16

On the efficiency of statistical–bitrate service for video

G. Karlsson
Swedish Institute of Computer Science
Box 1263, SE–164 29 Kista, Sweden

G. Djuknic
Lucent Technologies
67 Whippany Road, Room 1A–234
Whippany, NJ 07981–0903, U.S.A.

Abstract

The provisioning of quality of service by means of statistical multiplexing has been an alluring research idea for the last decade of teletraffic research. In this paper we question the efficiency of statistical bitrate service which is the standardized representation of this operational mode for ATM networks. Our argument is that the amount of information needed about a traffic source in order to attain a fair multiplexing gain is beyond what is captured in the standard's three–parameter traffic descriptor.

Keywords

Statistical bitrate, variable bitrate, deterministic bitrate, ATM, video, teletraffic

1 INTRODUCTION

Statistical multiplexing with quality guarantees is often seen as the prime service offering of asynchronous transfer mode networks that should justify the introduction of

Performance of Information and Communication Systems U. Körner & A. Nilsson (Eds.)

ATM in relation to existing deterministically multiplexed telephony networks as well as statistically multiplexed networks without quality guarantees, such as most local area networks and internet protocol based wide area networks. There has consequently been a remarkable interest in the research community on the definition and evaluation of this operational mode.

Statistical multiplexing has been successfully used for data communication during the last three decades and more recently in radio networks. In both cases there is a division of responsibility: the network provides fair access to the transmission capacity and routing; the end–equipment is responsible for the quality of the transmission by means of retransmission and forward–error correction. ATM is breaking this division by asking the network to provide quality guarantees for statistically multiplexed channels. The implicit assumption is that the guarantees would come at only a small loss in multiplexing efficiency, which still would leave a large efficiency gain compared to the use of deterministic multiplexing. We will consider this latter comparison of statistical and deterministic multiplexing for the case when quality of service is required but the information about source behavior is limited to that provided by a leaky–bucket descriptor. Our main interest is video communication.

The Telecommunication Standardization Sector of the International Telecommunication Union has made Recommendation I.371 for the choices of traffic control mechanisms in B–ISDN. We will use the terminology of the Recommendation and refer to statistical multiplexing with quality guarantees as the statistical–bitrate service (SBR) and the deterministic multiplexing as deterministic–bitrate service (DBR). The ATM Forum has chosen the terms variable–bitrate service and constant–bitrate service for these two services.

For video communication over ATM networks there are primarily two causes of information loss to consider: quantization loss in the source coder, and cell loss due to multiplexing overload in the network. In general the quantization loss can be made less perceptual than the cell loss for comparable levels. This, in turn, means that it is better to reduce the bitrate by source coding and allowing at most a small amount of cell loss, compared to nearly lossless source coding and more cell loss in order to get more efficient multiplexing. For example, Heeke reports that the statistical multiplexing gain increases 20 percent for a video conference scene and 40 percent for a television scene when increasing the cell loss rate seven orders of magnitude from 10^{-9} to 10^{-2} [6]. An equally large increase in distortion caused by source coding would most likely allow a much higher reduction in needed transmission capacity for the signal.

The idea has been that going from a truly lossless network service to a virtually lossless one would open up for a reasonable statistical multiplexing gain without compromising the quality which ought to be determined by the source coding loss. In this study we show that this idea, although appealing, cannot always be realized. The ensuing risk is that the more complex SBR service is implemented and yet it does not give any performance improvements compared to a simpler DBR provisioning. The main reason why this risk is not negligible is that the call–acceptance control would need a fair amount of information about the source characteristics in order to ensure

quality at a high multiplexing efficiency. The recommended traffic parameters are often insufficient for this purpose. For example, Heeke's work relies on the measured average rate and its standard deviation in order to calculate the number of identical but independent streams that could be multiplexed onto a link. In reality the procedure would have been that the sender estimates a few traffic parameters for the traffic stream, and the call–acceptance control chooses the number of streams based on those parameters (where different calls would of course have different parameter values). The mean and standard deviations would thus have been calculated from the traffic descriptors, rather than being the true values of the source.

2 ATM TRANSFER CAPABILITIES

According to ITU Recommendation I.371 we may pose the following requirements on the parameters that would be used to describe a forthcoming call: they should be understandable by user or terminal to make conformance possible; they should be useful for the call–acceptance control to meet performance requirements, and finally, the parameters should be enforceable for user and network parameter controls.

