INTERNATIONAL JOURNAL OF COMMUNICATION SYSTEMS
Int. J. Commun. Syst., **11**, 29–41 (1998)

RESOURCE ASSIGNMENT FOR INTEGRATED SERVICES IN WIRELESS ATM NETWORKS

paramvir bahl[1], imrich chlamtac[2] and andrás faragó[3]

[1]*Microsoft Corporation, Redmond, WA, USA*
[2]*University of Texas at Dallas, Dallas, TX, USA and*
[3]*Technical University of Budapest, Budapest, Hungary*

SUMMARY

The task of supporting integrated multirate multimedia traffic in a bandwidth-poor wireless environment poses a significant challenge for network designers. In this paper we propose a novel bandwidth allocation strategy which partitions the available bandwidth amongst the different traffic classes in a manner that ensures quality-of-service guarantees for digital video while minimizing the maximum blocking probability for voice and data connections. At the connection level, near-optimum utilization of the reserved bandwidth for video traffic is achieved through an intra-frame statistical multiplexing algorithm, while at the system level the delicate task of partitioning the bandwidth between voice, video and data is accomplished by developing an efficient algorithm which uses traffic parameters consisting only of the aggregate traffic load and the total available bandwidth. The algorithm, built on non-trivial mathematical results is robust, easy to implement and has a geometric rate of convergence which ensures that the partitioning points are found quickly. These properties make it well suited for practical implementations, even for cases where changes in the aggregate traffic loads cause bandwidth allocations to be recomputed frequently. © 1998 John Wiley & Sons, Ltd.

key words: wireless multimedia; wireless video; quality-of-service; bandwith allocation

1. INTRODUCTION

Different broadband services require different amounts of bandwidth and have different priorities. For example, a connection for visual communications will in general require more bandwidth than one for data communications, and a voice connection will in general be of higher priority than either a data or a video connection. In response to these varied demands the network designer may choose to assign different amounts of bandwidth to different types of traffic. The motivation for such an approach stems from the desire to support a variety of multimedia services with a reasonable level of performance and without letting the demand from any one type shut out other types of services. Thus the challenge for the network designer is to come up with techniques that are able to balance the needs of the various applications with the need of the system to accommodate as many heterogeneous connections as possible. This task of providing guaranteed quality of service (QoS) with high bandwidth utilization while servicing the largest possible number of connections can be achieved through a combination of intelligent admission control, bandwidth reservation and statistical multiplexing.

Providing simultaneous support for real-time variable-bit-rate (VBR) video, real-time voice and data traffic over a bandwidth-constrained noisy channel continues to be a formidable problem. The difficulty arises primarily because VBR video is unpredictably bursty and because real-time visual services require timely delivery and performance guarantees from the network. While resource reservation schemes work best for constant-bit-rate (CBR) traffic, there is no consensus on which strategy should be used for VBR traffic. On one hand, since real-time VBR traffic is delay-sensitive, a resource reservation scheme seems to be the right choice; on the other hand, because VBR video is bursty if resources are reserved according to peak rates, the network may be under-utilized whenever the peak-to-average rate ratios are high. These two opposing characteristics have resulted in a common belief that it is unlikely that performance guarantees can be provided to such bursty sources with very high network utilization. Although the problem of minimizing wastage (or maximizing utilization) while supporting the largest number of heterogeneous connections is not limited to wireless networks, the need to solve this problem in such networks is much greater than in wireline networks owing to the characteristics of the underlying transmission channel.

The challenge can thus be stated as follows. *Can performance guarantees be provided to VBR video without significantly underutilizing the bandwidth* **and** *can this be done in conjunction with minimizing*

Correspondence to: P. Bahl, Microsoft Corporation, Washington, DC, USA. Email: bahl@microsoft.com

CCC 1074–5351/98/010029–00

the maximum blocking probability for voice and data connections?

In References 1 and 2 we tackled the first part of this question. The approach we advocated in those papers was to reserve channel bandwidth for the lifetime of the video connections. In order to minimize wastage of reserved resources, we suggested the use of either an adaptive joint source–channel multiscale, spatially segmented video codec[2] or a standards-based video codec enhanced to include a content-sensitive region segmentation preprocessor.[1] With such a codec in place we proposed that the application reserves the peak bandwidth for the primary subband (or region) while letting the secondary and tertiary subbands (regions) compete for bandwidth dynamically. We circumvented the problem of underutilizing the reserved bandwidth by stuffing or statistically multiplexing the remaining (secondary, tertiary, etc.) subbands (regions) into the non-used portion of the bandwidth after the primary subband (region) had been dispatched. With this scheme, called *intra-frame statistical multiplexing*, all the reserved bandwidth is used up efficiently and wastage is avoided.[1]

The focus of this paper is on tackling the second part of the challenge outlined above; that is, to develop an algorithm that partitions the system bandwidth according to some useful and well-defined optimality criterion. We will show that by using an appropriate bandwidth partitioning scheme, we can ensure that the maximum blocking probability for the various traffic classes can be minimized. In contrast with the solutions proposed in References 1 and 2, which are implemented in the host system, the solution offered in this paper is implemented at the system level.

We develop a simple yet effective algorithm that partitions the available bandwidth in a manner that minimizes the maximum blocking probability for voice and data connections while providing guaranteed QoS to VBR video connections. It should be noted that even when the distribution of the different traffic types is available finding the optimal partitioning of the bandwidth is a very difficult task, and for the general case can be modelled by an NP-complete graph colouring problem. The intractability of finding the optimum is present already in the simplest situation when the traffic consists of voice and data connections only and the statistics of the offered traffic are completely known. However, the problem becomes even more difficult when the wireless network carries integrated non-homogeneous traffic, an integral feature of wireless ATM networks. In this case, estimating the blocking probability of connections and its application in resource allocation strategies is further complicated because of two fundamental reasons.

- Although there are methods for computing blocking probabilities for integrated systems under specific statistical assumptions[3] (e.g. multirate Poisson models), there are no simple closed-form formulae that can easily be applied to optimizing resource allocation.
- It is realistic to expect that traditional statistical asumptions will not describe the traffic load precisely. Therefore it is injudicious to make concrete assumptions based on any advance knowledge regarding the detailed statistical properties of traffic in a wireless multimedia network. This calls for a bandwidth management methodology that works under incomplete information and does not critically depend on specific statistical assumptions.

In the subsequent sections we propose a solution for the allocation of transmission resources among different traffic classes under incompletely known conditions. Our solution has the following main properties.

