Published in the Proceedings of the SPIE International Symposium
on Voice, Video and Data Communications.
Boston, Mass, September 1999

# Perceived quality and bandwidth characterization of layered MPEG-2 video encoding*

Jun-ichi Kimura†, Fouad A. Tobagi, Jose-Miguel Pulido♮, Peder J. Emstad‡

Dept. of Electrical Engineering, Stanford University, Stanford, CA 94305

## ABSTRACT

The current Internet is not well suited for the transmission of high quality video such as MPEG-2 because of severe quality degradation during network congestion episodes. One possible solution is the combination of layered video coding with the Differentiated Services (DiffServ) architecture; different video layers are mapped into different priority levels, and packets with different priorities receive a different dropping treatment in the network. It is expected that with layering and priority dropping, graceful degradation of video quality will be experienced during congestion episodes.

We consider various layering mechanisms defined in the MPEG-2 standards; namely, temporal scalability, data partitioning (DP) and Signal to Noise Ratio (SNR) scalability. The main issue in this paper is how layers should be created to maximize perceived video quality over a given range of network conditions. Key to our study is the use of real life video sequences and a video quality measure consisting of a perceptual distortion metric based on the Human Visual System (HVS). Our results show that video quality is sensitive to how layering is accomplished, and that there is an optimum layering that maximizes the quality for a given network condition. Our results also show that layering can achieve higher network loading for a given minimum quality target than non-layered video, and can achieve graceful degradation over a wider range of network conditions. We have also seen that the wider the range of network conditions is, the higher is the number of layers required in order to remain at the highest possible quality level for each network condition. In particular, we demonstrate how three or four layers achieve better results than two layers; however, additional layers beyond four provide marginal improvement. Therefore, from a practical point of view, three or four layers are sufficient to attain most of the benefits of layering. We compare the various scalability techniques in terms of complexity and video quality. Temporal scalability, which restricts the layering to be done at frame boundaries, is the simplest to implement and introduces no overhead, but performs poorly compared to data partitioning, which allows the grouping of coefficients into layers independent of the frames they belong to. This shows that, contrary to customary belief, dropping data in B frames prior to dropping data in P or I frames is a poor layering technique. DP is much simpler to implement and introduces significantly lower overhead than SNR scalability. However, SNR scalability provides higher quality than DP when network conditions are particularly poor.

**Keywords:** Video transmission, MPEG-2 video, layered video coding, video scalability, temporal scalability, SNR scalability, data partitioning, Differentiated Services networks.

## 1 INTRODUCTION

The best effort service provided by the current Internet is not well suited for the transmission of high quality video such as MPEG-2. The burstiness of data traffic combined with limited network resources produce congestion episodes during which packets are dropped leading to severe quality degradation. It has been observed in [1], that packet loss incurred in the transmission of MPEG video over the public Internet varies over a wide range, with the peak rate well above the average. It has also been observed that a mere 3% packet loss affects 30% of the frames (due to dependencies in the encoded video bit stream). To limit this degradation, the idea of dynamically modifying encoding parameters at the source based on feedback from the network has been considered. However, adaptation at the source cannot overcome the large variation in packet loss incurred when video and data traffic are mixed, as indeed emphasized in [2]. Video and data traffic must be separated into different classes of service, and video traffic must be shielded from the burstiness seen in data traffic. Even when video

* This work has been supported in part by 3Com Corporation.

† Visiting scholar at the Stanford Center for Telecommunications from Hitachi, Ltd., Central Research Laboratory.

♮ Supported by a fellowship for graduate studies from "la Caixa" Foundation, Spain.

‡ Visiting professor in the Department of Electrical Engineering, from the Faculty of Electrical Engineering and Telecommunications, Norwegian University of Science and Technology,

traffic is separated from data traffic, packet loss may still occur caused by the variability in the video traffic itself. As explicitly shown below in this paper, even when packet loss rate is small the dependencies in the encoded bit stream lead to significant quality degradation. One possible solution to this problem is to prioritize video data according to its relative importance and apply priority dropping within the network during congestion.

Priority dropping in the Internet will be made possible by means of the Differentiated Services architecture (DiffServ) currently being developed by the IETF [4]. In this architecture, packets are marked by the source, and network nodes treat them differently according to their marking. The Assured Forwarding (AF) type of service [5] defines four classes with three levels of drop precedence in each to provide priority dropping.

Prioritization of MPEG-2 video data may be achieved by means of any one of the four scalable coding techniques specified in the MPEG standards [6, 7]; namely: temporal scalability, data partitioning (DP), SNR scalability, and spatial scalability. However, the MPEG standards do not specify how layers should be created, as the focus in these standards is on the behavior of the decoder. With temporal scalability it is customary to separate I and P frames from B frames, and to drop B frames before I or P frames; and it is shown in [9, 10] that this indeed achieves better results than if no distinction among the different frame types were made. Aravind et al. [8] considered spatial scalability, DP and SNR scalability with the goal of comparing them in terms of performance and ease of implementation in a given network environment. More specifically, they considered the scenario consisting of the transmission of two-layered video streams over ATM networks. The base layer is transmitted over a virtual circuit of constant bit rate and is created by adapting the layering parameters so as to match the constant bit rate of the circuit. The enhancement layer is transmitted over a virtual circuit with variable bit rate to which random cell loss is applied. They show that DP is the easiest to implement but provides the lowest performance among the three; spatial scalability provides the best performance but has the highest implementation complexity, and SNR scalability is in between in terms of both implementation complexity and performance.

The goal of this paper is to determine how video streams should be divided into layers and how to prioritize these layers so as to maximize the perceived quality over a given range of network conditions. Our approach consists of examining all possible layering structures, assessing for each layering structure the contribution to quality and bandwidth of each of the layers, and then identifying the layering structure that provides the best quality-bandwidth tradeoff for a given range of available network bandwidth. Key to our study is the use of real video sequences, and a quality measure consisting of a perceptual distortion metric (PDM) based on the human visual system (HVS), designed specifically for MPEG-2 [14, 15]. The latter is a crucial aspect of our study as it allows us to accurately assess the user perceived quality in our experiments.

The paper is organized as follows. Section 2 is devoted to layering based on data partitioning (and temporal scalability as a special case). Section 3 is devoted to layering based on SNR scalability, and includes a comparison between DP and SNR layering. Section 4 uses a simplified DiffServ network model to illustrate the benefits of layering and priority dropping. Conclusions are presented in Section 5.