Source parameters

The peak rate is a mandatory parameter to specify for all calls.It is simply given as the inverse of the minimum cell distance, measured in time from first bit to first bit (T_{pcr}). The time is treated as a continuous variable despite the fact that ATM transmission is slotted (idle times are filled by empty cells to maintain link synchronization of cell boundary detection at the physical layer). However, the peak rate specification is quantized to 1 638 444 distinct rates (from 1 cell up to 4.3×10^9 cells per second).The peak rate is coupled to a tolerance value for the cell–delay variation which specifies the maximum deviation from the minimum cell–interarrival time specified by the peak rate.

The second rate–tolerance pair is the sustainable rate and the intrinsic burst tolerance ($1/T_{sbr}$ and τ_{ibt}). They are defined by a generic cell rate algorithm. There is also a tolerance value for the cell–delay variation with respect to the sustainable rate. The burst tolerance is measured in seconds. An equivalent burst measure in terms of cells, the so called maximum burst size, is given by $1 + \lfloor \tau_{ibt}/(T_{sbr} - T_{pcr}) \rfloor$.

The worst admissible behavior of a source that is specified by sustainable and peak rates and an intrinsic burst tolerance is an on–off behavior, transmitting τ_{ibt} seconds at peak rate followed by an idle period of $\tau_{ibt}(T_{sbr}/T_{pcr} - 1)$ seconds [13].

We will assume that the parameters describing a source are the peak and sustainable rates (in bits per second) and maximum burst size (in bits) and denote it by the triple $(\hat{R}, \overline{R}, b)$. We will disregard the two rate–tolerance values. No further information about the source can be assumed by the call–acceptance control. The bound can be illustrated by a so–called *arrival curve*: Let $\mathfrak{R}(t)$ denote the number of bits sent by

a source from time 0 to time t, then the arrival curve is given by $a(\tau) = \sup_{t \geq 0} \mathfrak{R}(\tau + t) - \mathfrak{R}(t)$. It is consequently a bound on the number of bits the source can generate in a period of τ seconds. The leaky bucket gives an upper bound to the arrival curve consisting of two lines, as shown in Figure 1.

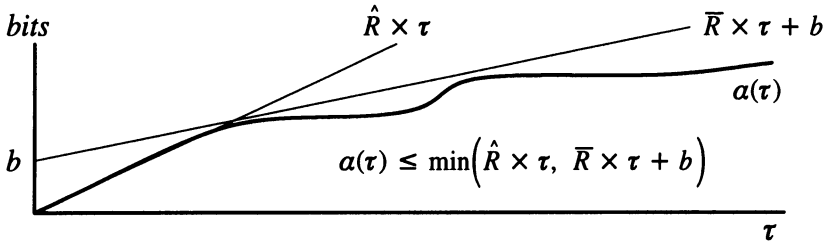

Figure 1 An arrival curve for a source bound by a leaky bucket.

Statistical and deterministic bitrate services

As mentioned before, there are two types of transfer capabilities that we consider in this study: the deterministic bitrate service and the statistical bitrate service. The former requires specification of the peak rate and will subsequently be denoted $(\overline{R}, \overline{R}, 0)$ (*i.e.*, $\hat{R} = \overline{R}$); the latter is specified by the full parameter triple $\left(\hat{R}, \overline{R}, b\right)$. Deterministic bitrate service means that the connection is assigned a capacity that is at least equal to the peak rate. The ITU Recommendation does not state the associated quality of service but loss–free service with low maximum delay is possible, and will henceforth be assumed [9]. The peak rate cannot be renegotiated during the session by any other means than signalling and network management procedures.

The statistical bitrate service means that a rate $\overline{R} < R^* < \hat{R}$ is allocated for the connection. The parameters are fixed for the duration of the call, or renegotiated by signalling or management. The number of algorithms for call–admission control in the literature is large [2]. Yet none, to our knowledge, handle call requests based on the leaky–bucket descriptor when not all calls have the same parameter values. Our study is consequently based on a homogeneous situation with identical and independent calls. The algorithm we have chosen considers only the peak and sustainable rates in the acceptance decision.

