

A BANDWIDTH ASSIGNMENT ALGORITHM ON A SATELLITE CHANNEL FOR VBR TRAFFIC

NEDO CELANDRONI,* MARCO CONTI, ERINA FERRO, ENRICO GREGORI AND
FRANCESCO POTORTÌ
CNUCE, Institute of National Research Council (CNR), Via S. Maria 36, I-56126 Pisa, Italy

SUMMARY

Variable bit rate (VBR) video is currently by far the most interesting and challenging real-time application. A VBR encoder attempts to keep the quality of video output constant and at the same time reduces bandwidth requirements, since only a minimum amount of information has to be transferred. On the other hand, as VBR video traffic is both highly variable and delay-sensitive, high-speed networks (e.g. ATM) are generally implemented by assigning peak rate bandwidths to VBR video applications. This approach may, however, be inefficient in a satellite network based on a TDMA scheme. To overcome this problem, we have designed a demand assignment satellite bandwidth allocation algorithm in TDMA, named V2L-DA (VBR 2-Level Demand Assignment), which manages the VBR video traffic according to a dynamic bandwidth allocation algorithm. In this paper we discuss how to tune the proposed algorithm in order to optimize network utilization when MPEG-1 VBR video traffic is being transmitted. Our results indicate that most of the time only 40% of the peak rate bandwidth is needed to satisfy the VBR source, so the remaining 60% of the peak rate bandwidth can be used to transmit the datagram traffic queued in the network stations. © 1997 John Wiley & Sons, Ltd.

KEY WORDS: satellite; TDMA assignment; real-time traffic; non-real-time traffic; VBR traffic; MPEG coding; traffic model

1. INTRODUCTION

A variety of new applications, such as the transport of pictures, teleconferencing and video, and a large volume of interactive computer data must be supported in an integrated manner by today's high-speed networks. These applications have diversified quality-of-service (QoS) requirements and traffic statistics (ranging from the high burstiness of video applications to the smooth continuous traffic generated by large file transfers).

Variable bit rate (VBR) video is currently by far the most interesting and challenging real-time application. A VBR encoder attempts to keep the quality of video output constant and at the same time reduces bandwidth requirements, since only a minimum amount of information has to be transferred.

As VBR video traffic is both delay-sensitive and has a high degree of burstiness, it is commonly believed that a bandwidth corresponding to the source peak rate must be reserved for this application to satisfy its QoS requirements. In our TDMA satellite network the peak rate allocation is extremely inefficient, because only the station reserving the bandwidth is authorized to utilize it. Taking into consideration that the ratio between the

peak and the average bit rate for VBR video is generally high (for example, in the MPEG-1 movie used in this paper it is equal to five; see Section 3.2), this implies that a large portion of the network bandwidth remains unused unless the station transmitting the movie has enough low-priority traffic as well. To increase the efficiency in bandwidth allocation, we designed an algorithm which dynamically allocates bandwidth to VBR video on the basis of the actual source rate. This algorithm is integrated into a *centralized control* demand assignment satellite access scheme, named V2L-DA (VBR 2-Level Demand Assignment). The results presented here show that V2L-DA can efficiently and simultaneously support two classes of traffic, called *datagram* and *stream* respectively.

According to the traffic categories as defined in the ATM Forum TM4.0 ('ATM service categories'),¹ the first class includes all the jitter-tolerant applications (unspecified bit rate (UBR) and available bit rate (ABR) service categories), while the second includes all the real-time applications (constant bit rate (CBR) and variable bit rate (VBR) service categories). This paper focuses on the efficiency of V2L-DA when the stream service is used to transmit the VBR traffic, though the allocation scheme is also suitable for CBR traffic.

The paper is organized as follows. Section 2 presents the satellite bandwidth allocation scheme together with the criteria to dimension the buffers needed to compensate for the jitter of VBR data. Section 3 describes the MPEG traffic characteristics

* Correspondence to: N. Celandroni, CNUCE, Institute of National Research Council (CNR), Via S. Maria 36, I-56126 Pisa, Italy. Email: n.celandroni@cnuce.cnr.it
Contact grant sponsor: National Research Council (CNR)