1. It provides guaranteed QoS for on-going real-time visual communication sessions. This guarantee does not come at the expense of wasted bandwidth, since all the reserved bandwidth is used up through intelligent connection-level statistical multiplexing.
2. It is robust and insensitive to statistical assumptions as it depends only on the average rates of the aggregated flow of traffic types, but not on detailed statistics of the traffic mix and of the arrival process. From a practical viewpoint this insensitivity is highly advantageous, since the detailed statistical information is typically unavailable or uncertain.
3. The resulting allocation is based on minimizing a bound on the blocking probabilities that is proven to be asymptotically optimal. The optimality is also important as it signifies that for large systems it is sufficient to know aggregated flow rates since the detailed knowledge of the traffic mix would not significantly contribute to achieving smaller loss.

2. BANDWIDTH PARTITIONING

While the simultaneous support of integrated heterogeneous traffic is important, it is also necessary that the network exercise traffic access control so as not to allow any one type of traffic to unfairly dominate and possibly shut out other classes of traffic. Towards this end, researchers have studied and proposed bandwidth partitioning strategies that 'fairly' allocate bandwidth to different traffic classes while maximizing the overall network throughput.[4,5] Previous studies of these techniques in wireless networks have focused on the coexistence of data and voice traffic, while packet video has generally been ignored. We make the logical extension to include packet video and present some pros and cons of these strategies.

2.1. *Complete Sharing, Complete Partitioning and Priority Sharing*

At the two extremes are the Complete Sharing (CS) and Complete Partitioning (CP) (also called Mutually Restricted Access (MRA)) strategies, and in between are the rest generally referred to as hybrid strategies. Figure 1 illustrates the three general bandwidth partitioning strategies.

As the names suggest, in CS all traffic classes share the entire bandwidth. Although trivial to enforce, the main drawback of this strategy is that a temporary overload of one traffic class results in degrading the connection quality of all other types. In CP, bandwidth is divided into distinct portions, with each portion corresponding to a particular traffic class. CP is wasteful of bandwidth if the predicted bandwidth demand for a particular traffic type is greater than the actual bandwidth demand. A compromise between CP and CS is a strategy in which bandwidth is allocated dynamically to match the varying traffic load. Put another way, an attempt is made to achieve statistical multiplexing at the burst level rather than at the connection level. One such technique is called Priority Sharing (PS). A moving boundary exists between the bandwidth allocated for the various traffic types, and priority users (voice traffic) are allowed to borrow bandwidth from non-priority users (data traffic). It has been shown that this hybrid scheme provides better performance than both CS and CP, over a range of offered loads in both microcellular and macrocellular environments.[6] Table I lists the main differences between the three strategies outlined above. The reader is referred to Reference 4 for a more comprehensive survey of such schemes.

Table I. Comparison of bandwidth partioning policies

Complete Sharing (CS)	Complete Partitioning (CP)	Priority Sharing (PS)
No protection from overloading of other classes	Complete protection from over loading of other classes	Protection from overloading of other classes can be built in
Blocking probability for each traffic class is not adjustable	Blocking probability for each traffic class is easily adjustable	Blocking probabilities are tunable, depending on flavour of PS scheme
Bandwidth-efficient	Bandwidth-inefficient	Bandwidth-efficient

2.2. *Priority Sharing with Restrictions*

Good bandwidth allocation schemes rely on dynamic allocation of bandwidth to achieve high utilization. While dynamic allocation at the burst level provides good statistical multiplexing, it performs poorly for connections that require a certain quality of service. It would not be too extreme to claim that the only practical way to guarantee quality of service is by providing bandwidth reservation for the entire lifetime of the connection. In previous studies, bandwidth allocation for VBR video at connection establishment time was not seen as an interesting and viable alternative, since no real technique had been developed that would prevent wastage of reserved bandwidth. Since it is very difficult to accurately predict at connection establishment time the bandwidth requirement of a VBR video connection, static reservation of bandwidth was generally ignored. However, using the technique from References 1 and 2, static reservation can be provided without wasting precious bandwidth (Figure 2). With this technique we can build into a medium access protocol provisions for both static (lifetime) and dynamic bandwidth reservation.[7] With static bandwidth reservation we are able to guarantee a quality of service and with dynamic reservations we are able to improve the visual quality of the images when the bandwidth is available.

A natural question to ask is: which is the best bandwidth allocation scheme (CS, CP or PS)? From the point of view of providing guaranteed quality of service for visual communications, Complete Sharing is not suitable. Complete Partitioning, on the other hand, can deliver but, as noted earlier, is