For the sake of discussion, we briefly describe the ATM block transfer capability (ABT) although it is not a part of our study. A block consists of a group of cells between two resource management (RM) cells. The first RM cell establishes the block cell rate for the group, which essentially is a peak rate for the block. The second RM cell releases the resource or changes the reserved rate to suit the following block. The service is therefore a DBR service with piecewise fixed rates. The parameters for a connection are: the overall peak cell rate that never may be surpassed by the block cell rate, the peak cell rate for the RM cells which gives the minimum renegotiation interval, and the sustainable cell rate. The sustainable rate can be used to lower the blocking probability for the renegotiations: if the mean rate up to a renegotiation

point is below the SBR, then an increase will be accepted without blocking. The rate may be set to zero.

The ABT is a flexible service that can be well suitable for video [3]. However, it is not clear how the sender would be able to choose appropriate renegotiation intervals and sustainable rate for a live transfer. We will discuss the ABT option further in the final conclusions.

3 THE VIDEO SYSTEM

A video communication system is shown in Figure 2. The digitized video is first passed to a source coder. It is often built with three system components: energy compaction, quantization and entropy coding [10].

The energy compaction aims at putting the signal into the form most amenable to coarse quantization. Common methods for video include discrete cosine transform, subband analysis, and prediction, possibly motion estimated. The quantizer reduces the number of permissible amplitude values of the compacted signal and introduces round–off errors. The entropy coding, finally, assigns a new representation to the signal which represent the data more efficiently but there is no longer a constant number of bits per picture, and the bit rate becomes temporally varying.

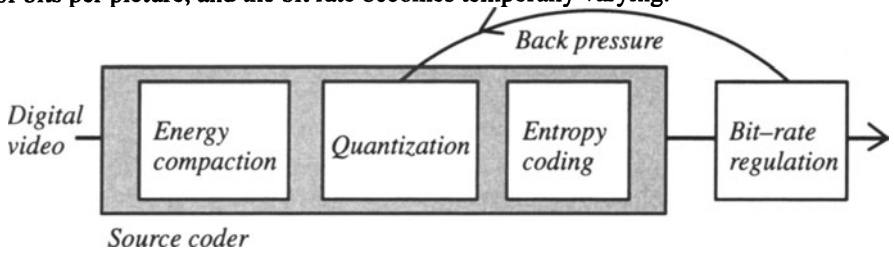

Figure 2 The sending side of a a video communication. The bit rate regulation consists of a smoothing buffer with back pressure to avoid overflow.

The bit–rate regulation is used to adapt the varying bit rate to the channel in the network. The regulation flattens the bit–rate variations by buffering and may regulate the compression to avoid overflow. The feedback reaches the quantization of the encoder and enforces a higher step–size with increased round–off error as a consequence. If the quantizer step–size is throttled frequently and heavily it may lead to visible quality fluctuations in the reconstructed signal.

Leaky bucket descriptors have recently been studied for regulated video. It is clear that the feedback makes it possible to regulate the bit rate from the coder in order to fit any choice of leaky–bucket parameters. Whether a particular set of parameters is good or not can only be determined by subjectively evaluating the encoding quality. Hsu *et al.* have established that a smoothing buffer of size B_{sbr} together with the leaky–bucket descriptor $\left(\hat{R}, \overline{R}, b\right)$ yields the same quality as a system with a buffer of size $B_{dbr} = B_{sbr} + b$ and the single upper bound descriptor $(\overline{R}, \overline{R}, 0)$ (again, this

notation means that $\hat{R} = \overline{R}$) [8]. The gain is therefore a lower delay in the former case since the buffer can be b bits smaller without any effect on quality. It should be noted, however, that the first case also requires a higher capacity allocation due to the allowed burstiness, and it is the issue we are considering in this study.

