A Simulative Study of Analysis-By-Synthesis Perceptual Video Classification and Transmission over DiffServ IP Networks

F. D'Agostino, E. Masala, L. Farinetti, J. C. De Martin*
Dipartimento di Automatica e Informatica / *IEIIT/CNR
Politecnico di Torino
C.so Duca degli Abruzzi, 24 — I-10129 Torino, Italy
Email: [dagostino | masala | farinetti | demartin]@polito.it

***Abstract*— This paper presents the results of transmission of video data on 2-class DiffServ IP networks using perceptual packet classification and slicing. An analysis-by-synthesis technique to identify perceptually important video regions, to create optimal video slices and to assign the resulting packets to the appropriate DiffServ classes is described. The proposed technique was implemented using the ISO/IEC MPEG-2 video coding standard. Several transmission scenarios, including homogeneous video traffic and interfering FTP traffic, were simulated using Network Simulator (NS). The proposed perception-based video transmission approach outperformed classical data partitioning in all tested network conditions, while delivering greater flexibility both in terms of network usage and potential to match time-varying channels. Substantially higher PSNR values than the regular best-effort case were also obtained assigning to the high-QoS class as little as 10% of the traffic. Demo sequences are available at http://multimedia.polito.it/icc2003.**

I. Introduction

Video streams are becoming an important portion of the overall traffic transmitted over IP networks. Real-time traffic, however, must deal with the limitations of IP networks, above all the lack of quality-of-service (QoS) guarantees [1]. Several approaches have been proposed to enhance the perceptual quality achieved by real-time multimedia transmissions over IP networks. Some techniques aim to optimize the encoding/packetization process to improve error resilience. Others suggest to modify the architecture of IP networks to provide some form of QoS guarantees [2], as in the case of the Differentiated Service (DiffServ) architecture [3].

According to the DiffServ Model, packets are classified setting an appropriate field of the packet IP header. Network routers then apply different packet forwarding behaviors depending on the value of such field. The two-bit network architecture [4] is one of the simplest DiffServ scenarios. It defines three classes of QoS, i.e., a high-cost *premium* service with no losses and low delay, an *assured* service with very low drop probability and a regular *best-effort* service that provides the behavior of the current Internet.

One of the simplest approaches to assign multimedia packets to the various DiffServ classes is to assign the entire flow of packets to a single DiffServ class. High–QoS bandwidth, however, is a *limited* as well as an *expensive* resource that needs to be employed efficiently. A more advanced prioritization approach could be, in the case of motion-compensated video, to classify packets depending on the frame type (I, P or B) they belong to. If data are coded using a scalable encoder, the resulting layers could be mapped to the DiffServ classes according to their importance.

Given the highly *non-uniform perceptual importance* of speech, audio, image and video data, however, a more efficient approach is possible, one that takes into account the perceptual importance of each single packet to be transmitted. Frameworks factoring the perceptual importance of individual multimedia packets were proposed in [5] and [6]. An *analysis-by-synthesis approach* to packet classification for multimedia signals has been proposed [7] [8]. The technique consists of replicating the decoder behavior in case of packet losses, and to assign to high–QoS classes those packets that, if lost, would generate high distortion levels at the decoder.

In this paper, the focus is on video transmission. The ISO/IEC MPEG-2 encoder [9] was modified in order to implement an analysis-by-synthesis slicing and classification technique that exploits the assured and the best-effort services defined in [4]. Performance results obtained by simulating video transmission over a DiffServ network for different traffic scenarios are presented. The proposed perception-based video transmission approach outperformed classical data partitioning in all tested network conditions, while delivering greater flexibility in terms of both network usage and potential to match time-varying channels. Substantially higher PSNR values than the regular best-effort case were also obtained assigning to high-QoS classes as little as 10% of the video traffic.

The paper is organized as follows. In Section II the DiffServ network architecture is briefly reviewed. In Section III the analysis-by-synthesis approach to packet classification is presented for the specific case of MPEG-2 video coding. Section IV describes the simulation setup and Section V presents the performance results of the proposed technique. Finally, conclusions are presented in Section VI.