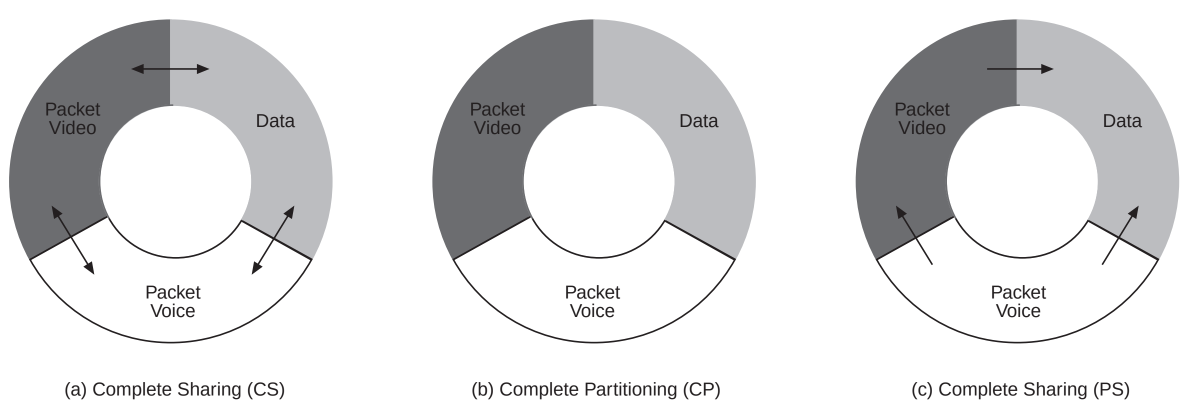

Figure 1. Flavours of bandwidth partitioning schemes

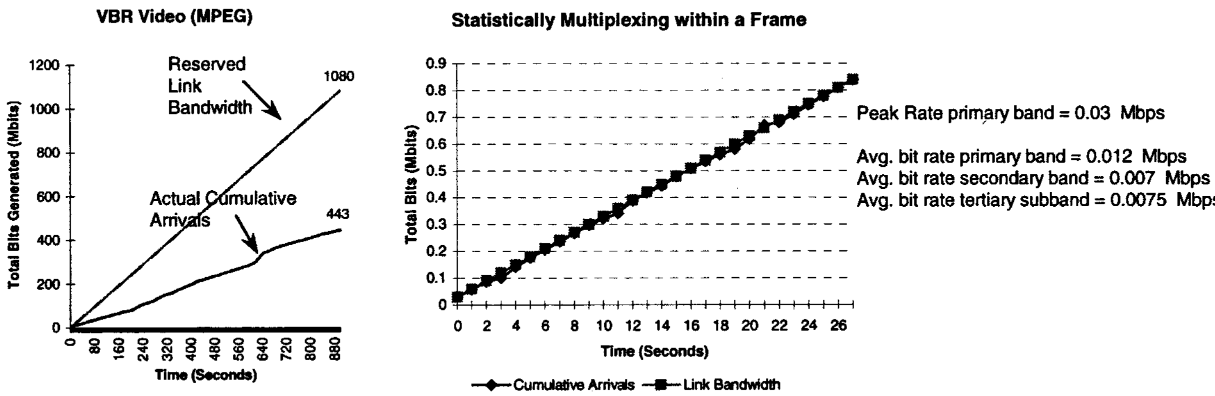

Figure 2. Bandwidth utilization for VBR video with and without intra-frame statistical multiplexing

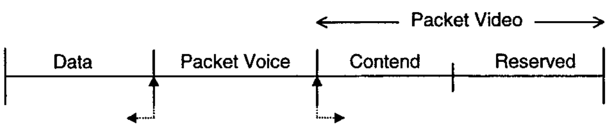

Figure 3. Priority Sharing with Restrictions

wasteful of bandwidth. Priority Sharing, which is essentially a compromise between CS and CP, is thus the most viable candidate. We have extended Priority Sharing to include static bandwidth reservation with a moving boundary. We call this new scheme *Priority Sharing with Restrictions* (PSR). Figure 3 illustrates the PSR scheme.

In this scheme, only real-time video connections are allowed to make lifetime (or static) reservations; voice and data connections are allowed dynamic reservations only. At connection establishment time, bandwidth for video connections is allocated (for the main subband or region) from the Reserved portion of the available spectrum. The amount of bandwidth allocated for such static reservations is determined by the network designers (see Section 3). The remaining spectrum is divided between voice, data and video (for secondary, tertiary subbands/regions) users and is used for dynamic burst-level reservations. In terms of priority, voice users have the highest priority, followed by video and data users in that order. Table II provides the rule-table of who can borrow from whom. For example, data users may borrow from both voice and video, but they will be pre-empted if voice or video users need the bandwidth (this is similar to what happens in the currently popular Cellular Digital Packet Data (CDPD) protocol[8]). Similarly, new video connections cannot reserve bandwidth allocated for data or voice users.

Table II. Rules for Priority Sharing with Restrictions

	Data	Voice	Video-dynamic	Video-static
Data	–	BP	BP	BP
Voice	B	–	B	BP
Video-dynamic	B	BP	–	BP
Video-static	X	X	X	–

B, borrowing allowed; BP, borrowing allowed, pre-emption possible; X, borrowing not allowed; –, don't care.

Figure 4 shows a sample of the performance data for our Priority Sharing with Restrictions scheme as compared with the basic Priority Sharing and Complete Sharing schemes. Details of the experiment are provided next to the graph. The video codec we used for the experiment was a region-segmented H.263 video codec.[1] A spatial segmenter preceded the H.263 codec, dividing the image into five distinct regions before compression. Video traffic load was increased by using the same compressed video stream multiple times and the average PSNR was computed by calculating the average PSNR of each bitstream at the output of the decoder and then taking the average of all these averages. The channel access protocol used in the simulation was a reservation-random time division multiplexing-based access control protocol called ARMAP (Adaptive Reservation Multiple Access Protocol).[7] ARMAP allows terminals to communicate with multiple traffic types, including data, voice and digital video. Voice and data are supported in a manner similar to DRMA,[9] while visual communications are supported through a combination of static and dynamic reservations. The ARMAP protocol monitors and then exploits the regularity in the video packet generation process to provide contention-free channel access to on-going real-time video connections at appropriate times. A novel adaptive reservation-slot generation algorithm ensures optimal bandwidth usage and optimal power consumption in the wireless device.

Looking at Figure 4, we can see that as the traffic load for video increases, the PSR scheme outperforms both the CP and PS schemes. As expected, the PS scheme outperforms the CP scheme. The difference between the three schemes becomes more apparent as the voice (and/or data) traffic load is increased while keeping the channel capacity constant. This is shown in Figure 5.

The general lowering of PSNR values for all three schemes can be explained on the basis of