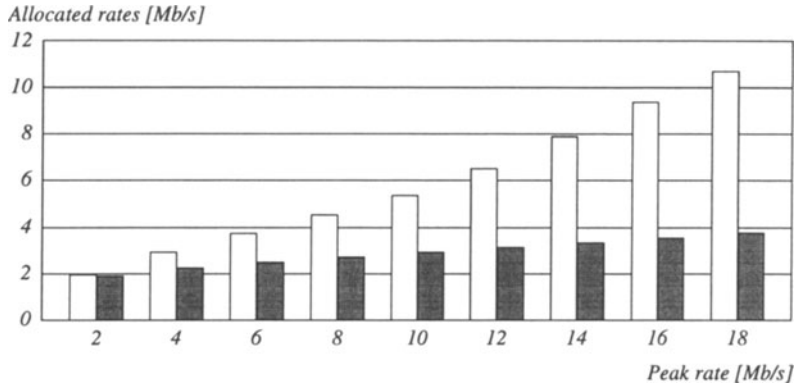

Figure 3 The allocated rate as a function of the source's peak rate. The sustainable rate is 1.8 Mb/s, and the link rate is 150 Mb/s for white bars and 620 Mb/s for grey bars.

4 COMPARISON OF EFFICIENCY

We now study the two cases given above. For the leaky–bucket characterized source, one cannot justify any helpful assumptions about the variations within the bounds and an on–off pattern must consequently be assumed to be safe (even though such behavior is not observed for variable bit rate video) [13]. Following Hamdi *et al.* [5], the allocation for identical connections can be approximated as $R^* = C/N$, where N is the largest value such that the target loss probability is not exceeded:

$$p_{loss} = \frac{\sum_{j=\lceil C/\hat{R} \rceil}^{N} \left(j\hat{R} - C\right)\binom{N}{j}\left(\overline{R}/\hat{R}\right)^{j}\left(1 - \overline{R}/\hat{R}\right)^{N-j}}{N\overline{R}}.$$

This expression assumes a fluid–flow model of the traffic and a bufferless multiplexer (thus, b is not appearing in the expression; it could, however, be considered in a tariff structure if call acceptance would be based on this formula). It follows that the utilization is $\overline{R}/R^* < 1$ when $\hat{R} > \overline{R}$. A unity utilization would mean that all connections are allocated their sustainable rates. This should not be mistaken for the actual usage of the link which could be arbitrarily much lower since the declared sustainable rate might be well above the actual mean rate due to uncertainty in the sender's parameter estimation. The allocations needed for a cell loss probability of 10^{-6} are plotted in Figure 3 as functions of the peak rate for link rates of 150 and 620

Mb/s. We only consider the 150 Mb/s link rate in what follows since it is more realistic for access links, which easily become the bottlenecks for video services.

A valid question is what gain can be achieved by SBR over DBR for a given source. If delays are not of primary importance then obviously DBR is more efficient than SBR in terms of capacity allocation (and quality of service) since it requires an allocation of $\overline{R}$ compared to $R^* > \overline{R}$ for SBR. Recall that the encoding quality of the two cases is comparable if $B_{dbr} = B_{sbr} + b$. The delay difference is at most $b/\overline{R}$ seconds, if we assume that the network delays are as short for SBR as for DBR.

A more interesting comparison is to keep the capacity allocation equal in the two cases and to compare the resultant smoothing delays. Thus, for DBR we have a descriptor $(R^*, R^*, 0)$ and for SBR the usual $\left(\hat{R}, \overline{R}, b\right)$ descriptor which also leads to an allocation of R^*. We would like to determine the buffer size B_{dbr} and the buffer plus burst size, $B_{sbr} + b$, such that $P(Q > B_{dbr} \mid R^*) = P(Q > B_{sbr} + b \mid \overline{R})$. This means that the probability of the queue exceeding B_{dbr} serviced at a rate R^* should be equal the probability of exceeding B_{sbr} with a burst of b bits when serviced at rate $\overline{R}$. Functions like $P(Q > B \mid R)$ have been studied by Chong and Li under the name *probabilistic burstiness curves* [1]. Given the values for B_{dbr} and $B_{sbr} + b$, we can determine the difference in smoothing delay for a given maximum burst size.