## 2 MPEG-2 VIDEO LAYERING BASED ON DATA PARTITIONING

### 2.1 Data Partitioning

Data partitioning is a technique that divides the encoded video bit stream into two or more layers simply by allocating the data in the bit stream (motion vector information and DCT coefficients) to the various layers. As specified in the MPEG standards, Priority Break Points (PBPs) are introduced in each slice header to specify the grouping of DCT coefficients into the layers. Thus in data partitioning, PBPs are the parameters used in the layer creation process. To form n layers, n-1 PBPs are selected. Only headers (namely, sequence, group of pictures, picture, and slice headers) are included in all layers; thus the overhead introduced in the creation of layers is relatively negligible. Note that data partitioning is performed at the system level rather than in the encoding process itself. Accordingly, it may easily be used with pre-encoded video. For a more detailed description of data partitioning, see [6].

We consider only a subset of the PBPs specified in the MPEG standards; they are selected so as to have minimal sets of DCT coefficients the loss of which will have a perceptible quality effect. Note that the sets are not equal in size due to the relative contribution of the DCT coefficients to quality; e.g., dropping the DC coefficients may be more noticeable than dropping the $16^{th}$ to the $64^{th}$ coefficients. The selected PBPs are shown in Table 1 and are renumbered for simplicity by drop codes (dp_codes), ranging from 0 to 7.

**Table 1 Definition of dp_code**

| dp_code | priority_breakpoint | definition |
|---|---|---|
| 0 | 127 | syntax elements up to 64th (run,level) DCT coefficient pairs are in the layer |
| 1 | 79 | syntax elements up to 16th (run,level) DCT coefficient pairs are in the layer |
| 2 | 75 | syntax elements up to 12nd (run,level) DCT coefficient pairs are in the layer |
| 3 | 71 | syntax elements up to 8th (run,level) DCT coefficient pairs are in the layer |
| 4 | 67 | syntax elements up to 4th (run,level) DCT coefficient pairs are in the layer |
| 5 | 64 | syntax elements up to DC coefficient and 1st (run,level) DCT coefficient pair |
| 6 | - (*) | syntax elements up to DC value are in the layer (*) |
| 7 | 1 | only headers are in the layer (whole picture is lost and previous field data is used if enhancement layer stream is not received) |

(*) This mode is prohibited in MPEG-2 standards because of risk of startcode emulation.

Note that dp_codes may be different for different types of frames. Thus for a two-layer data partitioning, we use a triplet of dp_codes corresponding to picture types (I, P and B) in that order. For example, the triplet "135" assigns dp_code 1 to I pictures, dp_code 3 to P pictures and dp_code 5 to B pictures. This means that all the DCT coefficients up to the 16th coefficient of macroblocks pertaining to I frames are included in the base layer. Similarly, up to the 8th coefficient of P frames and up to the 1st coefficient of B frames are included in the base layer, respectively. If a particular picture type (P and/or B) is not used in the encoding of the video, the corresponding dp_code is set to 0. For n-layer data partitioning, a vector of n-1 triplets is used.

### 2.2 Optimum layering in two-layer data partitioning

We consider here the case of two-layer data partitioning, and present the quality and bandwidth results corresponding to different dp_code triplets for various sequences. All sequences we consider are PAL sequences, with frame size 720 x 576 pixels. The group of pictures (GOP) structure is IBBPBBPBBPBBI… i. e., it comprises 12 frames between I frames (N=12), and 3 frames between anchor frames (M=3). The picture structure is frame structure. The sequence is encoded with an open-loop Variable Bit Rate (VBR) control scheme [11]; we use a fixed quantizer step size Q for I and P pictures and 1.4Q for B pictures. Motion estimation is performed with full search, +-16 pixel/frame (H), +-16 pixel/frame (V). The slice size is 45 macroblocks, which corresponds to an entire row of macroblocks in the picture. A corrupted frame is concealed as follows: it is reconstructed with higher layer data (e.g., base layer) if such data exists, otherwise the previous field is copied. Quality measurements are obtained by means of the perceptual distortion metric PDM and converted to the rating defined in ITU-R recommendation BT.500 [15]. ITU-500 rating ranges from 5 to 1, where 5 is excellent, 4 good, 3 fair, 2 poor and 1 bad.

In Figure 1(a) we present quality-bandwidth results for the “flower” sequence. Both quality and bandwidth values are averages calculated over the entire video sequence. Black circles correspond to non-layered streams encoded with various values of Q. White circles correspond to the base layer of the video encoded with Q=16 and layered according to different dp_codes, except for the rightmost white circle marked [base + enhancement]. Note that the latter is very close to the non-layered case with Q=16, confirming the fact that the overhead associated with data partitioning is negligible. The difference in bandwidth between the white circle marked [base + enhancement] and any other white circle represents the size of the corresponding enhancement. In Figure 1(b) we present similar results for two other sequences, namely barca and basket.

We say that a layering structure dominates another if it achieves the same or higher quality for the same or lower bandwidth. We identify in Figure 1(a) the dominant layering structures and find that they correspond to the following dp_code triplets: 003, 014, 005, 015, 016, 036, 136, 046, 146, and 156. These are also the dominant structures for all other sequences we studied, including those shown in Figure 2(b). From these results we deduce that there is an order of importance to quality among motion vector information and DCT coefficients in the various picture types which is independent of the video content. This order of importance is explicitly shown in Figure 3.

We find it worthwhile to point out that dropping entire B pictures (dp_code triplet 007) prior to dropping any information in P or I pictures leads to very bad quality, while dp_code triplet 036 gives a quality rating of 3.8 for the same resulting bandwidth. The reason is that the inclusion of motion vector information and the DC coefficients of B pictures in the base layer allows the decoder to reconstruct B pictures, contributing significantly to perceived quality.

Finally, we note that to achieve a given average bandwidth, a non-layered OL-VBR stream with the appropriate value of Q provides better quality than a two-layer stream with the appropriate dp_code triplet. As we can see from Figure 1, the degradation in quality is different for different sequences, with the degradation seen in Figure 1(b) with barca and basket being among the worst observed over all sequences studied.