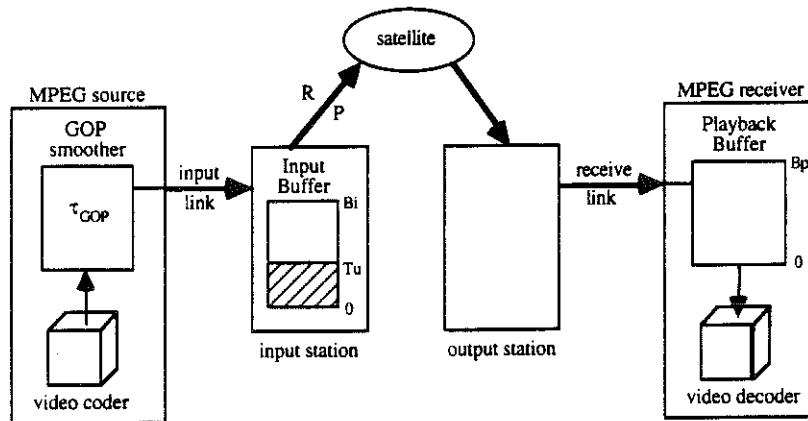
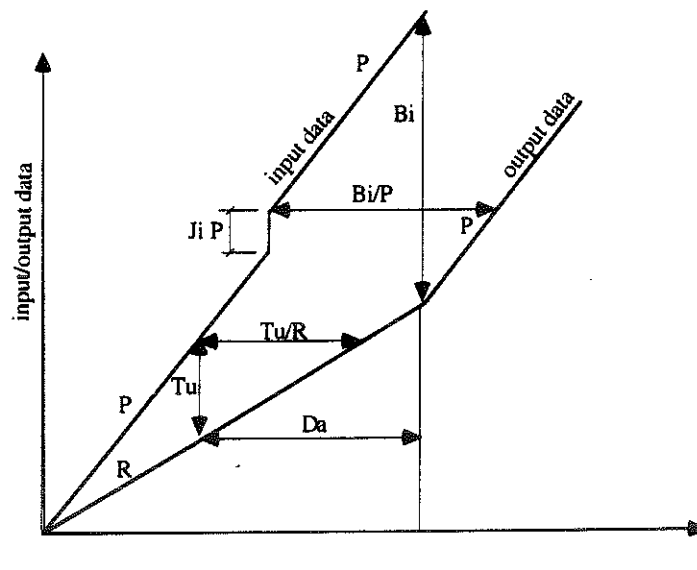


Figure 3. Communication chain

Figure 4. Buffer requirements and queuing time relationships in transition from allocation R to allocation P (worst case)

from the output of the MPEG coder to the input of the video decoder is evaluated as

$$D_t = \tau_{\text{GOP}} + \tau_i + J_o + \tau + \tau_r + J_r + D_p \quad (6)$$

where τ_{GOP} is the GOP duration time and τ_i and τ_r are the latencies of the input link and receive link respectively. To account for the image delay, from the moment a picture is taken to when it is shown, one should also add the coding and decoding delays of MPEG. Also, delays induced by framing and packetization should be accounted for in (6) if a detailed estimation of the end-to-end delay is required.

We have followed the data path from the MPEG source to the MPEG receiver and have computed the overall link delay, taking into account what happens when the allocation switches from R to P . Now we consider the opposite switch, which happens when the current allocation is at the high level P . In this state the station maintains a *virtual input*

queue, i.e. a counter which is incremented at the rate of the input traffic and decremented at rate R . The counter is never incremented above zero, so it always contains a non-positive number. When the input traffic has a rate greater than R , the virtual queue is zero. When the virtual queue drops below the threshold $-T_d$, the station issues a request for the low allocation level R . This mechanism is specular with respect to the one used for switching from allocation R to allocation P , so the threshold is computed with the same criterion, and

$$T_d = J_i R \quad (7)$$

3. TUNING THE PARAMETERS OF THE ALLOCATION ALGORITHM FOR MPEG APPLICATIONS

Below we study how the parameters P and R must be set to optimize the utilization of the satellite network capacity when the stream traffic is an MPEG-1-encoded movie.

3.1. MPEG-1 traffic characteristics

An uncompressed video source may generate bits at rates as high as hundreds of Mbps. Data compression techniques are therefore used to reduce the video source bit rate which is transmitted over the network.

MPEG-1 is a specification for coding video, developed by the ISO Joint Motion Pictures Experts Group.^{9,10,13} The standard is well suited for a large range of video applications at a variety of bit rates. Compression of a combination of video and audio information, particularly for 'movie' applications, is also possible. Typical compression ratios are in the range from 50:1 to 200:1.¹¹

MPEG-1 is an interframe coder. Coders in this class exploit, in addition to intraframe coding, the temporal redundancy that exists between adjacent frames by predicting the next frame from the current one. A key feature that distinguishes MPEG-1 from previous coding algorithms is bidirectional temporal prediction. For this type of prediction, some of the video frames are encoded using two reference frames, one in the past and one in the future, which leads to higher compression gains.