0-7803-7802-4/03/$17.00 (C) 2003 IEEE

II. DiffServ IP Networks

A DiffServ network is an IP network in which packets receive a different forwarding behavior on nodes along their path depending on the service class they belong to. This architecture [3] defines different classes, each one suitable for specific purposes. One of the simplest cases uses three QoS levels[4]: a *premium* service, an *assured* service and a *best-effort* service. The first class is meant to transmit delay- and losses-sensitive traffic, such as interactive speech or video; the second class is characterized by a low drop probability, while the third is for less demanding types of traffic. One field of the IP packet header is designated to contain the identifier of the class the packet belongs to. IETF documents [10] specify it for both Version 4 and Version 6 of the IP protocol.

The scheduling algorithms used to manage the DiffServ router queues heavily influence the characteristics of the different QoS classes, because packets are placed by routers in different queues depending on the class they belong to. Implementations of the assured forwarding must attempt to minimize the long-term congestion within each class, while allowing short-term congestion resulting from bursts [11]. This requires an active queue management algorithm, e.g., RED [12]. When packets have to be dropped the DiffServ router selects more frequently the packets belonging to lower QoS classes. This behavior does not affect the order in which packets are forwarded, a fact that is significant for many applications.

III. Analysis-by-Synthesis packet classification

Highly efficient use of DiffServ networks resources in the case of multimedia transmission can be obtained if packet classification is performed taking into account the perceptual importance of the data to be delivered. The perceptual importance of a multimedia data block can be defined in terms of *the distortion that would be introduced by its loss* [7]. An analysis of the distortion introduced in case of loss can be carried out simulating the decoding process, including the concealment technique used to replace the missing data at the decoder. In the case of video, the reconstructed video segment is compared to the original material; the resulting distortion value can be used as an indication of its perceptual importance. Figure 1 shows the block diagram of the described analysis-by-synthesis approach.

The complexity and delay of the analysis-by-synthesis classification technique strongly depends on the frame types the sequence is composed of. If only I-type frames are present the analysis-by-synthesis technique can be greatly simplified. Due to the absence of temporal error propagation, the distortion caused by the loss of an elementary video element such as a macroblock can be easily computed. Assuming that the decoder applies a temporal concealment technique replacing the missing macroblocks with the macroblocks in the same position in the previous frame, the potential distortion associated to a macroblock is the Mean Squared Error (MSE) value between that macroblock and the macroblock that would be used to conceal its loss. If the sequence contains also P- and

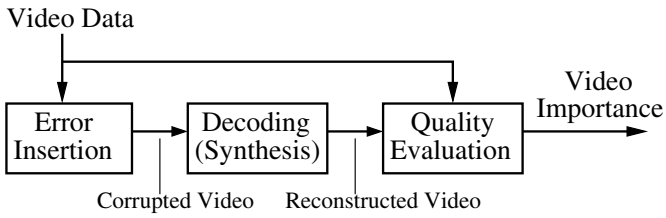

Fig. 1. Block diagram of the analysis-by-synthesis perceptual classification approach.

B-type frames, the algorithm becomes more complex because temporal propagation of the distortion must be taken into account. The delay, too, increases, because a certain number of subsequent frames must be considered, at least until the intra refresh procedure has significantly reduced error propagation. In both cases the error concealment technique used by the decoder has to be known to ensure proper simulation of the decoding process.

We chose to implement adaptive analysis-by-synthesis packet classification algorithm for the MPEG-2 video format. The MPEG-2 standard [13] defines many different basic syntax elements. The basic element for the description of texture and motion information of each picture is the *macroblock* (MB), which refers to an area of 16×16 pixels. An arbitrary number of consecutive macroblocks belonging to the same row is called a *slice*; this is the smallest unit which can be independently decoded. A packet should contain an integer number of slices since, in case of packet losses, a partial slice is useless. To minimize the effect of packet losses, only one slice per packet should be sent. On the other hand, transmitting several slices per packet would reduce overhead.

The perception-based video classification algorithm operates at macroblock level, computing a distortion measure for each of them using the analysis-by-synthesis approach, simulating the concealment in case of loss. Macroblocks belonging to the same frame are then analyzed to find a classification which minimizes total picture distortion in case of loss of one or more macroblocks given a certain constraint on the maximum rate. This can be accomplished with an initially empty set of macroblocks to be assigned to an high-QoS class, then incrementally adding to this set the macroblocks of the frame in decreasing order of distortion, until reaching the maximum rate constraint. The remaining macroblocks are assigned to the best-effort class. Then adjacent macroblocks belonging to the same class are grouped into single slices to minimize slice header overhead. As final step, slices of the same class are encapsulated into packets that are assigned to the corresponding service class. A more detailed description of the algorithm can be found in [8].