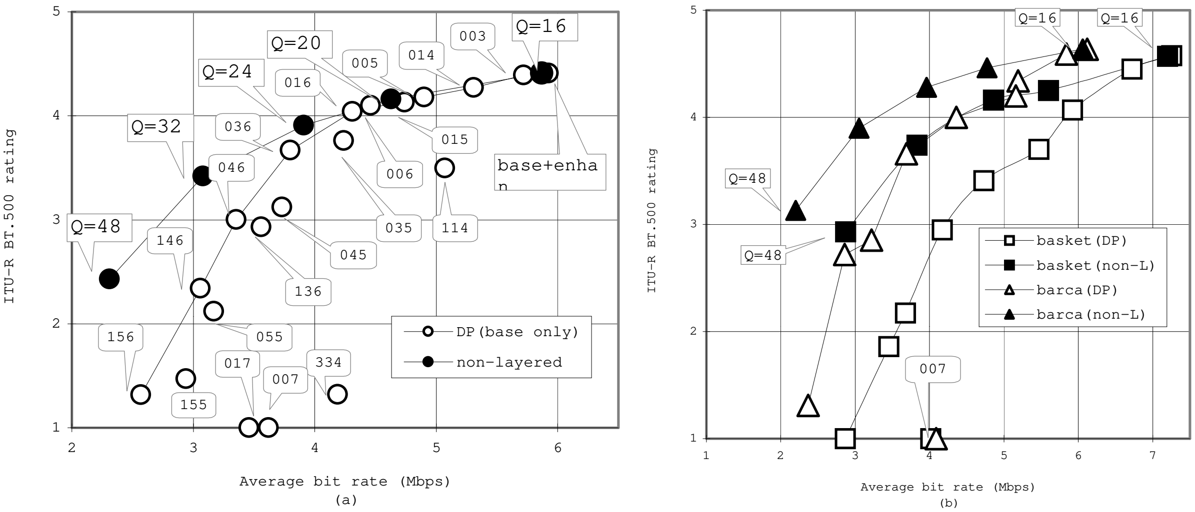


**Figure 1. Picture quality versus mean bit rate for 2-layer data partitioning and GOP structure M=3 N=12. (a) "flower" sequence, (b) "basket" and "barca" sequences**

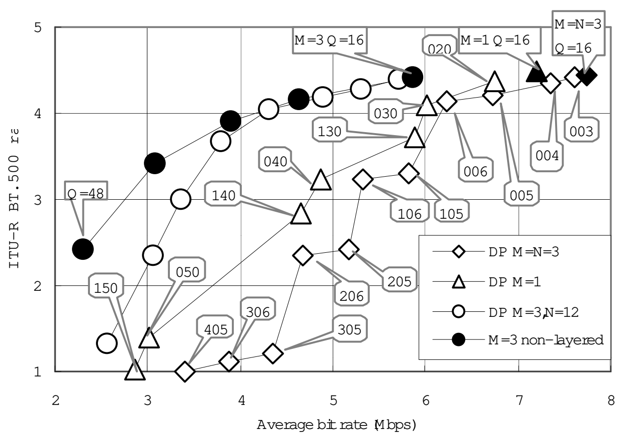


**Figure 2. Performance comparison at different GOP structure; 2-layer data partitioning, "flower" sequence.**

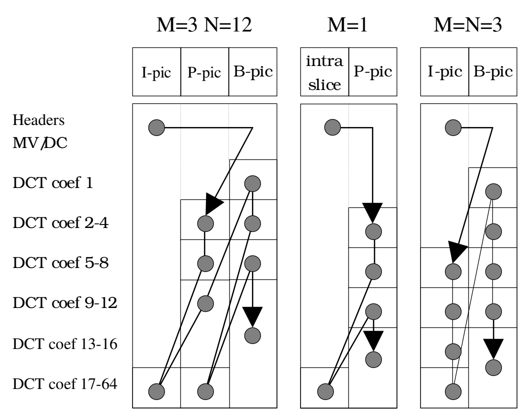


**Figure 3. Order of importance of bitstream data elements with data partitioning. (Line connected from the most important element to the least important one.)**

For completeness, we assess the effect that the GOP structure may have on our results. Figure 2 compares different GOP structures using data partitioning. For (M, N) = (1, inf.), (i.e., only P pictures,) rotational intracoding of groups of three consecutive slices from frame to frame is considered, with the dp_codes used for I pictures being applied to the intra slices. We observe in Figure 2 that (M, N) = (3, 12) provides the best performance even in the range of lower bandwidth. This result indicates that the GOP structure with the highest coding efficiency is also the one that provides the most graceful quality degradation. We also observe that the case of (M, N) = (3, 3), (i.e., a GOP of IBBIBBI...), provides the poorest results; we note that in this case, eliminating even the highest DCT coefficients in I pictures has a significant effect on quality degradation. Finally, we note that the case (M, N) = (1, inf.) provides a graceful degradation that is comparable to the case (M, N) = (3, 12), except for the fact that it has a lower coding efficiency. Hereafter, we limit ourselves to the GOP structure (M, N)=(3, 12).

### 2.3 Two-layer data partitioning with random packet loss

We examine in this section the effect of random loss in the enhancement layer on the perceived quality. For this purpose, we packetize the two-layer bit streams, and apply random packet loss to packets carrying data belonging to the enhancement layer. We consider that the streams are packetized according to the RTP/UDP transport protocol [12]. The maximum packet size is 1440 bytes. Packets are considered formed and transmitted when the packet buffer (of size 1440 bytes) for a layer is filled up, or when a new picture begins, whichever occurs first. (In this model, no instantaneous rate restrictions are assumed. If such restrictions were to be considered, then a packet buffer of size larger than a single packet would need to be used instead). We adopt the MPEG-2 specific RTP header format [13] to allow the decoder to synchronize to the next slice header even if the picture header information is lost due to packet loss.

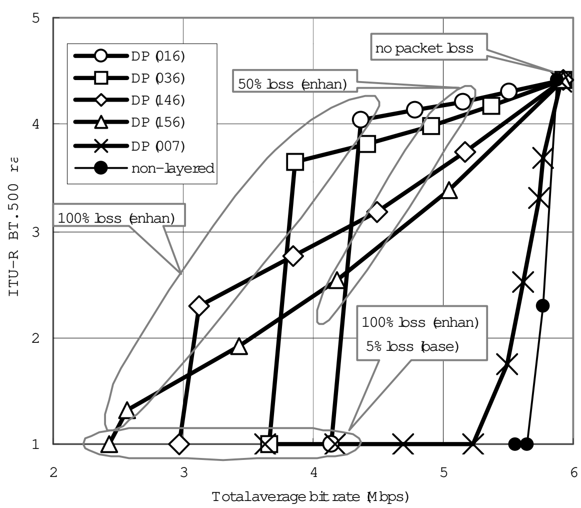


**Figure 4. Picture quality versus mean bit rate for 2-layer data partitioning with random packet loss in the enhancement layer; ("flower", M=3 N=12.)**

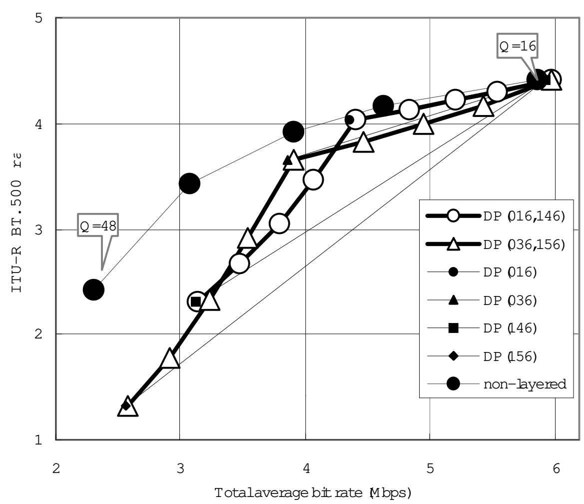


**Figure 5. Picture quality versus mean bit rate for 3-layer data partitioning with prioritized random packet loss in the lower layers ("flower", M=3 N=12.)**