As indicated above, when applying MPEG-1 to video, one of three different coding modes can be used for each frame. The terminology used for the resulting frame is related to the model used as follows:

- I-frame—intraframe-coded,
- P-frame—predictive-coded with reference to the previous P- or I-frame,
- B-frame—bidirectional predictive-coded.

I-frames provide access points for random access but only with moderate compression. Predictive-coded frames are generally also used as a reference for future P-frames. The frames of type B provide the highest amount of compression but require both a past and future reference prediction.

In the encoded sequence the frames are arranged into groups as shown in Figure 5. In this case a group consists of 12 frames—one I-frame, three P-frames and eight B-frames. Figure 5 also shows the relationship between the frames. We can see that I-frames are independent, P-frames are predicted and B-frames are bidirectionally predicted.

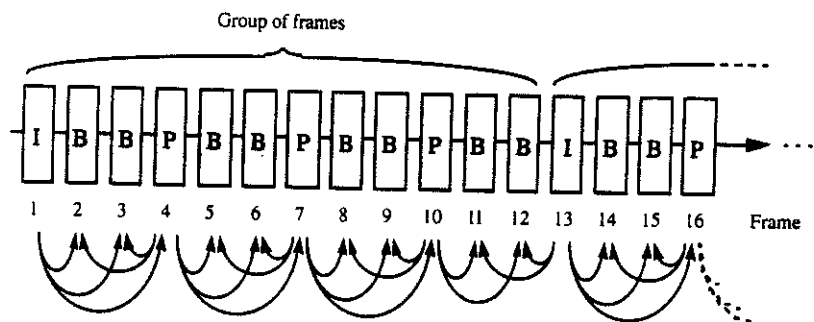


Figure 5. Sequence of MPEG-1 video frames and their relationship

Figure 6 shows a small extract from the output of the MPEG-1-coded Star Wars movie released by M. Garret at Bellcore. Specifically, frames are coded into GOPs as defined in Figure 5 (i.e. the frame pattern is IBBPBBPBBPBB).

As shown in Figure 6, the bandwidth required to transmit consecutive frames is highly variable and very much depends on the frame types I, P and B. Furthermore, as expected (owing to the coding scheme algorithm), the shape of the output is repeated every 12 frames.

To simplify the study of the bandwidth allocation algorithm, we assume a 12-frame pre-buffering before the transmission and we only look at the aggregate bit rate produced by the coding of a group of 12 frames. Thus, hereafter, we only consider the aggregate sequence obtained by summing the amount of bits generated in every GOP. This aggregate sequence has a period of 12 frames. In Table I the basic statistics of the MPEG-1 Star Wars aggregate sequence are presented.

Below we present a model developed to characterize the aggregate sequence obtained by the output of an MPEG-1 coder. More details on the modelling of an MPEG-1 video source can be found in Reference 12. The model is used as a synthetic traffic descriptor for analysing the performance of various bandwidth allocation schemes without requiring the huge amount of data describing the actual traces.

3.2. The model

The analysis presented in Reference 12 shows that in the aggregate sequence there is both a short-range dependence which lasts for a small number of groups (15 s) and long-range dependences which last for thousands of groups (10–20 min). To capture both types of dependences, a bidimensional Markov chain $\{L_k, H_k | k \geq 0\}$ is used, where H_k is the k th GOP size and L_k is the status of a low-frequency process modulating the k th GOP size.

$\{H_k | k \geq 0\}$ describes the bit rate per group of an MPEG encoder. To avoid unnecessary complexity (in the state space of $\{H_k | k \geq 0\}$), we quantize the bit rate in a uniform way into a number of levels. The number of quantization levels for the process

