-ahead processing [11]). Their cost and the small number of inputs (up to 15 or 24, respectively) make them unsuitable for large-scale deployment. It is also significant to mention that none of the previously proposed mechanisms consider end-user perceived quality in the adaptive process, although its maximization should be the goal of any multimedia delivery system.

The new solution proposed in this paper and described in detail in the next section is the Quality-Orientated Adaptation Scheme (QOAS) for multimedia streaming. QOAS is much simpler to deploy and involves little additional cost. It is also demonstrated that by using QOAS, the efficiency of link utilization increases and a greater number of simultaneous clients are allowed to share the same bandwidth. At the same time the overall end-users' viewing satisfaction increases because estimations of end-users' perceived quality are actively taken into account during the adaptation process.

III. Quality Oriented Adaptation Scheme (QOAS) for Multimedia Streaming

During nonadaptive multimedia streaming end-user perceived quality could be reduced due to server-related problems (e.g., server load, machine, software), network-related problems (e.g., congestion, extremely variable traffic, equipment failures) and/or client-related problems (e.g., slow or incompatible software, old hardware). These problems directly or indirectly cause some Periods of Unpredictable Delay and Loss (PUDLs) that affect the overall quality of delivery. The proposed QOAS reacts to these PUDLs to try to maximize the end-user perceived quality in existing network conditions. It increases or decreases the transmitted quantity of multimedia data by dynamically adjusting the quality of the streamed multimedia. If the adaptation is performed while maintaining the continuity of the streaming process and the quality is varied in a controlled manner, the end-users benefit in terms of their perceived quality [12].

A. QOAS-Based System Architecture

Fig. 2 shows a local system for QOAS-based adaptive multimedia delivery. It involves a server-located QOAS controller ap-

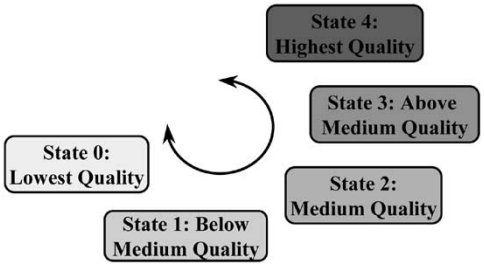

Fig. 3. Different quality versions of the same multimedia content associated with IA-QOAS server states.

plication and multiple instances of feedback-controlled QOAS client and server applications.

These client and server applications implement the QOAS-based multimedia delivery mechanism, allowing for a coarse adaptation process that involves only the delivery of the current stream. The QOAS client application uses a Quality of Delivery Grading Scheme (QoDGS) to assess the delivery quality and a feedback mechanism to inform the server application about this. The QOAS server application uses a Server Arbitration Scheme (SAS) to filter the received quality reports based on which it takes adaptive decisions. This quality adaptive process that involves only a single delivery process is denoted IntrA-stream QOAS-based adaptation (IA-QOAS).

The QOAS Adaptive Server Controller (ASC) application which is in permanent contact with all the IA-QOAS server application instances makes fine adjustments to all the adaptation processes by looking at the delivery process globally and improves the link utilization. This process is denoted InteR-stream QOAS (IR-QOAS) adaptation.

B. IntrA-Stream QOAS

IA-QOAS adjusts the delivery of a single stream in reaction to PUDLs by varying its quality and consequently the quantity of transmitted data, regardless of the evolution of other streams under delivery. This requires extra storage space at the server for the pre-recorded case [13] and an extra processing stage for live streaming [14].

IA-QOAS adaptation requires the definition of a number of different quality versions for each multimedia stream. Each version is then associated with an IA-QOAS server state, as in the five-state example shown in Fig. 3. During transmission the server varies its state according to the received stream quality at the client. A feedback mechanism that takes advantage of the fast network delivery infrastructure informs the server about the end-user's perceived quality of service and allows it to take the necessary adjustment decisions.

The different quality versions are chosen such that they are highly graded on the subjective testing scale standardised in the ITU-T R. P.910 [15] and widely accepted in the engineering community (see Table I). The no-reference moving picture quality metric (Q) proposed in [16] is used both for choosing the operating region for the adaptive scheme and for assessing the effect of the proposed scheme on the client perceived quality. More details about this are presented in the next subsection.

TABLE I
ITU-T R.P. 910 QUALITY SCALE FOR SUBJECTIVE TESTING

Rating	Impairment	Quality
5	Imperceptible	Excellent
4	Perceptible, not annoying	Good
3	Slightly annoying	Fair
2	Annoying	Poor
1	Very annoying	Bad

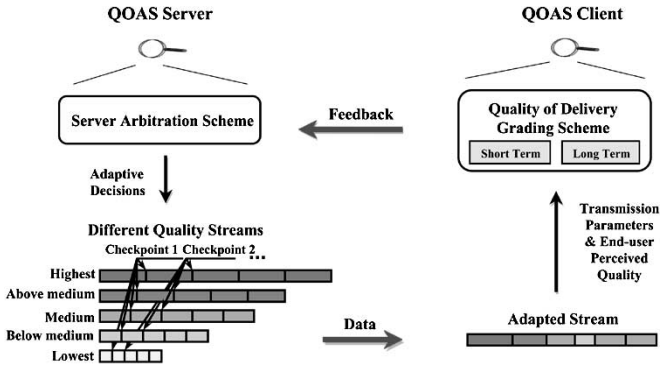

Fig. 4. Schematic description of IA-QOAS adaptation principle.