The results are shown in Figure 4 for several dp_code triplets. For each dp_code triplet, we plot the perceived quality versus the resulting bandwidth for the situations where 25%, 50%, 75 % and 100% of the packets in the enhancement layer are lost. We also plot the perceived quality for the case where 5% of the base layer packets are lost in addition to all enhancement layer packets. It is clear that packet loss in the base layer results in severe quality degradation. For the case of dp_code triplet 016, a mere 5% loss in the base layer translates to a drop in quality from an ITU-500 rating of 4 (good) to an ITU-500 rating of 1 (bad). This clearly establishes the need to protect the base layer from loss in order to take advantage of layering and priority dropping. The quality degradation incurred with partial loss of packets in the enhancement layer varies almost linearly with the rate of packet loss. In contrast, we observe a very steep quality degradation in the non-layered stream, similar to that observed when loss occurs in the base layer with layered video streams; this is so precisely because both data that contribute significantly to quality and data with marginal contribution see the same dropping probability. Note also that the layering structure corresponding to dp_code triplet 007, (i.e., considering the entire B pictures to be in the enhancement layer as is done with temporal scalability), is only slightly better than the non-layered stream. This confirms the fact that some data in the B frames (such as motion vectors and DC coefficients) contribute significantly to quality and should not be assigned to the enhancement layer.

### 2.4 Multi-layer data partitioning

The dominant layering structures obtained with two-layer data partitioning represent an upper bound on the quality-bandwidth trade-off achievable with layering. From the results obtained so far however, it is clear that the choice of a single two-layer data partitioning structure does not by itself allow one to achieve the best quality for the entire range of network bandwidth. As seen in Figure 4, dp_code triplet 016 provides good quality (a rating of 4 and above) but over a somewhat limited bandwidth range (from 6 to 4.5 Mbps). dp_code triplet 146, on the other hand, covers a much wider range of bandwidth (from 6 to 3 Mbps), but provides a linearly degrading quality from a rating of 4.5 (very good) to a rating of 2 (poor). That is, at an average bandwidth of 4.5 Mbps, dp_code 146 achieves a level of quality equal to only 3 (fair). To be able to achieve the highest quality level possible over the entire range of bandwidth (from 3 to 6 Mbps), as determined by the dominant solutions, a layering structure with more than two layers is needed. This is illustrated in Figure 5 in which we plot the quality bandwidth tradeoff for the flower sequence with 3-layer data partitioning, and random packet loss prioritized over the three layers. (The three layers are referred to here as base layer, middle layer, and enhancement layer.) Two layering structures are shown in Figure 5, corresponding to dp_code triplets (016,146) and (036,156). The first triplet denotes the partition between the enhancement and the other two layers, and the second triplet denotes the partition between the base layer and the middle layer. The middle layer and base layer combined provide acceptable quality, but the base layer alone does not. We have chosen to create the middle layer from the original base layer in two-layer scalability so as to provide finer dropping granularity when loss goes beyond the enhancement layer. The middle layer will then act as a "shield" to protect important information carried by the base layer. This allowed us to prevent the significant quality degradation that occurred when 5% of the base layer is dropped, as shown in Figure 4. With random packet loss randomized over the three layers, the 3-layer structures achieve a piece-wise linear approximation of the upper envelope corresponding to the dominant solutions. The larger the number of layers, the closer is the approximation. The penalty from a layering overhead point of view with data partitioning is negligible and is not a factor. The only limitation we have on the number of layers comes from the limited drop code values that we chose to consider. From a practical point of view, however, we observe that any additional layer beyond 4 provides a rather limited incremental improvement. Furthermore, in real life scenarios the number of layers maybe limited by other factors, such as the complexity of implementing multiple-level priority dropping in network nodes.

# 3 MPEG-2 VIDEO LAYERING BASED ON SNR SCALABILITY

## 3.1 SNR scalability

Two-layer SNR scalability consists of encoding the video using a quantizer scale Qb in order to generate a base layer, and then encoding the difference between the original video and the base layer using a quantizer scale Qe to generate the enhancement layer. (Note that Qb is larger than Qe.) For more than two layers, the same process is applied recursively using additional quantizer scale parameters. Figure 6 shows a block diagram for a SNR scalable encoder as specified in the MPEG standards, which we refer to here as “SNR scalability”. In SNR scalability, the total bandwidth for the layered stream is higher than the bandwidth of the corresponding non-layered stream (obtained with the same quantizer step size as the enhancement layer) because the DCT coefficients are fragmented among layers, and information for every DCT coefficient is included in all the layers. The layering overhead for the MPEG standards “SNR scalability” is in the range of 5% to 20%.

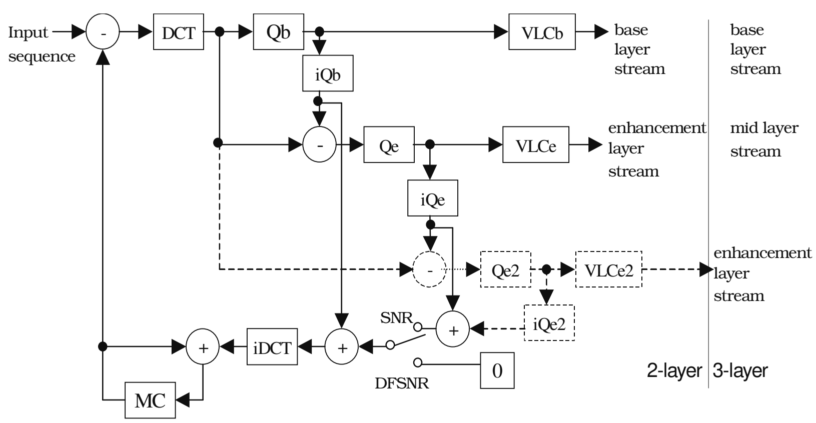


**Figure 6. Block diagram for SNR scalable encoder as in MPEG standards**

The SNR scalable encoder shown in Figure 6 is subject to drift due to the feedback from the enhancement layer to motion compensation. In [16], a drift-free SNR scalable encoder has been proposed. It uses prediction images that are reconstructed only from the base layer. Thus, even if a decoder loses part of the enhancement layer, errors are not accumulated. The diagram for this encoder is similar to the one shown in Figure 6 except for the feedback from the enhancement layer to motion compensation, which is omitted. Thus, this encoder is not compliant with the MPEG standards. We refer to it here as "DFSNR scalability". While DFSNR scalability is more resilient to loss in the enhancement layer than the standard MPEG SNR scalability, it requires more bits in the enhancement layer. A good example illustrating this fact is the encoding of a still sequence for which DFSNR encoders always transmit the difference between the (low quality) prediction images and input images, while the MPEG SNR scalability encoders transmit only skip information. Accordingly, the overhead for DFSNR is in the range of 30% to 70%.