Instead of using the more general probabilistic burstiness curves, we restrict our comparison to the equivalent capacity for a two–state Markov chain. The formula by Guérin *et al.* is well–known [4]:

$$R_{eq.} = \frac{a\hat{R}_s - B + \sqrt{(a\hat{R}_s - B)^2 + 4aB\overline{R}_s}}{2a},$$

where

$$a = -(\ln p_{loss})\, \overline{b}_s \left(1 - \overline{R}_s/\hat{R}_s\right).$$

The subscript 's' signifies that the parameters are for the source, and not the connection. In general we may expect that $\hat{R}_s \geq \hat{R}$ and $\overline{R}_s < \overline{R}$. The latter is simply a stability condition for the smoothing buffer. The parameter $\overline{b}_s$ is the average burst duration (in seconds) for the two–state chain.We solve the expression for the two cases $R_{eq.} = \{R^*, \overline{R}\}$ to find the corresponding buffer sizes $B = \{B_{dbr}, B_{sbr} + b\}$ for the given loss probability.

Figure 4 shows the resultant buffer sizes for $\overline{R}_s = 0.7\overline{R}$, $\hat{R}_s = \hat{R}$ and $\overline{b}_s = 40$ ms (the frame duration for European video formats). The actual utilization of the sustainable rate is consequently 70 percent. The quality in terms of loss has been fixed in the calculation (to $p_{loss} = 10^{-6}$ which could be the probability of overflowing the smoothing buffer or of regulating the quantizer). The delay for the SBR and DBR cases would be equal if $B_{sbr}/\overline{R} = B_{dbr}/R^*$. This gives the value for B_{sbr} and from the buffer values in Figure 4 for $B_{sbr} + b$ we find the minimum value of b for which SBR yields as low smoothing delay as DBR. These values are plotted in Figure 5 for in-

creasing peak rates. SBR is more efficient than DBR in terms of smoothing delay when the declared burst size of the leaky bucket is above the line in the plot.

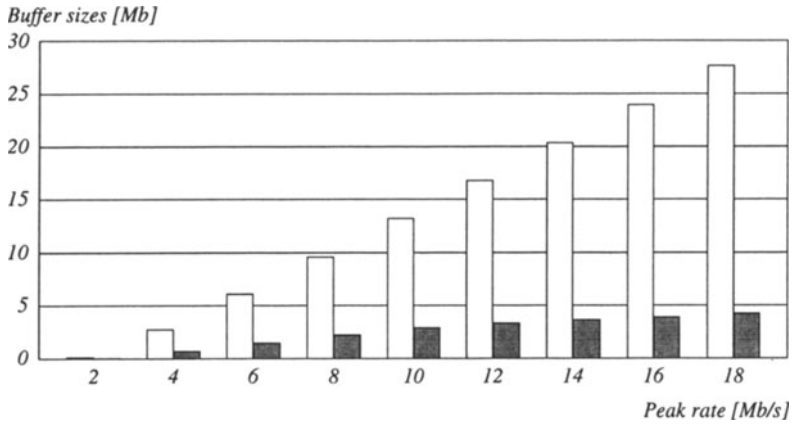

Figure 4 The needed buffer sizes for a given equivalent capacity. White bars are for $B_{sbr} + b$ and grey bars for B_{dbr}.

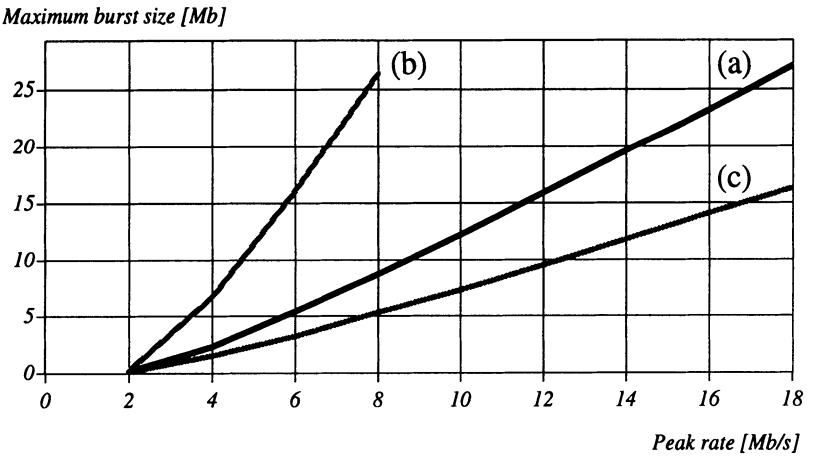

Figure 5 The maximum burst size, b, as a function of peak rate $\hat{R}$. The parameters are $\overline{R} = 1.8$ Mb/s, $\hat{R}_s = \hat{R}$ and $\overline{b}_s = 40$ ms. The average source rates are (a) $\overline{R}_s = 0.7\overline{R}$, (b) $\overline{R}_s = 0.9\overline{R}$ and (c) $\overline{R}_s = 0.5\overline{R}$.