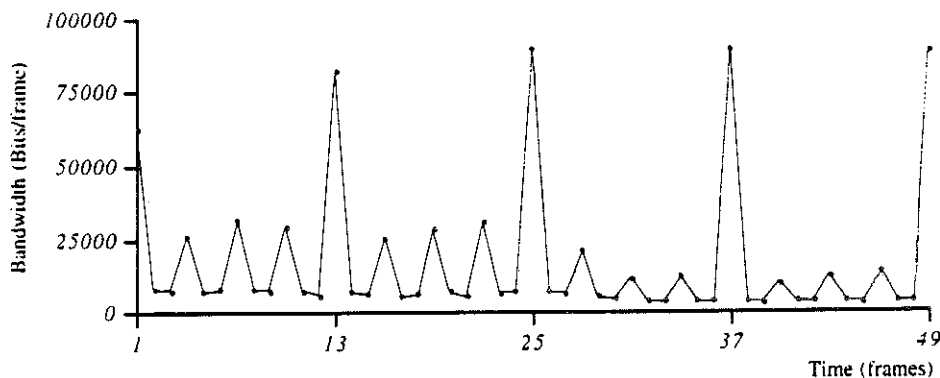


Figure 6. Part of MPEG-1 coder trace, revealing group length and frame pattern

Table I. Star Wars basic statistics

	μ	σ	<i>min</i>	<i>max</i>
GOP statistics (kbits)	187.2	72.5	77.754	932.71

will hereafter be denoted by N , i.e. $H_k \in \{0, 1, \dots, N-1\}$. Specifically, let *max* and *min* denote the maximum and minimum bit rates observed in the aggregate sequence; the possible bit rates are quantized with a constant step size $\Delta = (\text{max} - \text{min})/N$. By applying this quantization procedure, the average bit rate associated with $H = i$ is

$$\text{min} + (i + 1) \Delta \quad (8)$$

The *min* and *max* values are reported in Table I together with the average μ and standard deviation σ of the GOP size.

We use the GOP as the time unit. To represent the low-frequency component of our source, a modulating process $\{L_k | k \geq 0\}$ is included in the model as well ($L_k \in \{0, 1, 2, \dots, M-1\}$). In the trajectories of the Markov chain the H_k value frequently changes (every few time units, on average), while the L_k value changes on a much longer time scale (about 70–100 time units).

The transition probabilities of the Markov chain $\{L_k, H_k | k \geq 0\}$ are estimated from the MPEG 1 Star Wars trace by applying the procedure presented in Reference 12. Specifically, the model used for the results presented in this paper is obtained with the parameters $M=8$ and $N=8$. The accuracy of this model was investigated in Reference 12. The results obtained indicate that both the qualitative properties (i.e. burstiness and overall appearance of the traces) and the statistical properties (maximum, minimum, average, standard deviation and autocorrelation function) of the GOP size sequence generated with our Markov model are very similar to those of the real trace.

4. TUNING THE ALLOCATION ALGORITHM: A CASE STUDY BASED ON AN MPEG-ENCODED STAR WARS MOVIE

In this section we study the setting of the parameters P and R to transmit the MPEG 1 Star Wars movie on the satellite link, together with some low-priority data. The choice of the values of P and R is made by minimizing the bandwidth allocation cost. Minimizing the end-to-end delay is not a primary target, since the delays already caused by the MPEG coding, the pre-buffering and the satellite transmission make this technology unsuitable for interactive video applications. As described in Section 3, we divided the source bit rate into eight levels (i.e. $H=0, \dots, 7$). However, as shown in Reference 12, states with $H=5, 6$ and 7 are rare (i.e. they only occur a few times in the 2 h sequence, and never consecutively); thus we do not consider these states for bandwidth allocation. In fact, a two-level allocation scheme (like the one presented here) would be extremely inefficient if P were set equal to the highest bandwidth level (i.e. $H=7$). Hereafter it is therefore assumed that when a group with throughput higher than $H=4$ occurs, the traffic exceeding the allocated bandwidth is transmitted with a best-effort policy (e.g. as datagram traffic).

The bandwidth allocation problem, as shown in Table II, is thus reduced to the study of four cases. P is set to the bit rate corresponding to $H=4$, while the parameter R can be set equal to the bit rate corresponding to one of the states 0, 1, 2, 3.

To identify the optimal allocation parameter setting, we evaluate the cost of transmitting a VBR video source and some EDP data traffic. We assume

Table II. Cases of allocation positions

Case	Allocation level positions
1	$R=0$ and $P=4$
2	$R=1$ and $P=4$
3	$R=2$ and $P=4$
4	$R=3$ and $P=4$

a cost equal to one for each unit of bandwidth allocated to the VBR traffic. The cost for each unit of bandwidth reserved for the low-priority traffic is assumed to be less than one and will be denoted by β ($0 < \beta < 1$).

As stated in Section 2, the unused stream bandwidth (i.e. booked for VBR traffic but not used for it) can be used to transfer datagram data. The maximum amount of such data is U , which is the difference between the peak and the average bandwidth of a VBR video:

$$U = \sum_i \pi_i (P - i)$$

where π_i is the probability of a bit rate i for the VBR source.