Fig. 4 describes graphically the principle behind IA-QOAS. Multimedia data is uni-directionally transmitted to the client which both monitors transmission parameters and estimates end-user perceived quality during the streaming process. These are performed by the Quality of Delivery Grading Scheme (QoDGS) whose functionality is described in Section III-E. QoDGS grades the overall quality of multimedia streaming in terms of quality scores that are regularly sent as feedback to the server. The Server Arbitration Scheme (SAS), which is described in detail in Section III-F, analyzes these scores and suggests adaptive decisions to be taken by the server in order to maximize the end-user perceived quality in existing delivery conditions.

The IA-QOAS-based adaptation process achieves good adaptation to the available bandwidth, but due to the limited number of pre-defined quality versions of the streams taken into consideration, can result in sub-optimal utilization of the delivery network.

C. End-User Perceived Quality Assessment and IA-QOAS Operating Region Selection

Different factors may affect the end-user perceived quality of the multimedia streams, including the IA-QOAS-related quality adaptations. Therefore there is a need to quantify streaming quality, affected both by bitrate variations and packet losses, in order to determine the right balance between the server adaptations and end-user perceived quality.

For assessing the end-user perceived quality the no-reference moving picture quality metric (Q) proposed in [16], which describes the joint impact of MPEG bitrate and data loss on video quality, was used. Equation (1) presents the formula for Q, where PLR is the packet loss ratio, $\bar{R}$ is the stream's mean bitrate and the constant Q_0 has a value close to the maximum

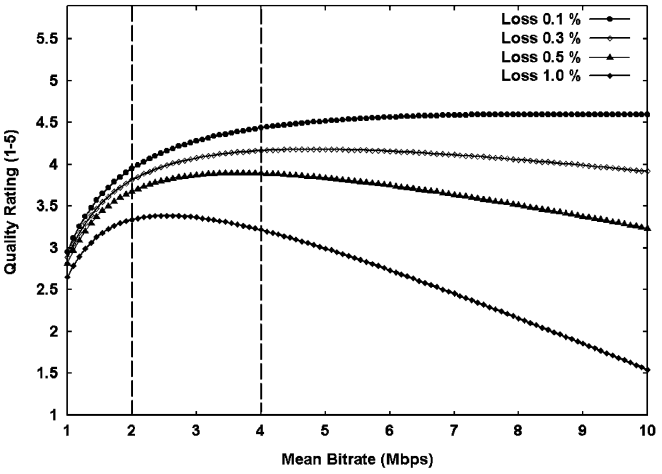

Fig. 5. Estimated end-user quality (Q) variation with mean bitrate for a multimedia stream with average motion content, plotted for different packet loss ratios.

quality 5. $\chi_{\rm q}, \xi_{\rm r}$, and $\chi_{\rm r}$are constants related to the complexity of the sequence, whereas $\chi_{\rm l}$ depends also on the average bitrate.

$$Q = Q_0 + \chi_q * \left(\frac{\bar{R}}{\chi_r}\right)^{-\frac{1}{\xi_r}} + \chi_l * \bar{R} * \mathrm{PLR} \qquad (1)$$

The curves in Fig. 5 show the variation of Q with mean bitrate for packet loss ratios between 0.1% and 1.0% when using average values for parameters related to the stream's complexity suggested in [16]. When the loss ratio increases, the end-user perceived quality decreases, therefore the IA-QOAS policy of reducing the transmitted stream quality—and consequently its bitrate—during congested periods may result in loss reduction and thus improved end-user perceived quality. In normal traffic conditions, characterized by low loss ratios, any transmitted stream quality upgrades yield increases in the perceived quality.

Since for very low loss rates (less than 0.1%), the benefit in the perceived quality with the increase in the stream bitrate above 4 Mbps (and consequent bandwidth consumption) is not significant, the higher limit of interest in this paper for the encoding rate was chosen to be 4 Mbps. Encoding multimedia below 2 Mbps makes the perceived quality drop below the "good" level even in very good delivery conditions, and therefore 2 Mbps was selected as the lower limit of interest. Since experimental testing was performed with MPEG-2 encoded streams with bitrates between 2 Mbps and 4 Mbps, the IA-QOAS's corresponding operating region is delimited in Fig. 5 by dashed lines.

D. Inter-Stream QOAS

The IR-QOAS implemented by the Adaptive Server Controller (ASC) application adjusts the overall adaptation process to yield better utilization of network resources.

The IR-QOAS is also responsible for preventing the IA-QOAS-based processes from reacting simultaneously to variations in the delivery conditions. Such synchronization may trigger IA-QOAS over-reaction resulting in both under-usage of the available bandwidth and reduced perceived quality for the remote viewers. Therefore the ASC application selects some of the multimedia sources to react to the received feedback, achieving near optimal link utilization and long-term fairness between the clients.

Based on the history of all the IA-QOAS-based delivery processes, the ASC application estimates the available bandwidth and the transmission conditions. According to the latter it regularly updates its working state between the two defined values: "normal" and "congested". If all the IA-QOAS-based multimedia deliveries are performed at maximum quality, the ASC state is set to "normal" and there is no interference between IR-QOAS and IA-QOAS processes. If some IA-QOAS-based streaming processes have adjusted downwards their transmission quality (and consequently have decreased their IA-QOAS server state), which suggests that there are some delivery problems, the ASC state is set to "congested". In these conditions IR-QOAS may interfere with individual IA-QOAS-based streaming processes.

In the "congested" state, the IR-QOAS can influence IA-QOAS processes on three occasions: during initialization, when a streaming process has ended, and when any IA-QOAS adaptive decision is taken.

Since the initial transmission rate of a requested multimedia stream is very important, the controller specifies a starting quality version during the initialization stage. This initial stream quality state is computed as the average of the states of the other streams currently being delivered. In order to prevent losses from occurring, the server controller forces a quality reduction on some of the streaming sources that are in a higher quality state than the average.