5 CONCLUSIONS AND DISCUSSION

We have given a straightforward example to illustrate that a DBR service may in many cases outperform an SBR service for variable bitrate sources, such as video. A given quality of service can always be assured at a lower allocation of capacity with DBR than SBR if delay is not of prime importance. The allocation for SBR service

would be $R^* > \overline{R}$ for a call described by the leaky–bucket parameters $\left(\hat{R}, \overline{R}, b\right)$, while the DBR allocation would be $\overline{R}$ at the cost of $b/\overline{R}$ seconds of more delay. If the burst size is considered by the call–acceptance control then the comparison could be skewed further in favor of DBR service since R^* would increase with increasing b.

By keeping the allocations equal and instead comparing delays we also show that DBR outperforms SBR for some reasonable parameter choices. We have used an on–off Markov model for the source which fits the on–off behavior that is assumed by the call–admission control. Yet the results show that SBR is not always yielding a lower delay than DBR. For instance, a call with 1.8 Mb/s sustainable rate and 8 Mb/s peak rate (a 5:1 peak–to–mean ratio is common for VBR video [6]) would require a declared maximum burst size of nearly 9 Mb in order to yield lower smoothing delay than a DBR connection with the same allocated capacity. This only assumes an average source rate that is 70 percent of the sustainable rate. The needed burst size depends on both the utilization, as can be seen in Figure 5, and on the average burst duration for the source. For $\overline{b}_s = 10$ and 80 ms, the needed burst sizes are 2.2 and 17 Mb, respectively, compared to 9 Mb for the case above (with 70 percent utilization in all cases).

It is not too surprising to find the low efficiency of SBR service for call–acceptance decisions based on leaky–bucket descriptions of the calls. All studies of statistical multiplexing gains have assumed source characteristics that are known to a very high degree: for instance, they could be captured by a stochastic model with parameters fitted to real data [7], or when only the first two moments are used, they are still actual values for a real source [6]. What we have shown here is that a leaky–bucket description of a source does not provide enough information about source characteristics to ensure a reasonable multiplexing gain in many cases. Our finding is supported by the study presented in [11].

Even such a simple descriptor as the triple $\left(\hat{R}, \overline{R}, b\right)$ causes problems for the sender to determine suitable parameter values before a call has commenced. The quality of a call may, on the one hand, not be as good as expected if the values are too small, since the bitrate regulation will ensure that the agreed parameters will not be exceeded. New parameter values can only be established by user signalling, which typically would have to be initiated manually. On the other hand, the call will be unnecessarily expensive if the parameters are only loosely fitted to the actual traffic. Thus, increasing the complexity of the descriptor is not a good solution: it would allow the network to operate more economically with more statistical multiplexing at the expense of the user who would have more parameters to estimate, allowing more room for mis–estimation. Methods for estimating even the simple leaky–bucket parameters for a call request are to our knowledge still lacking.

There are three possible solutions to this dilemma. The first is measurement–based admission control which enhances a user–provided traffic descriptor (typically only the peak rate) by information from measurements of on–going calls (there are already several proposals in the research literature, one example is [12]). Any quality

guarantee would be conditioned on the assumption that the measurements provide reliable information for predicting future network behavior.

The second possibility is to use the ATM block transfer capability for which the fixed rate can be renegotiated at need to suit variations in a bit stream, for instance as caused by scene variations in a video program. The study by Grossglauser *et al.* could serve as a good starting point for further investigations [3]. The third possibility is to offer a combination of only DBR and ABR/UBR transfer capabilities. Such service offering could still be very efficient since reserved but unused capacity of DBR calls would be available to ABR and UBR calls [9]. The sole advantage of ABT over this simple service offering is that ABT calls may share capacity between themselves through the renegotiations. In the simple case, unused capacity allocated to DBR calls is only shared with ABR/UBR calls. It is not clear how important this advantage is in practice.