In summary, we note that contrary to data partitioning which is accomplished at the system level, SNR scalability is an integral part of the encoding process; and contrary to data partitioning which introduces virtually no overhead, SNR scalability introduces a substantial overhead.

### 3.2 Two-layer SNR scalability

We consider here the case of two-layer SNR scalability. We show in Figure 7 the quality-bandwidth results for the “flower” sequence with GOP (M, N) = (3, 12). Figure 7(a) corresponds to MPEG SNR and Figure 7(b) corresponds to DFSNR. As in Figure 1 for data partitioning, black circles correspond to non-layered streams encoded with different values of Q. White circles correspond to layered video with Qe=16 and various values of Qb (20, 24, 32, and 48). [The convention used here to denote the quantizer scale parameters used in the creation of the layers is the pair (Qe, Qb)]. We see practically similar results as for the case of data partitioning shown in Figure 1. We note however the following differences. SNR scalability offers better quality in the low bandwidth range, indicating its higher resilience capability; and this is more so for DFSNR than for MPEG SNR scalability. SNR scalability requires higher bandwidth than data partitioning to achieve high quality levels, due to the overhead it introduces; and this is more so for DFSNR than for MPEG SNR scalability.

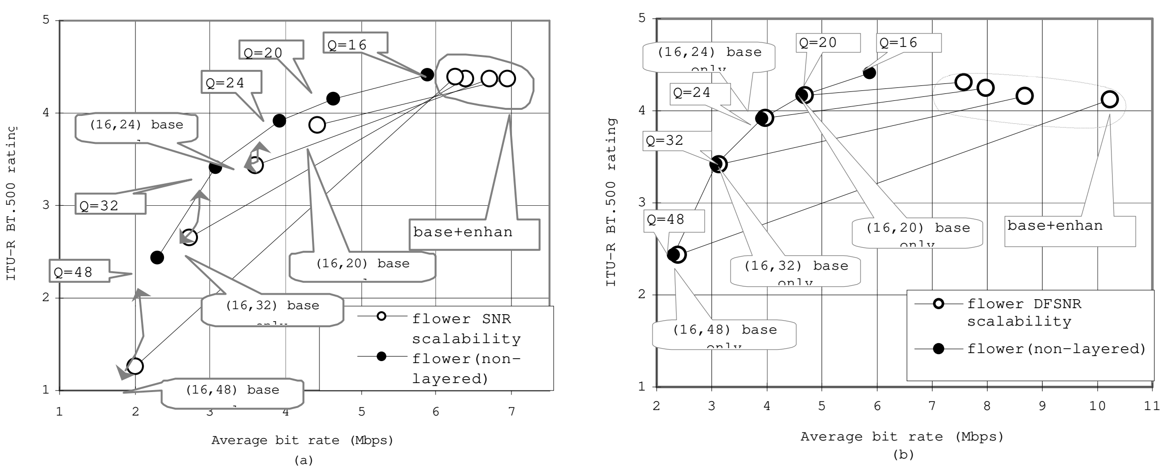


**Figure 7. Picture quality versus mean bit rate for 2-layer SNR scalability ("flower", M=3 N=12); (a) MPEG SNR scalability, (b) DFSNR scalability.**

We show in Figure 8 quality-bandwidth results with random packet loss inflicted on the enhancement layer only. Figure 8(a) corresponds to MPEG SNR and Figure 8(b) corresponds to DFSNR. We observe exactly the same behavior as with data partitioning, exhibiting a quality that degrades linearly with the rate of packet loss, and that drops to very poor rating as soon as data loss is incurred by the base layer. Furthermore, similar differences are seen between Figure 8 and Figure 4 as those seen between Figure 7 and Figure 1, due to the same reasons given above pertaining to overhead and loss resiliency.

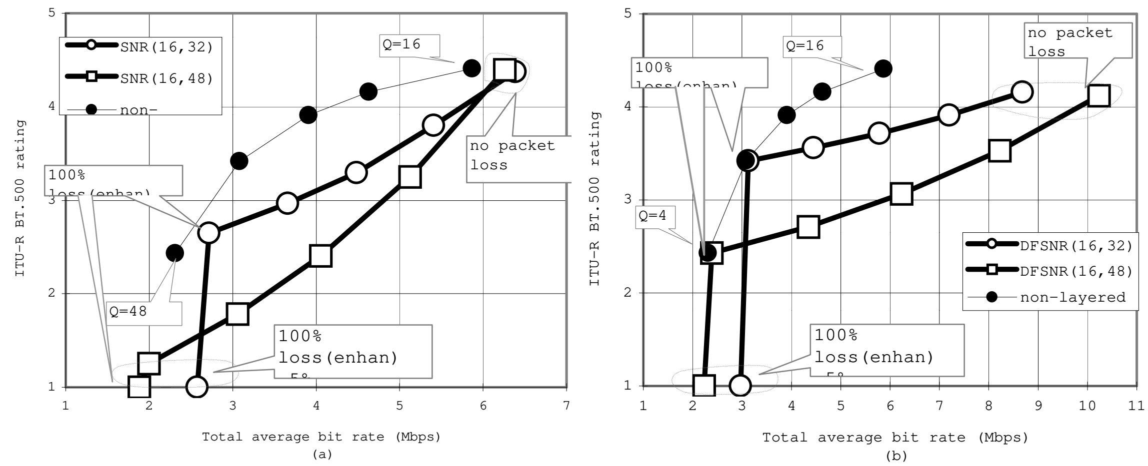


**Figure 8. Picture quality versus mean bit rate for two-layer SNR scalability with random packet loss in the enhancement layer; ("flower", M=3, N=12) (a) MPEG SNR scalability, (b) DFSNR scalability**