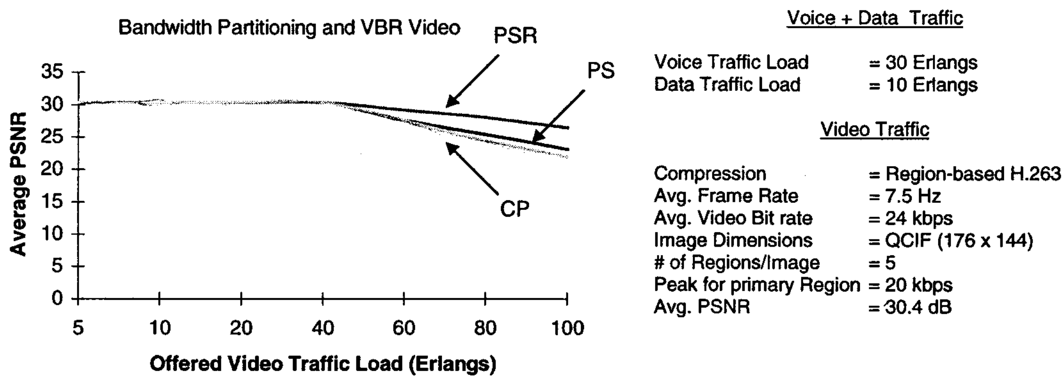

Figure 4. Comparison of bandwidth partitioning schemes

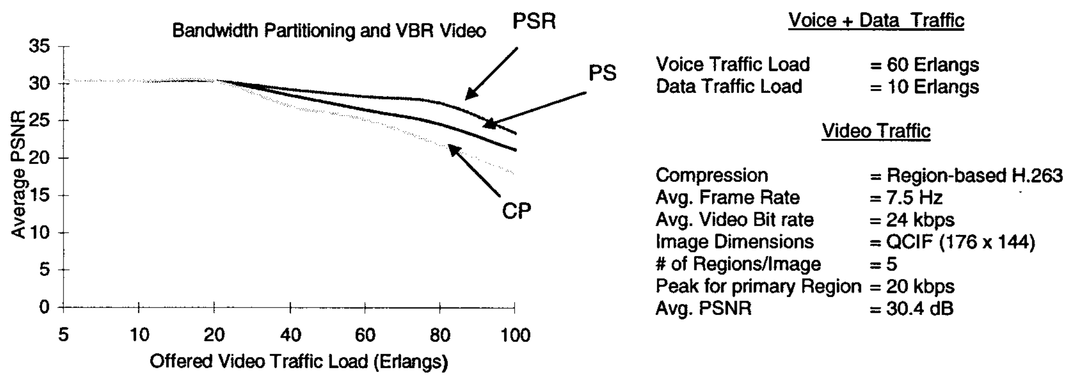

Figure 5. PSR performance with increased voice traffic

video packets (regions) not getting through to the receiver in time. In order to compensate for partially received regions, the video decoder uses previously received (older) regions from its region store[1]—maintaining the frame rate, but with lower PSNR values. As the demand from high-priority voice users is increased, more video packets are blocked and fewer regions (or subbands) reach their destination in time. The result is the lowering of PSNR values for all on-going video connections, but more so for the CS and PR schemes than for the PSR scheme.

The explanation for the generally higher PSNR values in the Priority Sharing with Restrictions scheme compared with the Complete Sharing and Priority Sharing schemes has to do with the availability of bandwidth assigned specifically for the main region (subband) of the video. The reserved bandwidth in PSR ensures that the negotiated QoS is continuously provided to on-going video connection. With at least a minimum number of video packets belonging to the main region getting through to the receiver, the computed PSNR values and the perceived quality for the PSR scheme tend to be higher. The argument that pre-reservation of bandwidth for video connections can also be included in CS and PS strategies is invalid, since reservation of bandwidth for video connections is not always possible in these strategies—if voice connections appear before video connections, they will take up most of the available bandwidth, leaving very little for the video connections (this effect is illustrated in Figure 5). In our simulations we computed the PSNR values for the three strategies with the same number of video connections.

3. BANDWIDTH DISTRIBUTION

So far we have developed a connection-level bandwidth reservation and a system-level bandwidth partitioning scheme which work well with multiresolution VBR video codecs. Together they ensure optimum resource utilization with guaranteed quality of service for video connections while simultaneously supporting voice and data connections. In this section we develop an algorithm that determines the amount of bandwidth to allocate for each traffic class. The objective of this algorithm is to partition the bandwidth such that maximum blocking probability for all connections is minimized.

3.1. *Optimality Criterion*

Let us consider a wireless multimedia network supporting N traffic classes (denoted by $T_1, \ldots, T_N$) and with a total available bandwidth B. Let us assume that users from an arbitrary finite population independently generate connection requests that may require different amounts of bandwidth. Also let us assume that the average aggregate load (or *traffic* *demand*) D_i for traffic type T_i is known ahead of time. Now, given this traffic demand, we wish to find a nominal allocation of the bandwidth B to the different traffic classes; that is, the values $B_1, \ldots, B_N$ such that T_i receives B_i under the constraint $\sum_{i=1}^{N} B_i = B$. (Note. Transmission bandwidth is assumed to consist of a number of basic bandwidth units (BBUs). If the access protocol is TDM-based, then the bandwidth resource B_i translates to B_i/S BBUs, where S is the number of bits in 1 BBU.)

Let $d_i(t)$ be a random variable describing the actual instantaneous aggregated demand by traffic type T_i. Assuming stationarity, the average demand D_i is the expected value of $d_i(t)$ independent of t, i.e. $D_i = E\{d_i(t)\}$. We measure the Grade of Service (GoS) for traffic type T_i by the saturation probability $\Theta_i = P(d_i(t) \geq B_i)$. This is the probability of the event that the instantaneous load for traffic type T_i exceeds the bandwidth B_i allocated for this traffic type. The system GoS is measured by the worst, i.e. the maximum, of these saturation probabilities:*

$$\Theta = \max_i \Theta_i = \max_i P(d_i(t) \geq B_i) \tag{1}$$

Thus our optimization task can be stated as follows. *Given the aggregate load $D_i = E\{d_i(t)\}$ for each traffic class T_i and the total system bandwidth B, determine the allocated transmission capacity B_i for each traffic class such that*

$$\Theta = \max_i P(d_i(t) \geq B_i) \text{ is minimized} \tag{2}$$

subject to the constraint $\sum_{i=1}^{N} B_i = B$

3.2. *Minimizing the Maximum Blocking Probability*

At first glance it appears impossible even to reasonably approximate the optimum in the above task, since the only quantities available for computing or at least estimating the probabilities $P(d_i(t) \geq B_i)$ for any given T_i are the D_i values. It is therefore valid to ask how one can tightly estimate the saturation probabilities from knowing merely the expected value of the randomly fluctuating traffic load from each traffic type, as the generally used estimations of such tail probabillities, known from the theory of large deviations, typically require much more information, e.g. the knowledge of the moment generating function. In what follows, we show that despite the presence of the unknown saturation probabilities in the problem it is possible to find a good practical solution which is based on transforming the problem into a well-defined optimization task, based on an asymptotically optimal estimation.

The key mathematical tool in this solution is a robust and tight estimation of the saturation probabilities, which is based on the following theorem.

Theorem 1

Let $X_1,\ldots,X_N$ be independent random variables taking their values from the interval [0,1]. Their probability distributions are otherwise arbitrary and not necessarily identical. Set $X = \sum_i X_i$ and $D = E(X)$. Then for any $C \geq D$ the following estimation holds:

$$P(X \geq C) \leq \left(\frac{D}{C}\right)^C e^{C-D} \tag{3}$$

Furthermore, this estimation is the best possible in the following sense. For any fixed $\epsilon > 0$ and for any fixed D and C with $C \geq D$ there exist infinitely many counter-examples for which

$$P(X \geq C) > \left(\frac{D}{C}\right)^C e^{C-D-\epsilon} \tag{4}$$

holds.

Remark. The optimality of the estimation, as stated in the second part of Theorem 1, means asymptotic tightness with respect to the exponent, as usual in bounding large deviations.

The proof of Theorem 1 is based on a bound due to Hoeffding,[10] which is a powerful generalization of the well-known Chernoff bound on the tail of the binomial distribution. For readability purposes we omit the proof of Theorem 1 here and present it in Appendix A.