A similar "imposed" adaptation process is performed when any IA-QOAS streaming process ends and some of the streams being delivered at a lower quality than the average will benefit from a quality state increase. The number of streams that will be affected by the forced adaptation is determined from an estimation of the available bandwidth, the number of the existing streams and their quality.

In order to reduce the synchronization between the IA-QOAS-based streaming processes, IR-QOAS spreads their adaptive reactions over a period of time, introducing random delays in their adjustment decision processes. If the feedback reports received by the IA-QOAS-based streaming processes do not indicate an improvement in the quality of delivery when some IA-QOAS processes have scaled back their transmissions, the downgrading in the quality of the streamed multimedia will continue. However if the delivery situation improves and the IA-QOAS-s that have requested downwards adjustments in their streamed quality send positive feedback reports, no further quality decreases are required. A similar process of avoiding synchronization occurs when the IA-QOAS-based streaming processes request increases in their streamed multimedia quality.

E. Quality of Delivery Grading Scheme

One of the most important components of the IA-QOAS mechanism is the client-located Quality of Delivery Grading Scheme (QoDGS) whose block-level architecture is presented in Fig. 6. QoDGS extends the grading scheme described in [2]

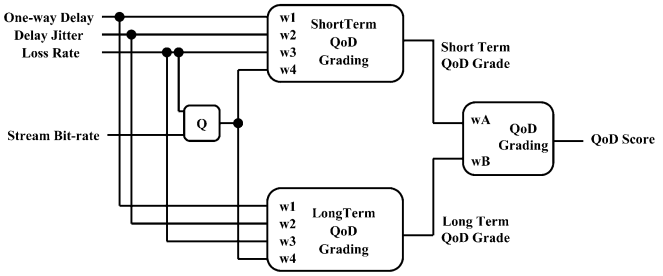

Fig. 6. The QoDGS takes into consideration traffic-related parameters and end-user perceived quality.

by taking into account end-user perceived quality in the grading process. It monitors some transmission related parameters (such as the number of packets lost or that arrived too late for play-out, packet delay, and delay jitter), and estimates the end-user perceived quality.

QoDGS regularly grades the received streams' Quality of Delivery (QoD) by taking into consideration the relative importance of each monitored parameter and the special characteristics of each transmission session. Short-term variations of parameters are monitored in order to learn quickly about sudden problems that may affect the quality of delivery, whereas long-term monitoring considers the effect of slow changes in the delivery conditions and introduces a degree of stability in the grading algorithm. In both cases, partial scores that reflect the values and the variations of the monitored parameters are determined and then used as shown in (2) and (3) to compute short-term $(\mathrm{QoD_{ST}})$ and long-term $(\mathrm{QoD_{LT}})$ QoD grades in the second stage of the QoDGS.

Finally, in the third stage, the computed $\mathrm{QoD_{ST}}$ and $\mathrm{QoD_{LT}}$ scores are weighted according to their relative importance and combined to determine the overall quality of delivery score $\mathrm{QoD_{Score}}$, as shown in (4). These $\mathrm{QoD_{Score}}$-s are regularly sent to the server in the feedback messages.

$$\mathrm{QoD_{ST}} = w_1 * \mathrm{DelayGrade_{ST}} + w_2 * \mathrm{JitterGrade_{ST}} + w_3 * \mathrm{LossGrade_{ST}} + w_4 * \mathrm{QGrade_{ST}} \quad (2)$$

$$\mathrm{QoD_{LT}} = w_1 * \mathrm{DelayGrade_{LT}} + w_2 * \mathrm{JitterGrade_{LT}} + w_3 * \mathrm{LossGrade_{LT}} + w_4 * \mathrm{QGrade_{LT}} \quad (3)$$

$$\mathrm{QoD_{Score}} = w_A * \mathrm{QoD_{ST}} + w_B * \mathrm{QoD_{LT}} \quad (4)$$

In order to determine the weights in (2)–(4), extensive tuning was performed taking into account different types of multimedia streams and with different degree of motion content. The best results in terms of adaptiveness, responsiveness to traffic variations, stability, shared link utilization and end-user perceived quality were obtained for $\mathrm{w_1} = 0.4, \mathrm{w_2} = 0.3, \mathrm{w_3} = 0.2, \mathrm{w_4} = 0.1, \mathrm{w_A} = 0.75$, and $\mathrm{w_B} = 0.25$.

F. Server Arbitration Scheme

Another major component of the IA-QOAS is the Server Arbitration Scheme (SAS). SAS takes into account feedback reports from the client and, in order to minimize the effect of noise in the QoD scores, it bases its adaptive decisions on an arbitration process.

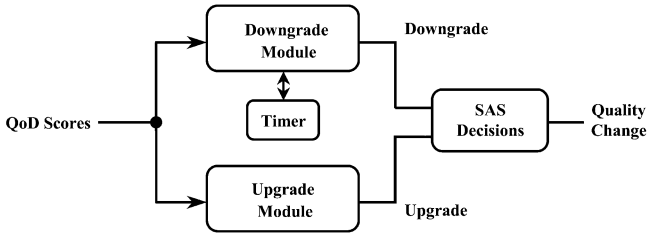

Fig. 7. Server Arbitration Scheme.