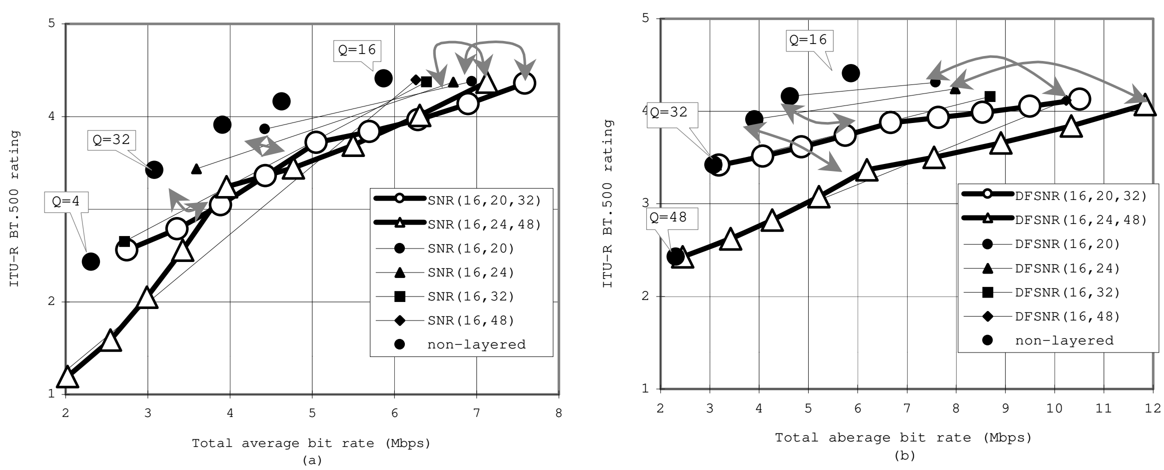


**Figure 9. Picture quality versus mean bit rate for three-layer SNR scalability with prioritized random packet loss in the lower layers ("flower", M=3 N=12) (a) SNR scalability, (b) DFSNR scalability**

## 3.3 Three-layer SNR scalability

In three-layer SNR scalability, the layering parameters are denoted by a triplet of quantizer scale values (Qe, Qm, Qb) corresponding to the enhancement layer, the middle layer and the base layer respectively. In Figure 9, we show quality-bandwidth tradeoff for three-layer SNR scalability and random packet loss prioritized over the three layers. Figure 9(a) is for MPEG SNR scalability, and Figure 9(b) is for DFSNR. We note that, contrary to data partitioning, increasing the number of layers from 2 to 3 not only did not offer the improvement over the 2-layer case (selected so as to cover the same intended range of bandwidth), but also degraded the picture quality. This is mainly due to the additional overhead in SNR introduced with the additional layer. This symptom is particularly noticeable in the case of (16,20,32) because of the rather fine quantization used in the middle layer which introduces relatively larger overhead. In DFSNR scalability, the increase in overhead is even greater than for MPEG SNR scalability (up to 80%). Clearly, in both cases, we cannot expect any advantage from more than three layers. In summary, three-layer data partitioning provides the best quality-bandwidth tradeoff of all cases studied, except in the low bandwidth range where DFSNR is superior due to its high error resiliency.

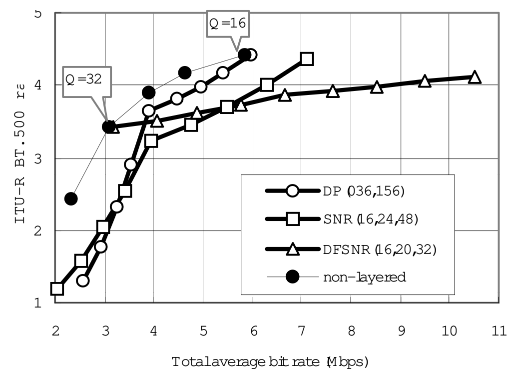


**Figure 10. Comparison of three scalable codings; "flower", M=3 N=12**

## 4 SIMULATION STUDY WITH A SIMPLE BUFFER MODEL

Results obtained in the previous sections correspond to a single stream with prioritized random loss in the lower layers. The next step is to see if these performance results hold in the transmission of layered video streams over a network where limited resources must be shared among several streams. If these streams are transmitted in VBR mode, bursts from different streams arriving at a network buffer at the same time may lead to buffer overflow. Important quality degradation may then occur if such bursts pertain to the base layers of the video streams. We show Figure 11 the traffic characteristics for three-layer DP and SNR layered streams, in terms of the instantaneous rate per layer produced by the encoder. Obviously, data partitioning contains higher peaks in the base layer than SNR because all DCT coefficients of I pictures are included in it. There are more bits in I pictures in SNR scalability than in data partitioning, but they are split over the three layers. Therefore, data partitioning is expected to be more sensitive to synchronization of I pictures from different streams, because in such events there is the risk of packet loss in the base layer.

### 4.1 Simulation with a buffer model

We simulate the transmission of several layered streams encoded with the same layering parameters over a simple differentiated services network. A single FIFO buffer of size B is used, and packets belonging to all layers are input to the same buffer. Priority dropping is implemented as follows. Packets in the base layer are dropped upon arrival to the buffer if the buffer is full, a mechanism known as tail drop. Packets in the middle layer and the enhancement layer are dropped upon arrival if the buffer occupancy is larger than ¾B and ½B, respectively. Once a packet is accepted into the queue, it cannot be dropped. We choose B=50kbytes and output rate of 50 Mbps.

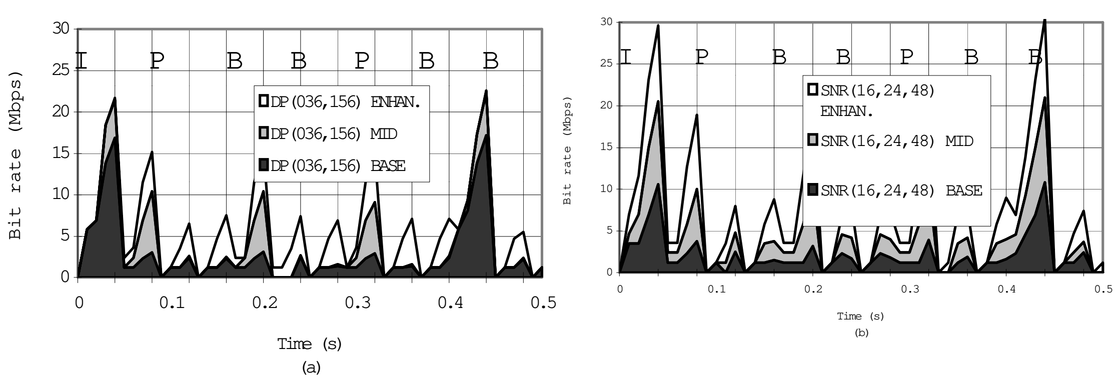


**Figure 11. Data rate; "flower", M=3, N=12**
**(a) data partitioning (036,156), (b) SNR scalability (16,24,48)**

Our simulation experiment is constructed as follows. We consider stream 1 to consist of the sequence "flower" (5 sec long) repeated over and over and we consider all other streams (2 to n-1) by concatenating random selections from the set of three sequences "barca" (4sec long), "basketball" (8.8sec long) and "baseball" (14.84sec long). Quality measurements are made on stream 1. To obtain stable results, we observe the picture quality of stream 1 ("flower") at 8 different starting points: every 50 seconds from 150 sec through 500 sec. All simulations choose sequences in the same order. We give in Table 2 the average bit rate averaged over all four sequences and for the "flower" sequence alone under various layering structures. We note that the total average bit rate for "flower" is 10-15% higher than the average for all four sequences. Ideally, one should be able to accommodate 8-9 streams over the 50Mbps link.