Figure 2 shows the result of the algorithm applied to a frame belonging to the video sequence known as *News*. The picture is obtained subtracting the previous frame from the current one. Darker levels of grey correspond to greater pixel-by-pixel differences. The slices enclosed by black outlines have been assigned by the classification algorithm to a high

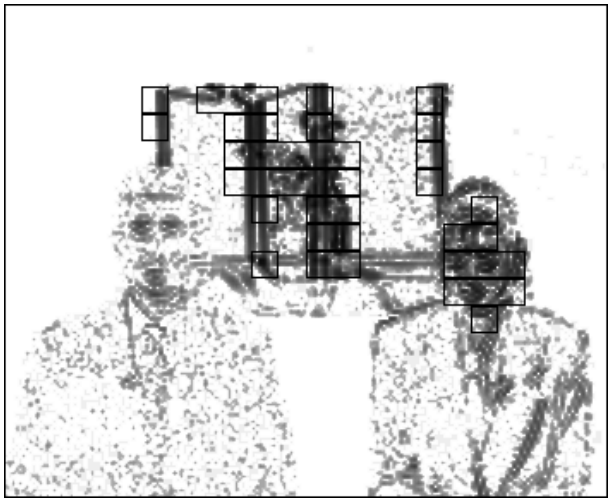

Fig. 2. Adaptive slicing of a *News* frame (grey levels emphasized to obtain a more distinct figure.)

QoS class. They correspond to dark regions of the frame, i.e., regions with strong differences with respect to the previous frame. If one of such slices were lost, it would cause high distortion at the decoder, because the concealment technique, which replaces missing macroblocks with the macroblocks in the same position in the previous frame, would generate poor estimates of the missing data.

As shown in Figure 2, the proposed classification technique can also be seen as a way to identify Regions Of Interest (ROI) inside the frame, e.g., the moving face of the speaker (right) and the video playing on the background (center). Stronger protection of the ROI's leads to overall better perceptual quality. Classic data partitioning, instead, uniformly protects all video regions, leading to a less efficient use of high-QoS bandwidth.

Subdividing the packet flow generated by a multimedia application into different DiffServ classes could lead to delays and reordering problems. A simple, widely-used solution is to introduce a playout buffer at the receiver; the decoder starts to play data a certain amount of time after the first packet has been received. We assume to use the IP/UDP/RTP [14] protocol stack; the packet reordering issue is easily solved by means of the sequence number present in the RTP header. Using this number, the decoder can place the packets in the playout buffer according to their original order. Besides, the playout buffer compensates, at least partially, the delay jitter affecting the packet arrival times. If packets suffer a high delay, they will be considered lost by the decoder. However, this is unlikely to happen for packets belonging to high QoS classes if the generated traffic respect the Service Level Agreement with the network provider.

IV. Simulation Setup

NS-2 package [15] was used to simulate a DiffServ network for different traffic conditions. The DiffServ routers employed the RED In Out (RIO) queue management algorithm [16].

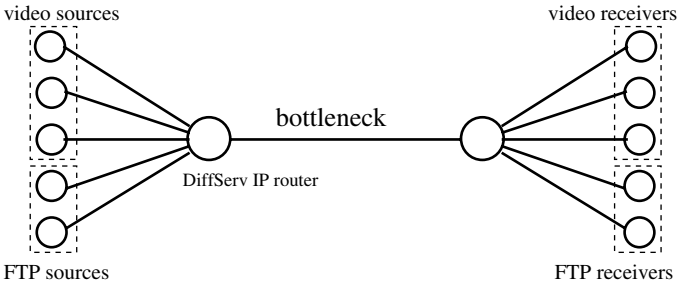

Fig. 3. Network topology.

The topology is shown in Figure 3. Several sources transmit video packets to their corresponding receivers. All flows must pass through the same bottleneck. In case of congestion, the DiffServ IP router drops best-effort packets to preserve the guarantees of the higher-QoS classes. The other links are oversized not to impact on the results.

We adopted the *Amélie* film trailer as test sequence. The sequence is characterized by several scene changes and, in the last section, high motion. The sequence was cyclically repeated to reach a length (296.4 s) adequate to obtain statistically valid results. Sequence size is 480×360; it was coded at 3.5 Mbit/s using I-frames only. Although the resulting compression ratio is generally lower than the one that can be obtained using also P- and B-type frames, I-frames-only compression has lower delay, lower complexity and high quality, and as such is used in a large number of applications.