Fig. 7 presents the block-level architecture of the SAS which includes the Downgrade and Upgrade modules, the SAS Decisions module and a Timer. Since SAS considers the values of a number of recent feedback reports, these received $\mathrm{QoD_{Score}}$-s are stored in different length sliding windows. The Downgrade and Upgrade modules are similar, the difference being the time scale on which quality adjustments are suggested and therefore the length of associated sliding windows. The average values of the most recently received QoD scores encompassed by these windows are compared with the current server quality state that determines the quality of the streamed multimedia clip (see Section III-B). This comparison allows the Decisions Module to take decisions regarding upgrades and downgrades in the transmitted quality.

The SAS-based arbitration process is *asymmetric*, requiring fewer feedback reports to trigger a decrease in quality than for a quality increase. This ensures a fast reaction during bad delivery conditions, helping to eliminate the cause of the PUDL that is reducing the end-user perceived quality. SAS response to improved delivery conditions is slow, allowing the distribution network to recover after PUDLs. The asymmetric arbitration process ensures both system stability (by minimizing the number of quality variations) and fast adaptive reactions to PUDLs, if and when they are necessary.

Since the deployment of QOAS is envisaged for broadband multi-service all-IP distribution networks where IP traffic will be the only traffic carried, SAS considers the late arrival of feedback messages as an indication of network congestion which leads to decisions involving decreases in quality.

IV. Experimental Results

In order to test the performance of QOAS, both a simulation model and a prototype system were designed. The simulation model, built using Network Simulator version 2 [17], was used to test QOAS behavior and performance in the presence of other similar processes. The QOAS prototype system, built using Microsoft Visual C++ 6.0 and tested in a Win32 environment, was mainly used for user-perceptual testing.

A. Simulation Topology, Model, and Multimedia Clips

The goal of the simulation tests is to show the significant increase in the number of clients that can be simultaneously served when using QOAS in comparison to using a nonadaptive solution. Another goal is to demonstrate that the user perceived quality does not significantly decrease during this adaptive process. A modest network capacity was chosen so that the simulations could be performed within reasonable times. It is

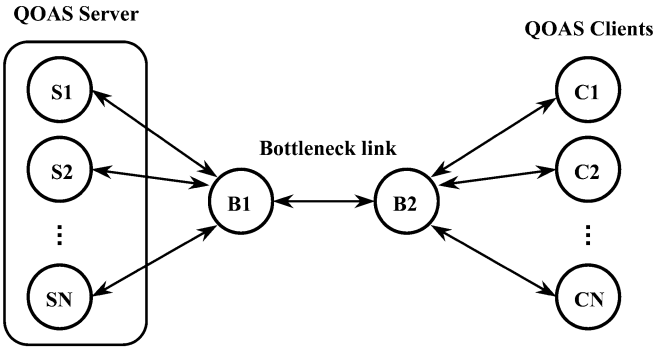

Fig. 8. Simulation topology which includes a QOAS adaptive server that communicates with N QOAS adaptive clients via a bottleneck link.

TABLE II
STATISTICS FOR DIFFERENT MPEG-2 ENCODED SEQUENCES FOR AN AVERAGE ENCODING RATE OF 2.5 MBPS

Sequence	Motion content	Peak Rate (bits/frame)	Peak /Mean Ratio
diehard1	High	860648	7.42906
jurassic3	Average-low	447528	4.37717
dontsayaword	Average	480840	4.51028
familyman	Low	322968	3.16777
roadtoeldorado	Average-high	693696	6.50675

expected that the results will scale to Gigabit systems without losing their essential characteristics.

The "Dumbbell" simulation topology shown in Fig. 8 consists of a number of QOAS clients connected to a QOAS server by a bottleneck link that has a bandwidth of 70 Mbps and a 0.1 sec propagation delay. A router with a drop-tail queue statistically multiplexes the data exiting the adaptive server onto the link. As mentioned earlier, the QOAS Adaptive Server Controller application controls all these QOAS server application instances. The length of the drop-tail queue was chosen after tests with different queue sizes and is proportional to the product of round trip time and bandwidth. If this queue was very small, even an adaptive system could not prevent loss that affects end-user perceived quality. If this queue were double the simulated size, both delays and delay jitters would increase significantly.

Each client is simulated by an instance of the QOAS client application. These clients randomly request movies from the QOAS server without taking into account the movies' popularity. QOAS server application instances then adaptively deliver the chosen movies to the clients.

Five five-minute long multimedia sequences were selected from movies with different degrees of motion content. The *diehard1* sequence includes a great deal of action, *jurassic3* and *dontsayaword* have an average amount of action, whereas *familyman* has very little movement in it. The *roadtoeldorado* sequence is from an animated cartoons movie. As shown in Fig. 3, five states were defined in the server adjustment space. Each state is associated with a different average encoding rate for the selected multimedia sequences. Each stream was then MPEG-2 compressed at 2.0 Mbps, 2.5 Mbps, 3.0 Mbps, 3.5 Mbps and 4.0 Mbps respectively and traces from the resulting files were used as streaming sources in the simulations. Some statistical characteristics of the 2.5 Mbps versions of the five sequences used are presented in Table II, while a comparison of the statistical information related to the *diehard1* sequence is presented in Table III.