When studying the bandwidth allocation, we considered U as the maximum amount of data that could be transmitted by the station. The bandwidth exceeding U always needs to be allocated as a datagram bandwidth for all possible parameter settings (see Table II) and thus its cost does not depend on the bandwidth allocation strategy. Hence we assume that the station has to transfer a percentage p of U , and pU is the amount of data traffic.

4.1 Ideal case

To simplify the presentation, in the computation of the bandwidth allocation costs we first focus on an ideal case in which the transients for switching between the two allocation levels (P and R) are negligible; that is, when the bit rate is greater or lower than R , the allocated level is P or R respectively.

Under this assumption a station with an MPEG-1 video source and pU data traffic has to pay the following two costs for the required bandwidth.

Excess stream cost

This cost takes into account only the real-time bandwidth that is not used by the video. As shown in Figure 7, this cost is the difference between the allocated bandwidth (bold line) and the source bit rate.

The amount of bandwidth used for transmitting the MPEG-1 traffic is obviously the same in all the allocation cases (see Table II). Unused video bandwidth can be used by the same station to transfer its datagram data (if any), but it is allocated as stream bandwidth and thus has a cost equal to 1. The excess stream cost is therefore given by

$$\lambda_u = \sum_{i=0}^4 \pi_i (A(i) - i)$$

$$A(i) = \begin{cases} R, & i \leq R \\ P, & i > R \end{cases} \quad (9)$$

where π_i is the probability of bandwidth level i and $A(i)$ is the allocated level. The probability π_i is computed from the Markov chain characterizing the source (see previous section) and is given by

$$\pi_i = \sum_l P\{L_k = l, H_k = i\}$$

Requested data bandwidth cost

A portion of the unused bandwidth in each TDMA frame (see Figure 1), i.e. the datagram bandwidth, can also be used for data transfer if the resources given by (9) are not sufficient to transmit all the low-priority data of the station. The additional bandwidth needed for the datagram is then

$$\lambda_d = [pU - \lambda_u]^+ \quad (10)$$

where $[y]^+$ is y for positive y and zero otherwise.

4.2 Real case

In a real case, as explained in Section 2, transient intervals occur both to obtain and to release the $(P - R)$ extra bandwidth. Figure 8 highlights the differences between the ideal case and the real one. Specifically, in the allocation and deallocation transient intervals the broken line represents the bandwidth allocation in the ideal case, while the full line shows the real bandwidth allocation level. Comparing Figures 7 and 8 shows that there are differences only in the two hatched areas which correspond to the allocation and deallocation transients. To take into account the effect of these transients, the allocation and deallocation costs must be added to the cost function.

Allocation cost

In the ideal case the excess stream cost was computed by assuming in this period an allocation level equal to P , and hence during this transient period it provides a cost overestimation with respect to the real case in which the allocation level is still R . The real cost is negative. During this transient the buffer size increases and this backlog is transmitted by using future unused bandwidth. For this reason it must be considered as a negative cost with respect to the excess stream cost computed in the ideal case. To remove this overestimation, we need to subtract the allocation transient area (see Figure 8). The allocation transient lasts for the sum of the time interval, say x , required to fill the buffer up to the level T_u , plus the allocation time D_a .