We simulated two different traffic scenarios:

1) Homogeneous traffic: eight video over IP (ViP) sources at 3.5 Mb/s; bottleneck bandwidth is 24 Mb/s.
2) Interfering FTP traffic: eight ViP sources at 3.5 Mb/s, and two FTP sources; bottleneck bandwidth is 24 Mbit/s.

The bottleneck bandwidth was chosen to cause about 15% of packet losses. Each source begins to transmit at a time randomly selected in an interval of 60 seconds to avoid potential synchronization problems. Packets are transmitted equally spaced in time for a frame rate of 25 frame/s. The end-to-end propagation delay is 5 ms. The size of the playout buffer is 20 ms, enough to attenuate most of the effects of delay jitter.

Encoding was carried out with the proposed analysis-by-synthesis packet classification strategy. We modified the ISO reference decoder [13] to implement a simple temporal concealment technique, i.e., lost areas are replaced with the corresponding ones in the previous frame, assumed to have been correctly received. When a packet containing an MPEG picture header is lost, all slices belonging to the same frame are considered lost.

V. Simulation Results

PSNR with respect to the original uncompressed video sequence was used as distortion measure; it was computed as the average of each frame's PSNR. For each traffic scenario, we computed the PSNR values corresponding to three different video transmission techniques. The first one simulates the

TABLE I

PSNR VALUES FOR SEVERAL CLASSIFICATION TECHNIQUES; PACKET LOSS RATE OF ABOUT 15%; VIP SOURCE #1.

Classification Technique	Share of High-QoS Traffic	Homogeneous Traffic Case	Interfering FTP Traffic Case
None (100% best–effort)	0%	35.09 dB	34.62 dB
Analysis-by-Synthesis	10%	37.77 dB	37.47 dB
Analysis-by-Synthesis	20%	38.57 dB	38.32 dB
Data Partitioning	20%	37.20 dB	37.54 dB

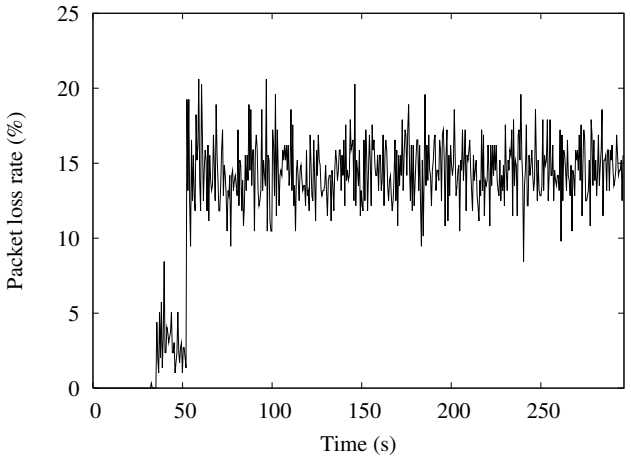

Fig. 4. Short-term packet loss rate as a function of time; ViP source #1; analysis-by-synthesis packet classification, 20% assured service; concurrent traffic: 7 other ViP sources.

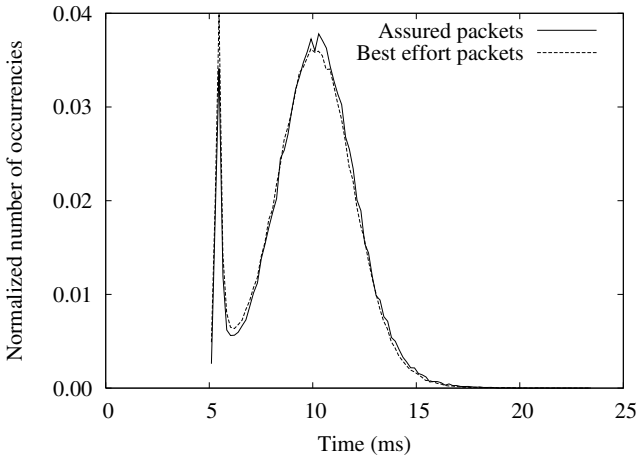

Fig. 5. Delay distribution for the two DiffServ classes; ViP source #1; analysis-by-synthesis packet classification, 20% assured service; concurrent traffic: 7 other ViP sources.

current Internet network, i.e., all traffic is sent as best-effort; video coding has been carried out by the standard ISO/IEC MPEG-2 encoder [9]. The second approach is the proposed technique that assigns the video packets to the available classes depending on their perceptual importance. For the second approach results were obtained both for 10% and 20% of assured traffic. The third one is the data partitioning technique. The DCT coefficients of each macroblock were split in two vectors, the first one always containing a fixed number of coefficients. The first vector, together with headers and motion vectors, is put in the base layer and sent as assured traffic; the remaining coefficients, which collectively can be viewed as a kind of enhancement layer, are sent as best-effort traffic. All simulations lasted enough to allow each ViP source to transmit the whole sequence at least one time.