TABLE III
STATISTICS FOR DIFFERENT ENCODED VERSIONS OF *DIEHARD1* SEQUENCE

Quality State (0-4)	Encoding Rate (Mbps)	Mean Rate (bits/frame)	Peak Rate (bits/frame)	Peak /Mean Ratio
0	2.00	92717.7	693872	7.48370
1	2.50	115848.8	860648	7.42906
2	3.00	135526.2	854536	6.30532
3	3.50	152440.4	860648	5.64580
4	4.00	170882.6	693872	4.06052

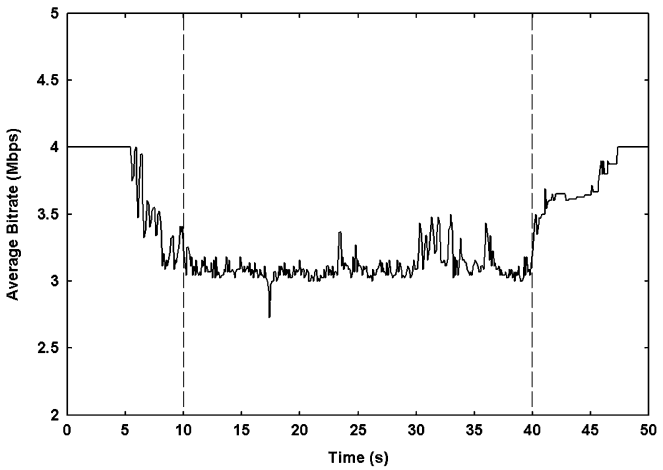

Fig. 9. Variation of the average server transmission rate during the adaptive delivery of 22 streams.

The QOAS model conforms to the description in Section III, with a server arbiter upgrade period of 6 sec and a downgrade timeout of 1 sec. The QoDGS short-term period was taken as 1 sec, and the long-term period was 10 sec.

B. Normal Loading Conditions

The simulations lasted 50 sec and involved the delivery of multiple multimedia streams over the bottleneck link using QOAS. The 50 sec multimedia sequences were randomly selected from within the five minute clips. The clients requested the movies at different times during an initial transitory period. There is a similar period at the end of simulation when the clients exit the system and the streaming stops. The fully loaded condition between these transitory periods is the area of interest, as marked for example in Fig. 9. This trace shows the variation in average transmission bit-rate of the delivered streams during tests with 22 QOAS-based adaptive clients.

For the case of 22 adaptive clients, a loss rate of 0.035% was observed. In the nonadaptive case, however, this loss rate was reached with just 16 clients, indicating that 37.5% more clients could be accommodated with a QOAS-based adaptive system. Since in these conditions all five QOAS server quality states are graded at least "good" on the ITU-T five-point perceptual quality scale presented in Table I, this shows a significant improvement in system efficiency while maintaining a high user perceived quality. Another important result is that the utilization of the bottleneck link is very high, reaching 99.9% during the QOAS-based transmissions. It is also important to

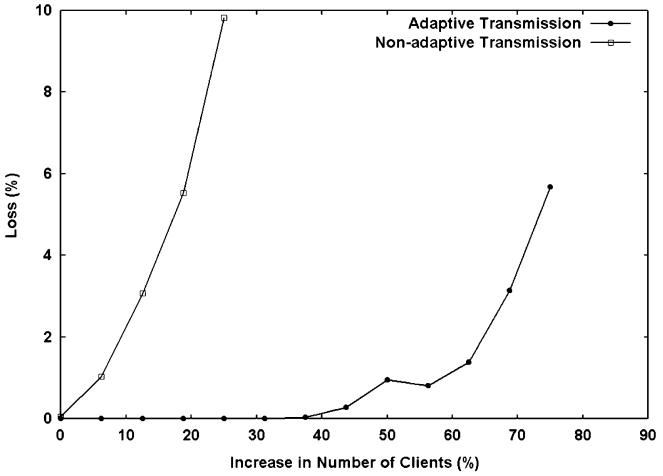

Fig. 10. Comparison of the loss rate caused by increasing the number of clients above a base line of 16 for adaptive and nonadaptive transmissions.

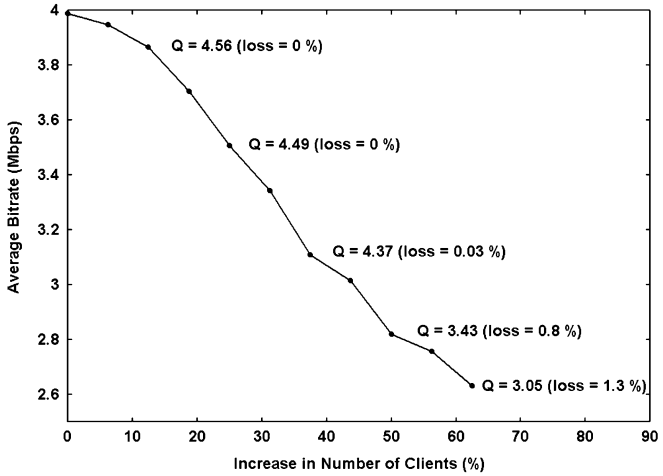

Fig. 11. The average bit-rate versus the increase in the number of clients above the baseline of 16 during QOAS-based adaptive streaming.

note that there is a very small quality variation between the different streams, indicating that good inter-stream fairness has also been achieved by using QOAS.

An important characteristic of this QOAS-based adaptive delivery scheme is that it permits a choice of the optimal operational point, according to economic, technical and quality goals, as indicated in Section III-C. It seems likely that companies with all-IP multi-service networks can maximize their revenues from VOD services by increasing the number of customers while maintaining an acceptable quality of delivery. Scaling the results obtained here from a 70 Mbps link to a fully deployed network would show that a single gigabit Ethernet connection could service 314 users with the QOAS-based adaptive solution compared to only 228 with a nonadaptive system.

C. Severe Loading Conditions

Next the behavior of the adaptive QOAS-based and nonadaptive systems under severe loading conditions is examined. The graph shown in Fig. 10 compares the average loss rate when the number of clients is gradually increased above the base line of 16. An increase of only 6% in the number of clients causes a loss over 1% in the nonadaptive case, and the loss rate reaches 10% when the number of clients has been increased by 25%. When QOAS is used, an increase of up to 40% in the number of clients has very little effect on loss rate. Increases of up to 60% result in a loss rate under 1%, which may be overcome by using post-processing techniques.