Hence the hatched area is

$$(D_a + x)(P - R)$$

The exact computation of x is complex. However,

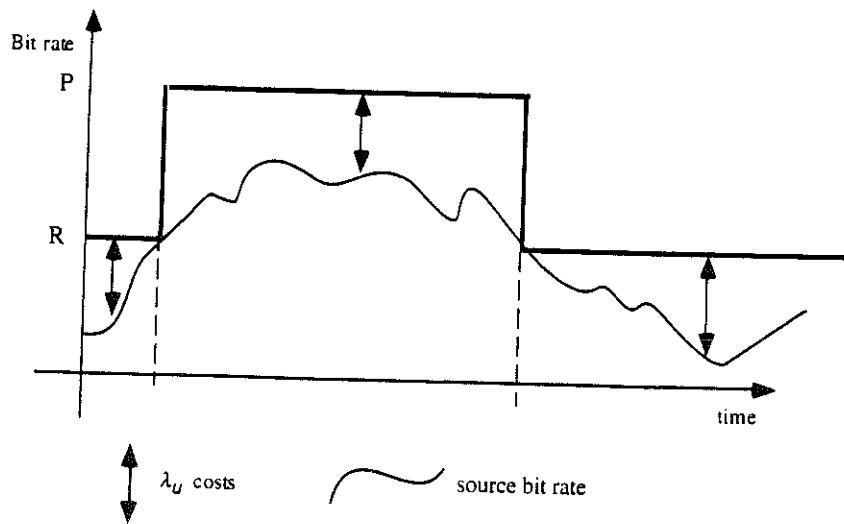


Figure 7. Allocation costs in ideal case

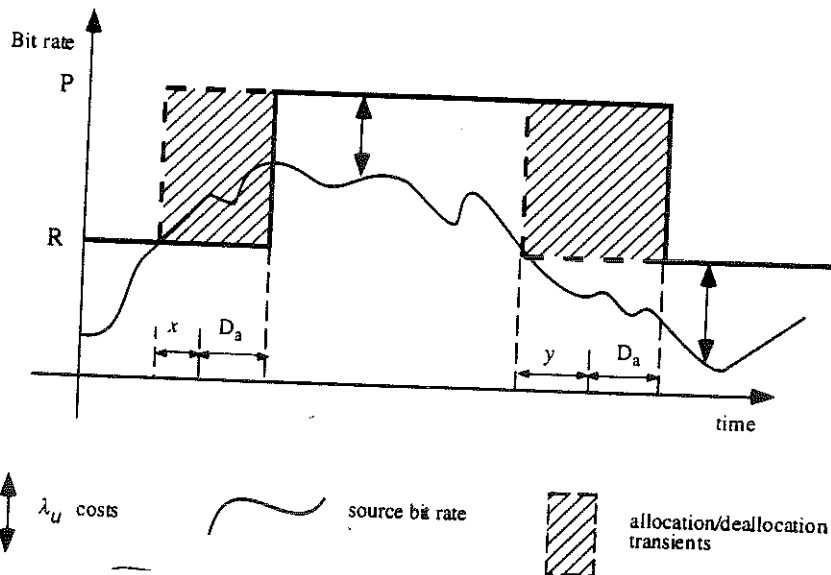


Figure 8. Real-time bandwidth lost owing to dynamic allocation

using our source model, it is possible to estimate the value of x when the source changes its level from i to j , i.e. $x_{i,j}$ *. In this case

$$x_{i,j} = \frac{T_u}{H_j - R}$$

and hence the average allocation cost for the bandwidth allocation (λ_{up}) is

$$\lambda_{up} \approx (P - R) \sum_{i=0}^R \sum_{j>R} (D_a + x_{i,j}) \pi_i q_{i,j}$$

where π_i is the steady state probability for the VBR video to be in level i , $q_{i,j}$ is the probability that the

video moves from level i to level j , and hence $\sum_{i=0}^R \sum_{j>R} \pi_i q_{i,j}$ is the frequency of the R -to- P transitions in the bandwidth allocation levels.

Deallocation cost

In the real case an additional cost is introduced whenever a station wants to release the $(P - R)$ bandwidth, because a $(D_a + y)$ delay (at least 500 ms) must be kept into account, i.e. the time between the transmission of a relinquish request and confirmation from the satellite network.

The situation is symmetrical with respect to the previous case, so the average cost for the bandwidth deallocation (λ_{down}) is

$$\lambda_{down} = (P - R) \sum_{i>R} \sum_{j \leq R} (D_a + y_{i,j}) \pi_i q_{i,j}$$

$$y_{i,j} = \frac{T_d}{R - H_j} \tag{11}$$

* This computation is performed under the assumption that during the transient the source level j does not change.

By exploiting the cost functions defined in this section, we are now ready to compare the global cost for each of the bandwidth allocation parameter settings defined in Table II. The cost is clearly the sum of the following components:

$$cost = 1 \cdot \lambda_{down} + 1 \cdot \lambda_u + \beta \lambda_d - 1 \cdot \lambda_{up} \quad (12)$$

Results obtained by applying formula (12) to the four allocation cases of Table II are presented in Figures 9(a)–9(c). These figures assume $D_a = 500$ ms and the price for a data bandwidth level, β , set to 0.4, 0.6 and 0.8 respectively.

Clearly, case 2 is the winner in this comparison, irrespective of β and p . Hence we can conclude that allocating to the VBR video a level corresponding to $H = 1$ as the minimum level (the R parameter) and a level corresponding to $H = 4$ as the maximum level (the P parameter) is the optimal solution, whatever the amount of data traffic to be transmitted by the station.