**Table 2 Average bit rate for various cases of scalable coding**

| scalable coding | mean bit rate (Mbps/sequence) 4 sequence | | | mean bit rate (Mbps) "flower" | | |
|---|---|---|---|---|---|---|
| | total | base layer | mid layer | total | base layer | mid layer |
| data partitioning(016,146) | 5.415 | 2.915 | 0.948 | 5.973 | 3.083 | 1.303 |
| data partitioning(036,156) | 5.415 | 2.451 | 1.021 | 5.973 | 2.517 | 1.354 |
| SNR scalability(16,20,32) | 6.598 | 2.728 | 1.855 | 7.586 | 2.672 | 2.339 |
| SNR scalability(16,24,48) | 6.255 | 2.108 | 1.590 | 7.125 | 1.942 | 1.966 |
| non-layered | 5.314 | - | - | 5.872 | - | - |

### 4.2 Results of the buffer model simulation

In Figure 12, we show the picture quality of stream 1 when n streams are input to the buffer for three-layer data partitioning and SNR scalability. In Figure 12(a) we show the average of 15 observations, and in Figure 12(b) we show the average minus the standard deviation for the 15 observations. The latter represents a lower bound on quality for at least 85% of the samples if one were to assume that the samples follow a normal distribution.

With layered coding, we obtain fair quality or better (above ITU-500 rating of 3) even if the combined input rate to the buffer is larger than the link capacity (at 8-9 sequences). Both data partitioning and SNR scalability provide stable picture quality up to n=5 sequences, then the quality gracefully degrades as compared to sudden quality drops in non-layered coding.

Data partitioning provides slightly more graceful degradation than SNR scalability when picture quality is greater than the quality threshold of ITU-500 3, i.e., up to n=10-11. Beyond that point, SNR scalability outperforms data partitioning although again it is not much of a factor because quality levels lower than ITU-500 rating 3 are not acceptable. Data partitioning is more sensitive to I picture synchronization. Figure 12 shows some sudden dips around n=8 and some for higher value of n. For n=8, the "flower" sequence starting at 140 sec synchronizes with three other streams at I pictures.

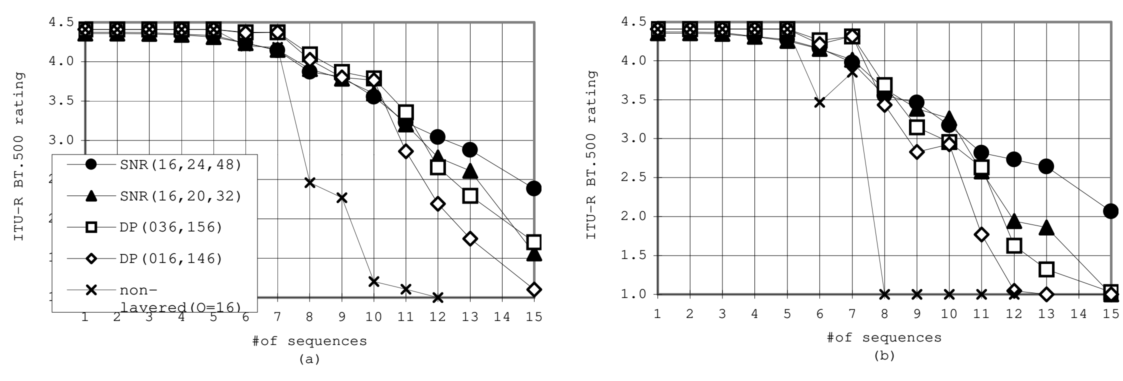


**Figure 12. Picture quality versus the number of sequences transmitted;**
**3-layer, M=3 N=12, link rate=50Mbps, buffer size=50kbytes;**
**(a) average quality, (b) lower bound (average-standard deviation) quality**

**Table 3 Statistical results of the simulation with the buffer model**

| # of seq. | DP(036,156) | | | | | SNR(16,20,32) | | | | | SNR(16,24,48) | | | | | non-layered(Q=16) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | total input rate(Mbps) | load | error (%) | | | total input rate(Mbps) | load | error (%) | | | total input rate(Mbps) | load | error (%) | | | total input rate(Mbps) | load | error (%) |
| | | | base | mid | enh | | | base | mid | enh | | | base | mid | enh | | | |
| 4 | 21.7 | 0.43 | 0.0 | 0.0 | 0.0 | 26.7 | 0.53 | 0.0 | 0.2 | 1.6 | 25.2 | 0.50 | 0.0 | 0.0 | 1.1 | 21.3 | 0.43 | 0.0 |
| 5 | 27.1 | 0.54 | 0.0 | 0.3 | 0.3 | 33.3 | 0.67 | 0.0 | 0.5 | 4.8 | 31.5 | 0.63 | 0.0 | 0.2 | 3.1 | 26.6 | 0.53 | 0.1 |
| 6 | 32.5 | 0.65 | 0.2 | 0.6 | 1.3 | 40.0 | 0.80 | 0.0 | 1.6 | 10.8 | 37.8 | 0.76 | 0.0 | 0.6 | 7.1 | 31.9 | 0.64 | 0.3 |
| 7 | 37.7 | 0.75 | 0.3 | 1.3 | 4.0 | 46.3 | 0.93 | 0.1 | 3.5 | 21.4 | 43.8 | 0.88 | 0.0 | 1.3 | 14.3 | 37.0 | 0.74 | 0.7 |
| 8 | 43.3 | 0.87 | 0.3 | 2.0 | 9.1 | 53.0 | 1.06 | 0.1 | 6.7 | 36.1 | 50.2 | 1.00 | 0.0 | 2.2 | 24.8 | 42.5 | 0.85 | 1.7 |
| 9 | 48.6 | 0.97 | 0.4 | 4.3 | 18.1 | 59.5 | 1.19 | 0.1 | 12.5 | 50.3 | 56.3 | 1.13 | 0.0 | 4.6 | 36.5 | 47.7 | 0.95 | 4.5 |
| 10 | 54.0 | 1.08 | 0.7 | 7.8 | 30.1 | 66.1 | 1.32 | 0.2 | 20.2 | 64.4 | 62.6 | 1.25 | 0.1 | 8.0 | 49.5 | 53.0 | 1.06 | 9.5 |
| 11 | 59.5 | 1.19 | 0.8 | 12.3 | 42.7 | 72.8 | 1.46 | 0.4 | 28.5 | 76.0 | 68.9 | 1.38 | 0.1 | 12.2 | 60.9 | 58.4 | 1.17 | 15.5 |
| 12 | 64.8 | 1.30 | 1.3 | 17.2 | 55.5 | 79.3 | 1.59 | 0.7 | 37.6 | 84.7 | 75.1 | 1.50 | 0.1 | 17.3 | 71.0 | 63.6 | 1.27 | 21.8 |
| 13 | 70.2 | 1.40 | 1.9 | 23.1 | 66.8 | 85.8 | 1.72 | 1.3 | 46.9 | 90.8 | 81.3 | 1.63 | 0.2 | 23.1 | 79.2 | 68.9 | 1.38 | 27.5 |
| 15 | 81.1 | 1.62 | 3.3 | 37.9 | 83.2 | 99.1 | 1.98 | 3.3 | 65.0 | 96.9 | 93.9 | 1.88 | 0.4 | 37.3 | 90.1 | 79.6 | 1.59 | 37.2 |