This substantial increase in the number of simultaneous clients while maintaining a very low loss rate comes with a voluntary and controlled degradation in the delivered quality. This degradation is shown in Fig. 11. A 12.5% increase in the number of clients yielded an average server state of 3.73, corresponding to a mean bit-rate of 3.86 Mbps and to a client-perceived quality rating of 4.56 out of 5.0. A 37.5% increase in the client population gave an average server state of 2.22, which corresponds to a client-perceived quality score of 4.37. Increases above 50% resulted in the client-perceived quality dropping below 4.0-considered here to be the lowest limit of interest, but perhaps acceptable for other transmission scenarios.

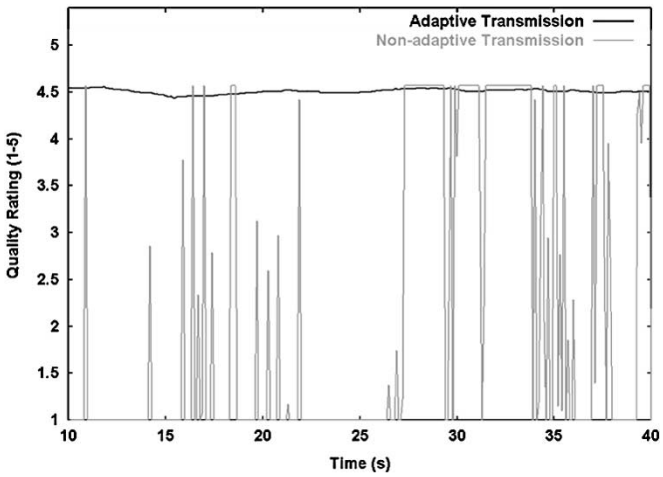

Fig. 12. The variation of the average perceived quality while concurrently transmitting 19 streams using the QOAS-based adaptive mechanism and nonadaptive solution.

The plot in Fig. 12 shows the variation of the perceived quality metric during a period when the delivery system was loaded with 19 users. It shows that the voluntary degradation in the quality of the QOAS streams does not significantly influence client perceived quality. The value of the metric remained better than "good" on the perceptual quality scale during the simulated period, with an average of 4.51. It is important to note that under the same network conditions, the nonadaptive delivery mechanism resulted in a loss rate that at times exceeded 10% and a perceived quality score below the level when impairments become annoying, severely affecting the perceived quality. During this transmission the perceived quality average was 2.10 and varied noticeably, both characteristics which negatively affect viewers.

The contrast is even clearer with an increasing number of simultaneous clients. For example, during the QOAS streaming with 23 clients, the average user perceived quality is lower than with 21 clients (Fig. 13), but still rated as 4.31 out of 5.0. In similar conditions, during nonadaptive transmission the average user perceived quality drops to 1.05, which represents "bad" user perceived quality on the ITU-T scale.

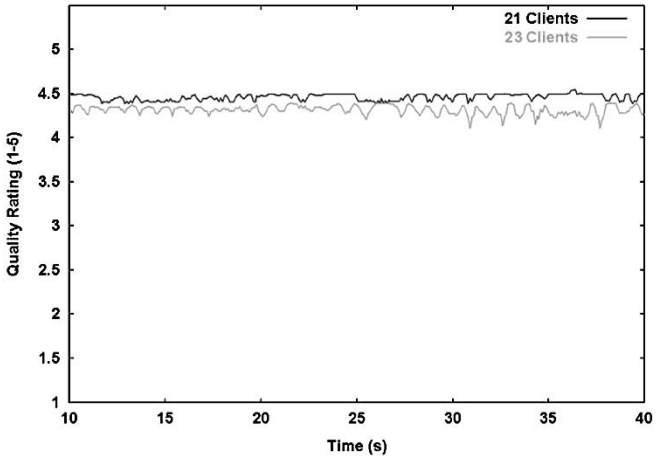

Fig. 13. The variation of the average end-user perceived quality while transmitting streams using the adaptive scheme.

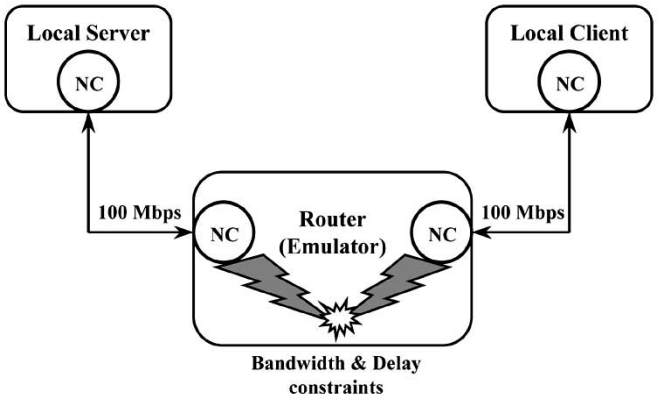

Fig. 14. Testbed setup consisting of a server and a client on different networks interconnected by a router on which a network emulator allows for bandwidth and delay variation.

D. User Perception-Based Test Results

The previous results showed that QOAS scored highly in terms of the objective no reference moving picture quality metric [16]. Subjective tests have also been carried out in order to confirm QOAS performance from this point of view. The goal of the user-perception tests is to establish the benefit of using the QOAS-based adaptive scheme and to determine whether the users are affected by the resulting slight variations in their perceived quality.