The above results show that the best values for

P and R must be calculated considering $H = 1$ and 4 respectively. Table III summarizes the results obtained so far for our case study, in which one GOP is equal to 12 frames and the frame rate is 24 frames/s. We assume that some sort of resource reservation is made on the input and receive links (see Figure 3), so that packets traversing those links have a bound on their maximum delay. The values assumed for the latency and the max jitter are typical of a wide-area terrestrial route.

It is interesting to note that the buffer sizes B_i and B_p are proportional to the data rate of the MPEG stream, while the delays are not.

5. CONCLUSIONS

The V2L-DA algorithm guarantees the peak bandwidth (P) for a VBR video application, while maintaining good efficiency in the overall channel bandwidth allocation. In fact, the throughput of a VBR application is often several times lower than its peak throughput (five times in our case), and this leads

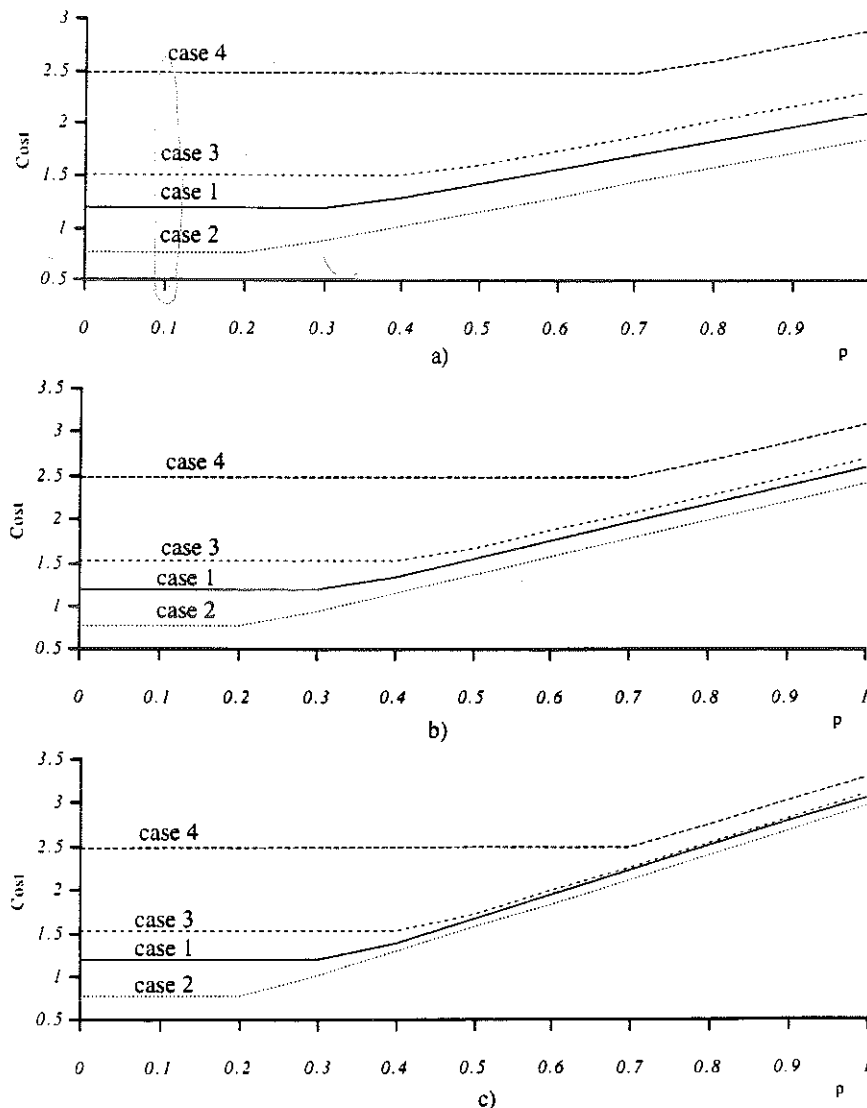


Figure 9. Cost with β equal to (a) 0.4, (b) 0.6 and (c) 0.8

Table III. Key parameter values for Star Wars case study

Parameter	Reference	Value
GOP peak rate of stream	Table I	1865 kbit/s
GOP mean rate of stream	Table I	374 kbit/s
Booked bandwidth P	Figure 2	1220 kbit/s
Low-level bandwidth R	Figure 2	583 kbit/s
GOP length τ_{GOP}	Figure 3	500 ms
Link latencies $\tau_i = \tau_r$	Figure 3	50 ms
Satellite link latency τ	Figure 3	250 ms
Maximum jitter induced by links, $J_i = J_r$	Figure 3	200 ms
Allocation delay D_a	Section 3	500 ms
Threshold levels $T_u = T_d$	Formulae (1), (7)	14 kB
Input station buffer size B_i	Formula (2)	83 kB
Playback delay at MPEG receiver, D_p	Formula (4)	757 ms
Playback buffer size at MPEG receiver, B_p	Formula (5)	226 kB
Reproduction delay D_s	Formula (6)	2360 ms