overloaded

For n = 5 or less, non-layered coding with uniform dropping provides the same performance than layered coding with priority dropping. One could use admission control to limit the number of streams thus achieving the same performance results without the complexity of layering and priority dropping. However, these results correspond to a single video sequence. Differences in sequences due to randomness of the content and burstiness of traffic characteristics make accurate provisioning very difficult, thus maintaining the need of a mechanism that provides graceful degradation under congestion occurrences.

In Table 3, we show how packet loss is distributed among the layers for data partitioning and SNR scalability. We also present the percentage of packet loss for the non-layered stream for comparison. We observe that indeed in all cases with priority dropping the packet loss rate is higher in the enhancement layer than in the middle and base layers. This pattern holds even when the input rate exceeds the output rate, i.e., when 10 or more sequences are present. That explains the graceful degradation even for high load. In this situation, the base layer only sees a 3.3% of packet loss for data partitioning and one case of SNR scalability, while the enhancement layer is being almost completely dropped.

When the size of the base layer is comparable to the size of other layers, protecting the base layer has a very significant impact on the reconstructed quality. We can see in Figure 12 that for a load of 8 or more streams DP and SNR combined with priority dropping outperform uniform dropping by more than 100%. Even at n=6, where the combined input rate is lower than the link capacity, priority dropping already outperforms uniform dropping because it is more resilient to instantaneous bursts. In contrast, if the base layer is small as assumed in [3], uniform and priority dropping would be closer because uniform dropping drops higher layer packets with higher probability. However, small layer size is an unrealistic assumption, at least for MPEG-2 layered coding: under DP, the base layer must carry motion vector information and DC coefficients for all frames to provide acceptable quality and that accounts for a base layer of a significant size. Similar arguments hold for SNR scalabilities, which are limited by their overhead. Only spatial scalability can provide a base layer of small size, but it is further limited by its complex implementation.

## 5 CONCLUSIONS

In this paper we have combined layered coding for MPEG-2 video with priority dropping as made possible by the Internet DiffServ architecture. We initially explored various layering parameters for two-layer data partitioning, and we found the order in which DCT coefficients should be assigned to the enhancement layer to provide optimum quality degradation. We have also found that dropping B pictures entirely provides unacceptable quality degradation and, instead, motion vector and DC coefficients should always be part of the base layer. The results also show that there is a choice to be made between graceful quality degradation and bandwidth reduction in two-layer data partitioning, and that multi-layer data partitioning provides graceful degradation over a wide range of bandwidth. We have compared data partitioning with two forms of SNR scalability, and we have seen that the advantages of layering are the strongest in data partitioning due to its negligible overhead and ease of implementation, and are weakened in SNR scalability because of its overhead. We have also quantified the effect of stream interaction on the performance properties. Our findings from a simple DiffServ network model show that data partitioning always provides slightly better quality than SNR scalability although it is more sensitive to synchronization of I frames. Our simulation also shows that priority dropping outperforms uniform dropping by more than 100% due to the protection of the base layer. In all our simulations we have used a perceptual quality metric based on the human visual system to provide accurate results.


## ACKNOWLEDGEMENTS

We wish to thank Stefan Winkler, EPFL for providing the source code for the perceptual distortion metric (PDM) we used in our study. We also thank Olivier Verscheure, EPFL and Christian J. van den Branden Lambrecht, Lucent Technologies for valuable advice and for providing video materials.


## REFERENCES


1. Boyce, J. Gaglianello, R. "Packet Loss Effects on MPEG video sent over the Public Internet", Proceedings of ACM MultiMedia, 1998.
2. J-C. Bolot, T. Turletti, Experience with rate control mechanisms for packet video in the Internet, Computer Communication Review, vol. 28, no. 1, January 1998
3. S. Bajaj, L. Breslau, S. Shenker "Uniform versus Priority dropping for Layered Video", ACM SIGCOMM 98
4. S. Blake et al. "An Architecture for Differentiated Services", IETF RFC-2475, Dec 1998
5. J. Heinanen, F. Baker, W. Weiss, J. Wroclawski, "Assured Forwarding PHB Group", IETF RFC-2597, Jun 1999
6. ISO/IEC, "Generic coding of moving pictures and associated audio information" (MPEG-2), ISO/IEC 13818-2, 1995
7. ISO/IEC, "Generic coding of audio-visual objects" (MPEG-4), ISO/IEC 14496-2, 1999
8. R.Aravind and et al., "Packet loss Resilience of MPEG-2 scalable video coding algorithms", IEEE Trans. on CSVT, Vol. 6, No. 5, Oct. 1996
9. D.Wilson and M.Ghanbari, "Transmission for SNR scalable two-layer MPEG-2 coded video through ATM networks", 7th Internarional Workshop on packet video, pp.37-42, March 1996
10. D.Wilson and M.Ghanbari, "An efficient loss priority scheme for MPEG-2 variable bit rate video for ATM networks", Proceedings of IEEE Global Telecommunications Conference, GLOBECOM'96, Nov. 1996
11. T.V.Lakshman, A. Ortega, and A.R. Reibman, "VBR video: Tradeoffs and potentials", Proceedings of the IEEE., Vol. 86, No.5, May 1998
12. H. Schulzrinne,S. Casner,R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", IETF RFC-1889, Jan. 1996
13. D. Hoffman,G. Fernando,V. Goyal, and M. Civanlar, "RTP Payload Format for MPEG1/MPEG2 Video", IETF RFC-2250, Jan. 1998
14. Stefan Winkler, "A Perceptual Distortion Metric for Digital Color Video", Proceedings of SPIE Human Vision and Electronic Imaging, vol. 3644, pp. 175-184, Jan. 1999
15. ITU-500-R, "ITU-500-R Recommendation BT.500-8 - Methodology for the subjective assessment of the quality of television pictures", 1998
16. Cheul-hee Hahm and Jae-kyoon Kim, "An adaptive error concealment in SNR scalable system", Proceedings of SPIE VCIP'95, pp.1380-1387, Apr. 1995