The tests involved the prototype QOAS system and five different versions of the same pre-recorded multimedia content selected from the *diehard1* movie with high motion content. The testbed presented in Fig. 14 was used during the subjective testing. It consists of a server and a client connected to different networks and a router that interconnects these networks. QOAS applications were deployed at both the server and the client. Congested network conditions were simulated using the Nistnet network emulator [18] deployed on the router. The QOAS server application switched its state in real-time, by changing which quality versions of the *diehard1* multimedia sequence it transmitted, causing the quality of the multimedia stream to vary. As a result the end-user perceived quality varied slightly.

Preliminary tests involved forty subjects who were already accustomed to remote multimedia streaming. Results showed that 82.5% of them liked the continuity provided by the proposed adaptive scheme, and although 46.3% noticed slight quality variations they did not find them disturbing. In spite of some implementation-specific factors (e.g., no post-processing algorithms to enhance the displayed quality were deployed), it seems reasonable to conclude that the variations introduced by the QOAS are not disturbing for the viewers.

Other perceptual test results are described in detail in [13].

V. Applicability Considerations

QOAS relies on feedback in order to learn about the quality of the streaming process. The results of research such as [7], [19] which studied feedback for performing quality adaptations show that the faster the feedback messages arrive at the server, the better the results of the adaptation process. Therefore QOAS is most strongly recommended for local or metropolitan area networks, local cable IP networks, or local all-IP broadband networks, where **fast feedback** is feasible.

The application of any adaptive scheme, including QOAS, is most recommended in **networks with a potential for congestion**. This is because adaptive schemes offer significant benefits in comparison to a nonadaptive approach only if shared resources are limited, even if only for certain periods of time.

It is crucial that the **viewers** and the **applications** targeted by QOAS are able to **tolerate a certain degree of quality variation.** Some multimedia systems viewing quality has life-threatening or precision-related consequences such as in some areas of Medicine (e.g., Surgery), Physics (e.g., atomic phenomena) or Transport (e.g., Radar systems). Therefore QOAS would most likely be applied in the entertainment industry, video-on-demand business applications, commercial presentations, and video-conferencing in which a slight decrease in play-out quality is much preferred to the buffering interruptions performed by many existing solutions.

VI. Performance Analysis

For streaming live content there are no extra storage requirements for the deployment of a QOAS-based solution. However, for the distribution of pre-recorded multimedia streams, QOAS trades bandwidth for the storage space required in the server's multimedia database. More than one quality version has to be encoded and stored for each multimedia stream. For instance, for the five different quality versions of the *diehard1* five-minute sequence presented in Table III, 562.5 MB are required instead of the 150 MB necessary to store only the highest quality stream, an increase of 275%. But during QOAS streaming, using the same distribution network with bandwidth 70 Mbps as in Section IV, 37.5% more customers could be served simultaneously than with a nonadaptive solution and each experiences good perceived quality. In order to serve 37.5% more customers with a nonadaptive system that transmits the highest rate stream all the time, the bandwidth has to be increased by 37.5% to 96 Mbps. Various reports [20], [21] show that bandwidth is still relatively more expensive than storage capacity. This makes a QOAS-based solution for multimedia distribution very attractive.

The significant advantages of a QOAS-based solution come with a cost in terms of extra processing requirements and some bandwidth used for feedback.

The fact that this processing is distributed among the QOAS clients whose QoDGS-s monitor and grade the quality of streaming at the receivers, significantly reduces the load of the QOAS server machine that runs only the SAS. The QOAS server has only to acquire the client transmitted $\mathrm{QoD}_{\mathrm{Scores}}$, to process them (which can be performed incrementally) and to take adaptive decisions (which do not involve excessive CPU load).

Regarding the feedback, it is significant to mention that each feedback report consists only of a $\mathrm{QoD}_{\mathrm{Score}}$. If RTCP packets are used, for standard values for the headers' sizes (20 Bytes-IP header, 8 Bytes-UDP header, 8 Bytes-RTCP receiver report packet header) and for a 4-Byte payload, the feedback packet size becomes 40 Bytes long. For a low inter-feedback transmission time of 0.1 sec the bandwidth used by feedback for a single client becomes $\mathrm{BW}_{\mathrm{feedback}} = 400$ Bytes/s. Since QOAS was designed for local broadband multi-service IP-networks, this represents an insignificant bandwidth usage. For example over 300 simultaneous customers on a gigabit Ethernet (as seen before in Section IV-B) consume only 0.1% of the available bandwidth for feedback.

VII. Conclusions and Future Work

The Quality-Oriented Adaptation Scheme (QOAS) for delivering high-quality multimedia streams over emerging all-IP network architectures has been presented. QOAS was compared to a nonadaptive solution for transmitting multimedia in the presence of a statistical multiplexer.

It was shown that by using QOAS the number of clients that can be served simultaneously from a finite bandwidth resource could be significantly increased. The increase is up to 40% in the absence of post-processing techniques and could be up to 60% if some post-processing techniques were used by clients to compensate for the resulting loss rates.

The simulation results also show that despite the large increase in the number of simultaneous clients for the adaptive case, the loss rate remained very small (0.035%), the link utilization is nearly optimal (99.9%) and the value of the end-user perceived quality metric has been maintained at high level (above the "good" level on the ITU-T five-point perceptual quality scale).

Preliminary perceptual tests with a prototype system were performed and their results validate those obtained via simulation. The test viewers both appreciated the continuity of the playout and were not disturbed by the slight quality variations introduced by QOAS.

QOAS is scalable to gigabit Ethernet with little additional cost and large potential benefit. Therefore the results presented here suggest that companies involved in multimedia delivery to home residences and business premises may benefit hugely from providing a QOAS-based solution.