Using Theorem 1, we can bound the saturation probabilities $P(d_i(t) \geq B_i)$ as follows. At any given time t let X_j be a random variable that takes the value of the bandwidth b required by user j. With b values being normalized to the interval [0,1], $X_j \in [0,1]$ holds. Then, according to our model, with $X = d_i(t)$, $D = D_i$ and $C = B_i$ we obtain an asymptotically optimal estimation

$$P(d_i(t) \geq B_i) \leq \left(\frac{D_i}{B_i}\right)^{B_i} e^{B_i - D_i} \tag{5}$$

The estimation (5) makes it possible to transform our original problem into a well-defined optimization task in which the unknown exact saturation probabilities are replaced by the optimal bound (5). *Given the aggregate load D_i for each traffic type T_i and the total system transmission bandwidth B, determine the allocated transmission capacity B_i for each traffic type such that*

* For example, $\Theta = 0{\cdot}01$ represents at most a 1 per cent probability that a given request will be blocked owing to unavailability of sufficient bandwidth.

$$\tilde{\Theta} = \max_i \left[\left(\frac{D_i}{B_i}\right)^{B_i} e^{B_i - D_i} \right] \text{ is minimum} \qquad (6)$$

$$\text{subject to } \sum_{i=1}^{N} B_i = B$$

The asymptotic tightness of the estimation (5) guarantees that in the asymptotic sense, i.e. for large user populations, the solution for (6) will be a very good solution to the original problem as well. The basis for solving (6) is provided by the following property.

Theorem 2
An allocation $B_i, \ldots, B_N$ with $\sum_{i=1}^{N} B_i = B$

is an otimal solution for (6) if and only if

$$\left(\frac{D_i}{B_i}\right)^{B_i} e^{B_1 - D_1} = \ldots = \left(\frac{D_N}{B_N}\right)^{B_N} e^{B_N - D_N} \qquad (7)$$

holds.

For the proof of theorem 2 see Appendix B.

In view of Theorem 2, all that remains to be done for solving (6) is to find an allocation $B_1, \ldots, B_N$ with $\sum_{i=1}^{N} B_i = B$ such that it makes the GoS bounds (5) equal. To derive an algorithm for this, we need an auxiliary function defined as follows. For any fixed $D_i > 0$ and $0 < \sigma \leq 1$ let $B_i(\sigma)$ be the unique solution for B_i of the equation

$$\left(\frac{D_i}{B_i}\right)^{B_i} e^{B_i - D_i} = \sigma \qquad (8)$$

The unique solvability of equation (8) follows from the facts that for $B_i = D_i$ the left-hand side is unity and otherwise it is a strictly decreasing continuous function of B_i that tends to zero as B_i grows (for proof see Appendix C).

3.3. *The Smart Allocate Algorithm*

It follows that the value of $B_i(\sigma)$, i.e. the solution of equation (8), can be computed with arbitrary accuracy by a simple iterative search (interval halving) that approaches the root at an exponentially decreasing error. Using the auxiliary functions $B_i(\sigma)$, we solve (8) such that we successively iterate a value $0 < \sigma \leq 1$ for which $\sum_i B_i = B$ holds within a given error bound $\epsilon > 0$. This algorithm, which we call *Smart Allocate*, is shown in Figure 6.

Looking at Figure 6, if $\sigma = 1$, then $B_i = D_i$ is the solution, otherwise start with a lower and an upper bound for the root, B'_i and B''_i respectively. Then, by iteratively halving the interval between the lower and the upper bound, we can approach the root at a geometric rate, i.e. with exponentially decreasing error. The initial values can be taken as $B'_i = D_i$ and a sufficiently large B''_i for which the left-hand side of equation (8) is already smaller than σ. This value can be found quickly by repeatedly doubling the initial value D_i until the left-hand side is made smaller than σ.

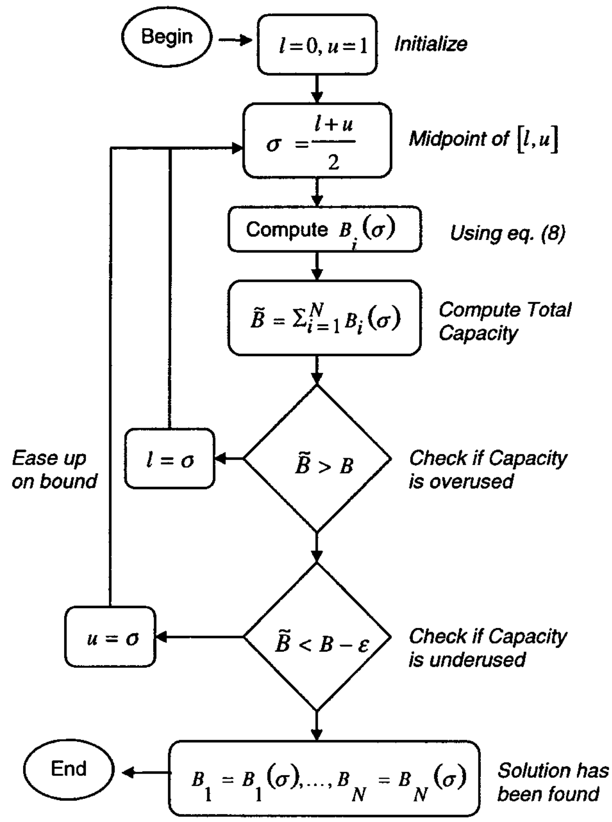

Figure 6. Algorithm *Smart Allocate*

The correctness of the algorithm and the rate of convergence are stated in the following theorem (for proof see Appendix D).

Theorem 3

Algorithm *Smart Allocate* converges to an optimal solution of equation (8) at geometric rate, i.e. the error decreases exponentially.

It is interesting to mention that in our situation, when the transmission capacities have to be allocated knowing only the aggregated average load in each cell, one could easily argue on a commonsense basis that without any other information the only reasonable solution is to allocate the capacities proportionally to the load values. The load-proportional allocation would mean that $D_1/B_1 = \ldots = D_N/B_N$ holds. On the other hand, we know from Theorem 2 that the optimal solution of equation (6), which is the asymptotically optimal GoS estimation, is obtained if and only if

$$\left(\frac{D_1}{B_1}\right)^{B_i} e^{B_1 - D_1} = \ldots = \left(\frac{D_N}{B_N}\right)^{B_N} e^{B_N - D_N} \qquad (9)$$

holds, which is different in general.

3.3.1. *Calculating Aggregate Load* (D_{video}) *for Digital Video.* We have seen that the *Priority Sharing with Restrictions* scheme requires that the system allocate some bandwidth for static reservations of VBR video connections. We have developed the *Smart Allocate* algorithm for determining this amount, based on aggregate load D_i. Thus the question is: how do we determine D_i for digital video?