A. Homogeneous traffic

The third column of Table I shows the PSNR values of the three approaches —best-effort, analysis-by-synthesis perceptual packet marking and data partitioning— when eight video sources are sending data to eight receivers. No other traffic is present in the network. The difference between the perceptual-based and the best-effort is clear, showing the effectiveness of the algorithm even when only 10% of traffic is sent with high-QoS. The proposed approach also outperforms data partitioning by about 1.3 dB when sending the same amount of assured traffic. Besides, analysis-by-synthesis classification assigning 10% of the traffic as high-QoS slightly outperforms data partitioning classification using twice as much high-QoS bandwidth.

Figure 4 shows the short-term packet loss rate seen by the receiver as a function of time. The initial period is loss free because the active sources do not exceed the bottleneck capacity. The loss probability increases when other ViP sources start to transmit. An average probability of about 15% is reached when all the sources are active. The packet loss value includes the share of packets discarded at the receiver because of excessive delay.

Figure 5 shows the delay distribution of the packets belonging to the assured-forwarding and best-effort classes, respectively. Values are normalized with respect to the total number of packets in each class. The two distributions are very similar: the assured class, in fact, has no precedence over the best-effort class, only a lower drop probability. The peak around 5 ms, equal to the propagation delay, is due to the first 50 seconds of the simulation, when overall traffic does not exceed the bottleneck capacity.

B. Interfering FTP traffic

The rightmost column of Table I shows the PSNR values of the scenario when eight video sources and two FTP sources are sending traffic on the same link. Analysis-by-synthesis packet classification outperforms best-effort transmission in this case too. The situation is similar to the preceding case

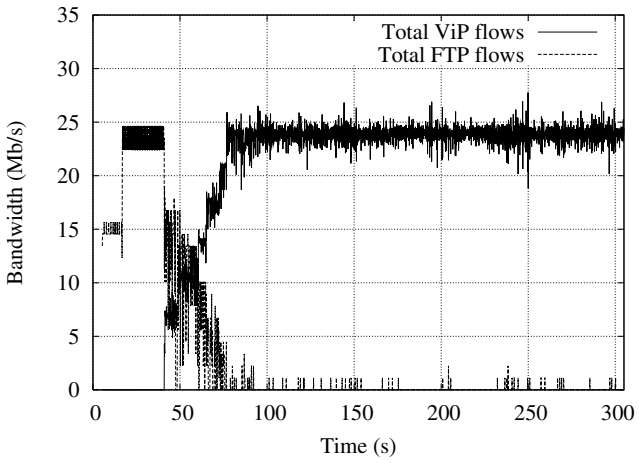

Fig. 6. Aggregate throughput as a function of time; analysis-by-synthesis packet classification, 20% assured share; concurrent traffic: 2 greedy FTP sources.

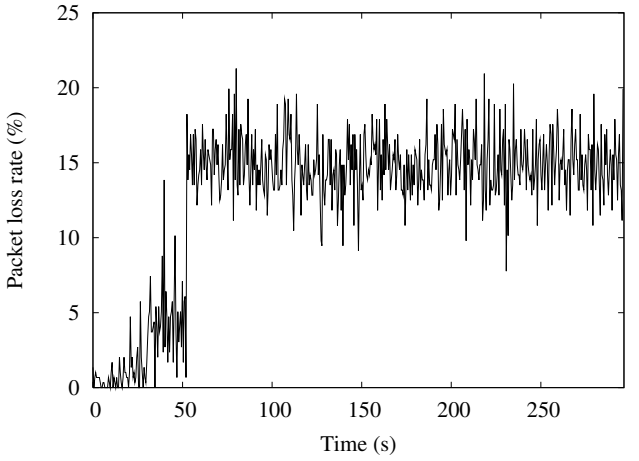

Fig. 7. Short-term packet loss rate as a function of time; ViP source #1; analysis-by-synthesis packet classification, 20% assured share; concurrent traffic: 7 other ViP sources, 2 FTP sources.

because the FTP flows, due to the adaptive behavior of the TCP protocol, nearly stop to transmit. In fact, TCP reacts to the high packet losses reducing its transmission window to the minimum, thus the FTP sources nearly turn off. Figure 6 shows throughput as a function of time. The number of the active ViP sources gradually increases until they saturate the bottleneck bandwidth, and at the same time the FTP reduces it throughput until nearly zero.