Although it is not clear what burst sizes are practically permissible in operational ATM networks, we are troubled by the very large sizes that are needed for SBR to compare favorably with DBR in terms of delay at equal bitrate allocations. This certainly weakens the case for promoting SBR, considering also that it is more complex to implement and that it yields an inferior transfer quality compared to DBR service. We hope that this paper may help directing the attention of traffic–control researchers away from the statistical bitrate service towards evaluating the merits and problems associated with measurement–based admission control, ATM block transfers, and the DBR/UBR service structure.

6 ACKNOWLEDGMENT

This study was done when G. Karlsson was visiting professor at the Telecommunication Software and Multimedia Laboratory at the Helsinki University of Technology, Finland. This support is gratefully acknowledged.

7 REFERENCES

[1] S. Chong and S. Li, "Probabilistic Burstiness–curve–based Connection Control for Real–time Multimedia Services in ATM Networks," IEEE Journal on Selected Areas in Communications, Vol. 15, No. 6, August 1997, pp. 1072–1086.

[2] E. Gelenbe, X. Mang, and R. Önvural, "Bandwidth Allocation and Call Admission Control in High–Speed Networks," IEEE Communications Magazine, Vol. 35, No. 5, May 1997, pp. 122–129.

[3] M. Grossglauser, S. Keshav, and D. Tse, "RCBR: A Simple and Efficient Service for Multiple Time–Scale Traffic," ACM Computer Communications Review, Vol. 25, No.4, October 1995, pp. 219–230.

[4] R. Guérin, H. Ahmadi, and M. Naghshineh "Equivalent Capacity and Its Application to Bandwidth Allocation in High–Speed Networks," IEEE Journal on Selected Areas in Communications, Vol. 9, No. 7, September 1991, pp. 968–981.

[5] M. Hamdi, J. W. Roberts, and P. Rolin, "rate Control for VBR Video Coders in Broad–band Networks," IEEE Journal on Selected Areas in Communications, Vol. 15, No. 6, August 1997, pp. 1040–1051.

[6] H. Heeke, "Statistical Multiplexing Gain for Variable Bit Rate Video Codecs in ATM Networks," International Journal of Digital and Analog Communication System, Vol. 4, 1991, pp. 261–268.

[7] D. P. Heyman, "The GBAR Source Model for VBR Video Conferences," IEEE/ ACM Transactions on Networking, Vol. 5, No. 4, August 1997, pp. 554–560.

[8] C.–Y. Hsu, A. Ortega, and A. R. Reibman, "Joint Selection of Source and Channel Rate for VBR Video Transmission Under ATM Policing Constraints," IEEE Journal on Selected Areas in Communications, Vol. 15, No. 6, August 1997, pp. 1016–1028.

[9] G. Karlsson, "Capacity Reservation in ATM Networks," Computer Communications, Vol. 19, No. 3, March, 1996, pp. 180–193.

[10] G. Karlsson, "Asynchronous Transfer of Video," IEEE Communications Magazine, Vol. 34, No. 8, August 1996, pp. 118–126.

[11] B. V. Patel and C. C. Bisdikian, "End–Station Performance under Leaky Bucket Traffic Shaping," IEEE Network, September/October 1996, pp. 40–47.

[12] H. Saito, "Dynamic Resource Allocation in ATM Networks," IEEE Communication Magazine, Vol. 35, No. 5, May 1997, pp. 146–153.

[13] T. Worster, "Modelling Deterministic Queues: The Leaky Bucket as an Arrival Process," in *Proc. ITC–14*, Elsevier Science, 1994, pp. 581–585.

8 BIOGRAPHIES

Gunnar Karlsson works at SICS since 1992. He holds a Ph.D in electrical engineering from Columbia University and a M.Sc. from Chalmers University of Technology. He has been project leader for the Stockholm Gigabit Network and conducts research on packet video communication, quality of service provisioning and switch architectures. He is a member of IEEE and ACM.

Goran Djuknic received his Diploma and MS degrees from the University of Belgrade, Yugoslavia, and a Ph.D. from the City College, New York, all in electrical engineering. He is with Bell Laboratories, Lucent Technologies, where he evaluates the potential of satellite–based and other innovative schemes for establishing wireless communications services. He also develops new wireless data applications. He is a member of IEEE and on the Board of the Tesla Society.