If we let $p(i,j,b)$ be the probability that user j with traffic class T_i demands bandwidth b at any given time and let $h(i,j,b)$ be the expected holding time of such a connection, then D_i can be expressed as

$$D_i = \sum_{j,b} p(i,j,b)h(i,j,b)b \tag{10}$$

In (10) it is assumed that there are finitely many possible b values and they are normalized such that $0 \leq b \leq 1$ always holds, otherwise they are arbitrary. Furthermore, stationarity is also assumed, so $p(i,j,b)$ and $h(i,j,b)$ are independent of time.

The problem in using (10) for computing D_i is that $p(i,j,b)$, $h(i,j,b)$ and b values and also the actual number of connections are *not assumed known*—fortunately, we do not really need these to determine D_i.

For CBR video connections, D_i is simply a multiple of the constant bit rate. For VBR video connections, demand can be estimated as the multiple of the average peak value of the primary subband (or region) in image frames from several video sequences. Thus D_i is determined by looking at the density functions for the primary subbands (regions) for several similar video sequences, using these to derive the distribution for the maximum values and then computing the mean of this distribution.

Let X_j, $j = 1,2,\ldots,n$, denote the size of the primary subband in the jth frame of n frames occurring in the video sequence. We are then interested in the probability distribution of Y_n in terms of the random variables X_j when $n \to \infty$. The random variable Y_n is defined as

$$Y_n = \max(X_1, X_2, \ldots, X_n) \tag{11}$$

For simplicity, and without losing generality, we assume that X_j are independent and identically distributed. The PDF and density function for Y_n are then given as

$$F_{Y_n}(y) = [F_x(y)]^n \tag{12}$$

$$f_{Y_n}(y) = n[F_X(y)]^{n-1} f_X(y)$$

While the distribution function $F_{Y_n}(y)$ becomes increasingly insensitive to the exact distribution of X_j as $n \to \infty$, no unique results can be obtained that are completely independent of the form of $F_x(x)$. From Reference 11 we know that the normal, gamma and lognormal distributions describe the size distributions best. Observing that all three distributions have right tails that are unbounded and are of the exponential type, i.e. for each case $F_x(x)$ approaches unity at least as fast as an exponential distribution, the cumulative distribution function can generally be described as

$$F_x(x) = 1 - e^{-g(x)} \tag{13}$$

where $g(x)$ is an increasing function of x.

Let $\lim_{n \to \infty} Y_n = Y$. Then, from Appendix E,

$$F_Y(y) = \exp(-e^{-\alpha(y-u)}), \quad -\infty < y < \infty \tag{14}$$

where u and α (> 0) are the location and the scale parameter of the distribution respectively. Here, u is obtained from u_n as $n \to \infty$ and is the value of X_j at which $P(X_j \leq u_n) = 1 - 1/n$. As n becomes large, $F_x(u_n)$ approaches unity or u_n is in the extreme right rail of the video frame size distribution. The scale parameter α is a limiting case of α_n and can be obtained as $\alpha_n = \mathrm{d}g(y)/\mathrm{d}y$ evaluated at $y = u_n$.

The mean of Y is given as

$$\eta_y = u + \frac{0{\cdot}577}{\alpha}$$

where 0·577 is Euler's constant, and

$$D_{\text{video}} = M \times \eta_y \tag{15}$$

where M is the estimated number of video connections to be supported.

3.3.2. *Determining Aggregate Load* (D_{voice}, D_{data}) *for Voice and Data.*

The aggregate load for voice is calculated easily, since the voice connection is assumed to be a CBR connection. Table III lists the output bit rates for some video codecs applied in popular cellular and cordless standards.

Thus, if R represents the constant bit rate at

Table III. Speech coders in popular cellular/cordless systems and their target bit rates

Cellular/cordless system	Coding method	Output bit rate (kbit/s)
GSM, DCS 1800	RPE-LTP	13
IS-95	CELP	1·2–9·6
USDC (IS-54)	VSELP	8
DECT, PACS, PHS, CT2	ADPCM	32

the output of the receiver, D_i is simply calculated by letting

$$D_{\text{voice}} = N \times R \tag{16}$$

where N is the target number of connections to be supported by the network. The remaining bandwidth is used for the data connection:

$$D_{\text{data}} = B - D_{\text{voice}} - D_{\text{video}} \tag{17}$$

Table IV. Blocking probabilities for load-proportional approach

Total channels/voice channels/video channels	Theoretical (T) Voice traffic	Theoretical (T) Video traffic	Simulation (S) Voice traffic	Simulation (S) Video traffic
80/20/60	~ 0·25	0·09	0·513	0·294
100/25/75	~ 0·05	~ 0·01	0·019	0·103
120/30/90	0·0085	< 0·001	0·001	0·042
140/35/105	0·001	<< 0·001	< 0·001	0·001

3.4. *Numerical Examples and Simulation Results*

To demonstrate the effectiveness of the *Smart Allocate* algorithm, let us consider a simple example in which the wireless multimedia network carries two types of traffic—voice (T_1) and video (T_2). Let us consider a TDMA system (similar to GSM, IS-136, PACS, etc.) and quantify the bandwidth by the number of basic bandwidth units (BBUs). Let us assume that we have altogether 99 BBUs which we want to distribute among the two traffic classes such that the maximum blocking probability for the connections is minimized. Furthermore, let the arrival process for connections be Poisson and the blocking probability for the two traffic types be computed by Erlang's classical B formula with the average demands being D_1 = 20 Erlangs for voice and D_2 = 40 Erlangs for video. In what follows, we show that our solution gives better results than the intuitive load-proportional allocation approach.

The load-proportional allocation would assign B_1 = 33 and B_2 = 66 BBUs to the respective traffic types. Then the largest blocking probability, computed from Erlang's formula, is 1 %. In contrast, using our *Smart Allocate* algorithm for this simple example, we obtain B_1 = 37 and B_2 = 62 BBUs. Then the upper bound on the largest blocking probability gives the value 0·57 %—an improvement of 43 % over the load-proportional approach.

Even though intuitively the load-proportional approach looks attractive, there are some problems that emerge when this approach is used for partitioning bandwidth for VBR video. To prove this point, we compared the theoretical blocking probabilities with ones obtained through simulation for the case where BBUs are assigned to each traffic class according to the load-proportional approach. For simulation we used a modified TDMA channel access scheme[7] where each time slot or BBU corresponded to a separate channel. Voice traffic was modelled at an on–off process with a CBR of 20 Erlangs and a voice activity factor of 0·36. Trace-driven simulation was carried out for video packet arrivals.