Figure 7 shows the short-term packet loss rate. The plot is similar to the homogeneous traffic case. It differs in the first 50 seconds where the packet loss probability is higher than the previous case because the ViP traffic has to compete with the TCP traffic present in the bottleneck.

All examined scenarios show that the proposed algorithm outperforms the best-effort case by 2.7 to 3.7 dB. With respect to the regular best-effort case, even an assured share of just 10% delivers a significant PSNR gain. Sending 20% of the traffic as assured delivers an additional perceptual gain of about 0.8 dB. Comparisons with classic data partitioning show that equivalent PSNR performances can be achieved transmitting only half of the traffic with high QoS.

VI. Conclusions

We presented simulation results of the analysis-by-synthesis approach to classification of video signals for transmission over DiffServ IP networks. Data packets are assigned either to the assured or to the best-effort service depending on their individual perceptual importance. A technique for adaptive classification of MPEG-2 video sequences was tested using objective quality measures. Simulation of DiffServ transmission scenarios in various traffic conditions showed that the analysis-by-synthesis packet classification greatly outperforms the best-effort case even when only a small amount (10%) of the traffic is transmitted with high QoS. Moreover, analysis-by-synthesis classification marking 10% of the traffic as high-QoS service outperforms data partitioning classification using twice as much high-QoS bandwidth.

References

[1] D. Wu, Y. Hou, and Y.-Q. Zhang, "Transporting Real-Time Video over the Internet: Challenges and Approaches," *Proceedings of the IEEE*, vol. 88, no. 12, pp. 1855–1875, December 2000.

[2] R. Hunt, "A Review of Quality of Service Mechanisms in IP-based networks – Integrated and Differentiated Services, Multi-layer Switching, MPLS and Traffic Engineering," *Computer Communications (Elsevier)*, vol. 25, no. 1, pp. 100–108, January 2002.

[3] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services," *RFC 2475*, December 1998.

[4] K. Nichols, V. Jacobson, and L. Zhang, "A Two-bit Differentiated Services Architecture for the Internet," *RFC 2638*, July 1999.

[5] J. K. J. Shin and C. Kuo, "Quality-of-Service Mapping Mechanism for Packet Video in Differentiated Services Network," *IEEE Transactions on Multimedia*, vol. 3, no. 2, pp. 219–231, June 2001.

[6] P. Chou and Z. Miao, "Rate-Distortion Optimized Streaming of Packetized Media," *submitted to IEEE Transactions on Multimedia*, February 2001.

[7] J. D. Martin, "Source-Driven Packet Marking For Speech Transmission Over Differentiated-Services Networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, Salt Lake City, Utah, May 2001, pp. 753–756.

[8] E. Masala, D. Quaglia, J.C. De Martin, "Adaptive Picture Slicing for Distortion-Based Classification of Video Packets," in *Proc. IEEE Workshop on Multimedia Signal Processing*, Cannes, France, October 2001, pp. 111–116.

[9] S. Eckart and C. Fogg, "ISO/IEC MPEG-2 software video codec," *Proc. SPIE*, vol. 2419, pp. 100–118, Apr. 1995.

[10] K. N. et al., "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers," *RFC 2474*, December 1998.

[11] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, "Assured Forwarding PHB Group," *RFC 2597*, June 1999.

[12] S. Floyd and V. Jacobson, "Random Early Detection Gateways for Congestion Avoidance," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 397–413, August 1993.

[13] ISO/IEC, "MPEG-2 generic coding of moving pictures and associated audio information," *ISO/IEC 13818*, 1996.

[14] R. F. H. Schulzrinne, S. Casner and V. Jacobson, "RTP: A transport protocol for real-time applications," *RFC 1889*, January 1996.

[15] UCB/LBNL/VINT, "Network Simulator – ns – version 2.1b8a," *URL: http://www.isi.edu/nsnam/ns*, 2001.

[16] D. Clark and W. Fang, "Explicit Allocation of Best Effort Packet Delivery Service," *Internet draft*, September 1997.