Further subjective tests would be needed to validate the preliminary results presented here, and a practical QOAS prototype system would need to be developed in order to explore scaling and implementation issues.

References

[1] S. Dravida, D. Gupta, S. Nanda, K. Rege, J. Strombosky, and M. Tandon, "Broadband access over cable for next-generation services: A distributed switch architecture," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 116–124, Aug. 2002.

[2] G.-M. Muntean and L. Murphy, "A quality-aware adaptive multimedia streaming scheme," Submitted to IEEE Trans. on Circuits and Systems for Video Technology, http://www.eeng.dcu.ie/~munteang/articole/CSVT_MunteanMurphy.pdf, 2003.

[3] S. A. Barnett and G. J. Anido, "A cost comparison of distributed and centralized approaches to video-on-demand," *IEEE Journal on Selected Areas of Communications*, vol. 14, no. 8, pp. 1173–1183, Aug. 1996.

[4] Harmonic Inc., "Network and access architecture for on-demand cable television," *Cable Telecommunication Engineering Journal*, vol. 24, no. 1, Mar. 2002.

[5] E. W. M. Wong and S. C. H. Chan, "Performance modeling of video-on-demand systems in broadband networks," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 7, pp. 848–859, July 2001.

[6] J. Y. B. Lee, "On a unified architecture for video-on-demand services," *IEEE Transactions in Multimedia*, vol. 4, no. 1, pp. 38–47, Mar. 2002.

[7] G.-M. Muntean, "Quality-Oriented Adaptation Scheme for Multimedia Streaming in Local Broadband Multi-Service IP Networks," Ph.D. thesis, Dublin City University, Ireland, Sept. 2003.

[8] L. Böröczky, A. Y. Ngai, and E. F. Westermann, "Statistical multiplexing using MPEG-2 video encoders," *IBM Journal of Research and Development*, vol. 43, no. 4, 1999.

[9] M. Perkins and D. Arnstein, "Statistical multiplexing of multiple MPEG-2 video programs in a single channel," *SMPTE Journal*, pp. 596–599, Sept. 1995.

[10] Cisco Systems, "Statistical Multiplexing: Increased Efficiency, Flexibility, and Quality for MPEG-2 Video Applications," White Paper, http://www.cisco.com, Oct. 2000.

[11] Harmonic Inc., "DiviTrackXE—Advanced Statistical Multiplexing," White Paper, http://www.harmonicinc.com, 2002.

[12] D. Wu, Y. T. Hou, W. Zhu, Y.-Q. Zhang, and J. M. Peha, "Streaming video over the internet: Approaches and directions," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 3, pp. 282–300, Mar. 2001.

[13] G.-M. Muntean and L. Murphy, "Adaptive pre-recorded multimedia streaming," in *Proc. IEEE GLOBECOM 2002*, Taipei, Taiwan, 2002.

[14] ——, "Adaptive traffic-based techniques for live multimedia streaming," in *Proc. IEEE International Conference on Telecommunication*, vol. 1, Beijing, China, 2002, pp. 1183–1187.

[15] "Subjective Video Quality Assessment Methods for Multimedia Applications,", ITU-T Recommendation P.910, Sept. 1999.

[16] O. Verscheure, P. Frossard, and M. Hamdi, "User-oriented QoS analysis in MPEG-2 video delivery," *Journal of Real-Time Imaging*, vol. 5, no. 5, Oct. 1999.

[17] Network Simulator-2. [Online]. Available:http://www.isi.edu/nsnam/ns/

[18] NIST Net, http://snad.ncsl.nist.gov/itg/nistnet.

[19] R. Rejaie, "An End-to-End Architecture for Quality Adaptive Streaming Applications in the Internet," Ph.D. thesis, University of Southern California, Dec. 1999.

[20] Evolution of Gigabit Technology, Intel, 2001. White Paper.

[21] R. J. T. Morris and B. J. Truskowski, "The evolution of storage systems," *IBM Systems Journal*, vol. 42, no. 2, 2003.

Gabriel-Miro Muntean is a Lecturer with the School of Electronic Engineering, Dublin City University, Ireland, where he obtained his Ph.D. degree in 2003 for research on quality-oriented adaptation schemes for multimedia streaming. He was awarded the B.Eng. and M.Sc. degrees in software engineering from the Computer Science Department, "Politehnica" University of Timisoara, Romania in 1996 and 1997 respectively. Dr. Muntean's research interests include QoS and performance-related issues of adaptive solutions for multimedia delivery. He is a Student Member of the IEEE.

Philip Perry (M'92) is a Senior Research Fellow in the Performance Engineering Laboratory, with responsibilities in both the Department of Computer Science at University College Dublin, Dublin, Ireland and the School of Electronic Engineering at Dublin City University, Dublin, Ireland. He obtained his Ph.D. in microwave engineering from the Department of Electronics and Electrical Engineering at University College Dublin, Dublin, Ireland in 1998. He studied for his Master's degree at the University of Bradford, Yorkshire, England (1989), while his primary degree is from the University of Strathclyde, Glasgow, Scotland. His current research interests are focused on the applications and enabling technologies for mobile systems.

Liam Murphy obtained a B.E. in electrical engineering from University College Dublin in 1985, and Master's and Ph.D. degrees in electrical engineering and computer sciences from the University of California, Berkeley, in 1988 and 1992 respectively. He is currently a Senior Lecturer in Computer Science at University College Dublin, where he is the Director of UCD's Performance Engineering Laboratory. His current research interests include performance issues in multimedia transmission, Voice over IP, and component oriented software systems. Dr. Murphy is a Member of the IEEE.