Table IV shows the blocking probabilities for the two traffic classes (VBR video and CBR voice) included in our simulation. Examination of the data in this table reveals that there is a notable discrepancy between the blocking probabilites calculated from the Erlang B formula and those calculated via simulation.

Several simulation runs indicated that this discrepancy is a function of the burstiness of the VBR video sequence. More bursty video leads to greater discrepancy. The problem may be explained on the basis that the load-proportional approach implicitly assumes that the aggregate value (D_i) of a traffic class is a good measure of the actual load offered to the system by that class most of the time. In reality, for most video sequences, especially those with a high degree of motion, the aggregate value is rarely reached. Hence the discrepancy between the theoretical and simulated blocking probabilities for video results. Figure 7 illustrates these discrepancies graphically.

For the case of voice traffic the discrepancy can be explained on the following basis. Even though we employ a CBR voice codec in our simulation, owing to the presence of silence detection circuitry and the medium access control protocol, a voice connection is forced to relinquish control of the channel assigned to it when it is not transmitting. The unused channels (time slots) are then grabbed by some of the on-going video connections. When the voice connection returns to its transmitting state, the video connection is forced to give up the channel, but not before finishing transmitting its current burst. Therefore, by the time the voice connection gets back its channel, it has experienced some blocking—giving rise to the discrepancy in blocking probabilities. Thus, in a practical situation, in addition to the number of BBUs assigned to voice traffic, the blocking probability for voice connection is affected by the burstiness of on-going video connections in the system as well.

4. CONCLUSIONS

Given a total bandwidth how should it be partitioned to accommodate different traffic classes such that utilization is maximized while connection blocking probability is minimized? This is the problem studied in this paper.

As a solution we presented a novel bandwidth partitioning strategy, called *Priority Sharing with Restriction*, that can support multirate multimedia traffic. We showed that through a combination of

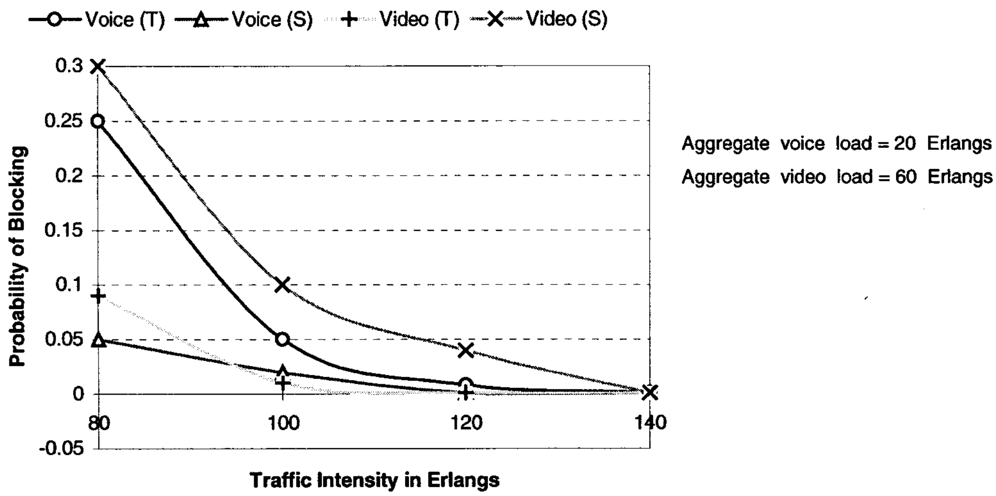

Figure 7. Performance of load-proportional approach with VBR video

connection-level improvements (low-bit-rate multi-resolution, spatially segmented VBR video, combined with appropriate bandwidth reservation and intra-frame statistical multiplexing strategy) and system-level improvements (appropriate partitioning of the available system bandwidth) a wireless network can be configured to simultaneously support digital video with QoS guarantees, along with voice and data communications. Using well-established analytical techniques, we developed the *Smart Allocate* algorithm, which allocates transmission capacities to different traffic classes without requiring detailed prior statistical knowledge of the traffic. Fortunately, the *Smart Allocate* algorithm is simple and easy to implement, and the geometric rate of convergence ensures that the allocations are found quickly. These properties thus make it well suited for practical applications, even for the case where the aggregate load values change frequently and the bandwidth allocation has to be recomputed. As a final step we showed that the *Smart Allocate* algorithm outperforms the popular load-proportional approach significantly.

APPENDIX A. PROOF OF THEOREM 1

The following result from Reference 10 is used to prove our theorem.

Let $X_1,\ldots,X_N$ be independent random variables, with $X_i \in [0,1]$ for each i; otherwise their distributions are arbitrary and possibly different. Let $Y = (1/n)\sum_{i=0}^{n} X_i$, $p = E(Y)$ and $q = 1 - p$. Then for each t with $0 \le t < q$ the following bound holds:

$$P(Y - p \ge t) \le \left[\left(\frac{p}{p+t}\right)^{p+t}\left(\frac{q}{q-t}\right)^{q-t}\right]^n \qquad (18)$$

Using

$$P(Y - p \ge t) = P(nY - np \ge nt) \qquad (19)$$

$$= P\left(\sum_{i=0}^{n} X_i \ge n(p+t)\right)$$

the bound can be reformulated as

$$P\left(\sum_{i=0}^{n} X_i \ge n(p+t)\right) \qquad (20)$$

$$\le \left(\frac{np}{n(p+t)}\right)^{n(p+t)}\left(\frac{1-p}{1-(p+t)}\right)^{n-n(p+t)}$$

Setting $X = nY = \sum_{i=0}^{n} X_i$, $D = np$ and $C = n(p+t)$, we have $D = E(X)$ and the first factor on the right-hand side of (20) becomes

$$\left(\frac{np}{n(p+t)}\right)^{n(p+t)} = \left(\frac{D}{C}\right)^{C} \qquad (21)$$

Similarly, the second factor can be reformulated as

$$\left(\frac{1-p}{1-(p+t)}\right)^{n-n(p+t)} = \left(\frac{n-D}{n-C}\right)^{n-C} \qquad (22)$$