to an inefficient use of the channel bandwidth if peak rate allocation is adopted. To increase the efficiency in bandwidth allocation, when the throughput of a VBR application is below a certain threshold R , only a bandwidth up to R is actually allocated to this application, while the difference $P - R$ is booked for this application but (until requested) is used by the channel dispatcher to satisfy the datagram traffic of all the network stations. As soon as the throughput of the VBR encoder exceeds R , the channel dispatcher allocates all the bandwidth P already booked by this application. We have discussed the setting of the parameters P and R in order to optimize the utilization of the network capacity. Specifically, by considering the transmission of the trace of a movie produced by an MPEG-1 encoder, the optimal bandwidth allocation for this VBR video application is obtained by setting R to about 40% of the booked bandwidth P . Taking into consideration that most of the time the source bit rate is below R ,¹² it follows that 60% of the bandwidth booked by a VBR video application can be used to satisfy datagram transmissions of all the stations.

In this work we have assumed that, when the traffic exceeds the maximum allocable bandwidth (*spikes*), a best-effort policy is used (e.g. by exploiting the datagram traffic). Although very rare, these events prevent professional quality transmissions. To avoid this limitation, future work will be devoted to extending the stream allocation policy to more than the two bandwidth levels currently used. The highest level should be used to manage the spikes.

ACKNOWLEDGEMENT

This work was carried out in the framework of the Italian co-ordinated project 'Advanced applications

for next generation packet-switching networks' funded by the National Research Council (CNR).

REFERENCES

1. ATM Forum, 'ATM service categories: the benefits to the user', *White Paper*, European Market Awareness Committee, 19XX.
2. N. Celandroni, E. Ferro and F. Potorti, 'DRIFS-TDMA. A proposal for a satellite access distributed-control algorithm for multimedia traffic in a faded environment', *Int. J. Satell. Commun.*, in press.
3. N. Celandroni, E. Ferro and F. Potorti, 'FEEDERS-TDMA. A distributed-control algorithm for satellite channel capacity assignment in a mixed traffic and faded environment', *Int. J. Satell. Commun.*, **15**, (4), 185-195 (1997)
4. N. Celandroni, E. Ferro, N. James and F. Potorti, 'FODA/IBEA: a flexible fade countermeasure system in user oriented networks', *Int. J. Satell. Commun.*, **10**, (6), 309-323 (1992).
5. T. Zein, G. Maral, T. Brefour and M. Tondriaux, 'Performance of the Combined/Fixed Reservation Assignment (CFRA) scheme for aggregated traffic', *Proc. COST-226 Final Symp.*, Budapest, May 1995, pp. 183-198.
6. Dornier Deutsche Aerospace, 'Advanced business communications via satellite', *System Description*, 1992.
7. J. I. Mohammed and Tho Ngoc, 'Performance analysis of Combined Free/Demand Assignment Multiple Access (CFDAMA) protocol for packet satellite', *ICC 94*, 1994, pp. 869-873.
8. Tho Ngoc and S. V. Krishnamurthy, 'Performance of Combined Free/Demand Assignment Multi-Access (CFDAMA) protocol with pre-assigned request slots in integrated voice/data satellite communications', *ICC 95*, 1995, pp. 1572-1576.
9. D. Le Gall, 'MPEG: a video compression standard for multimedia applications', *Commun. ACM*, **34**, (4), 46-58 (1991).
10. L. Chiariglione, 'The development of an integrated audiovisual coding standard: MPEG', *Proc. IEEE*, **83**, (2), 151-157 (1995).
11. D. Minoli and R. Keinath, *Distributed Multimedia through Broadband Communications Services*, Artech House, Boston, MA, 1994.
12. M. Conti, E. Gregori and A. Larsson, 'A study of the impact of MPEG-1 correlations on video-sources statistical multiplexing', *IEEE J. Select. Areas Commun.*, **SAC-14**, (9), 1455-1471 (1996).
13. P. Parcha and M. El Zarki, 'MPEG coding for variable bit rate video transmission', *IEEE Commun. Mag.*, **32**, (5) 54-66 (1994).

* As described in Section 4, the booked bandwidth P is set to 5/8 of the GOP peak rate.