To get rid of n, we set

$$x = \frac{n-C}{C-D}$$

Assuming $C > D$, we apply the inequality

$$\left(1 + \frac{1}{x}\right)^{x} < e\ [(n-C)/(C-F)]\ (C-D)$$

This yields

$$\left(\frac{n-D}{n-C}\right)^{n-C} = \left(1 + \frac{C-D}{n-C}\right)^{n-C} \qquad (23)$$

$$= \left[\left(1 + \frac{1}{x}\right)^{x}\right]^{C-D} < e^{C-D}$$

which together with (21) results in the bound

$$P\,(X \geq C) < \left(\frac{D}{C}\right)^{C} e^{C-D}$$

To prove the second part of the theorem, we use the fact that for binomial random variables Hoeffding's bound reduces to Chernoff's bound, which is known to have the said asymptotic optimality with respect to the exponent,[12] in the strong sense—for every asymptotic optimality with respect to the exponent, all cases with $n \geq n_0$ are counter-examples. Since the only step where we introduce additional error in the bound is the estimation (10), it is enough to show that this additional error tends to zero as n grows. This follows from the fact that for fixed D and C, $x \to \infty$ as $n \to \infty$, and therefore we have

$$\lim_{x \to \infty} \left[\left(1+\frac{1}{x}\right)^{x}\right]^{C-D} = e^{C-D} \tag{24}$$

APPENDIX B. PROOF OF THEOREM 2

Let us choose indices i, j such that the GoS bound $(D/B)^B\, e^{B-D}$ is the smallest for traffic type T_i and the largest for T_j for an allocation. Assume that

$$\left(\frac{D_i}{B_i}\right)^{B_i} e^{B_i - D_i} < \left(\frac{D_j}{B_j}\right)^{B_j} e^{B_j - D_j}$$

holds. Then, by continuity of the function $f(B) = (D/B)^B\, e^{B-D}$, we can decrease B_i with a sufficiently small $\epsilon > 0$ such that the above inequality still holds. In order to restore $\sum_i B_i$, we can increase B_j by ϵ. This, by the strictly decreasing nature of the function $f\,(B)$, yields that the maximum saturation bound is decreased, so the original allocation cannot be optimal. The same argument works if the maximum is achieved at more than one index, in which case ϵ should be distributed among the corresponding B_j values.

APPENDIX C. PROOF OF UNIQUE SOLVABILITY OF EQUATION (8)

Consider the function $f(B) = (D/B)^B\, e^{B-D}$ for a fixed D. We have $f(D) = 1$ by simple substitution. The derivative of $f(B)$ is computed as

$$\frac{df(B)}{dB} = \frac{d}{dB}\, e^{B\ln D - B\ln B + B - D}$$
$$= f(B)(\ln D - \ln B)$$

which is negative for $B > D$, so the function is strictly decreasing. To see that it tends to zero, it is enough to reformulate it as

$$f(B) = \left(\frac{D}{B} e^{1-F/C}\right)^{C}$$

This shows that for cases such as $B > 2eD$

$$f(C) < 2^{-C}$$

holds. These facts imply the unique solvability of equation (8) for $0 < \sigma \leq 1$.

APPENDIX D. PROOF OF THEOREM 3

The algorithm starts with the two extreme values of the GoS bound, i.e. zero and unity. Then it proceeds by iteratively halving the interval between the upper and the lower bound, and each time the $B_1(\sigma)$, …, $B_N\,(\sigma)$ values are computed for the middle point σ of the interval. If the capacity sum B is too small, then the new value of σ is taken as the middle point of the lower half-interval, otherwise as the upper half-interval. By the definition of $B_i(\sigma)$ and knowing that the function $f(B) = (D/B)^B\, e^{B-D}$ tends to zero in a monotonically decreasing way as B grows (see Appendix C), we have that a σ value is approached with the property that

$$\sum_{i=0}^{N} B_i(\sigma) = B$$

and

$$\sigma = \left(\frac{D_1}{B_1(\sigma)}\right)^{B_l(\sigma)} e^{B_l(\sigma) - D_1} = \dots$$
$$= \left(\frac{D_N}{B_N(\sigma)}\right)^{B_N(\sigma)} e^{B_N(\sigma) - D_N}$$

Theorem 2 implies that this is an optimal solution for (8).

The rate of convergence is geometric with respect to σ, since in each iteration the interval containing σ is halved. Then the continuity of the function $(f(B) = (D/B)^B\, e^{B-D}$ implies geometric convergence for the capacities as well.

APPENDIX E. DETERMINING PDF OF PEAK SIZE FOR VIDEO FRAMES

In order to find $F_Y(y)$, we do the following. Let us define a quantity α_n, known as the characteristic function of Y_n by

$$F_X\,(\alpha_n) = 1 - \frac{1}{n} \tag{25}$$

It is thus the value of X_j, $j = 1,2,\ldots,n$, at which $P(X_j \leq \alpha_n) = 1-1/n$. As n becomes large, $F_X(\alpha_n)$ approaches unity or α_n is in the extreme right tail of the video frame size distribution. It can also be shown that α_n is the mode of Y_n (this can be verified, in the case of X_j being continuous, by taking the derivative of $f_{Yn}(y)$ from equation (12) with respect to y and setting it to zero). If $F_X(x)$ takes the form of equation (13), then

$$1 - e^{-g(\beta_n)} = 1 - \frac{1}{n} \quad \text{or} \quad e^{g(\alpha_n)} = n \tag{26}$$

Substituting equations (25) and (26) in equation (12),

$$F_{Y_n}(y) = \left(1 - \frac{e^{-[g(y)-g(\alpha_n)]}}{n}\right)^n \tag{27}$$

Since α_n is the mode or the 'most likely' value of Y_n, the function $g(y)$ in equation (27) can be expanded in powers of $y - \alpha_n$ in the form

$$g(y) = g(\alpha_n) + \frac{dg(y)}{dy}\Big|_{y=\alpha_n} (y - \alpha_n) + \ldots \tag{28}$$

Let $dg(y)/dy|_{y=ga_n} = \lambda_n > 0$, since $g(y)$ is an increasing function of y). Retaining only the linear terms in equation (28) and substituting into equation (27), we obtain

$$F_{Y_n}(y) = \left(\frac{1 - e^{-\lambda_n(y-\alpha_n)}}{n}\right)^n \tag{29}$$

Here λ_n and α_n are functions only of n and not of y. Using the fact that $\lim_{n\to\infty}(1 - a/n) = e^{-a}$, as n becomes large ($n \to \infty$), equation (13) reduces to

$$F_Y(y) = \exp(-e^{-\lambda(y-\alpha)}), \quad -\infty < y < \infty \tag{30}$$

where α and λ (>0) are the location and the scale parameter of the distribution respectively; α is obtained from α_n as $n \to \infty$, and the scale parameter λ is a limiting case of λ_n. The above distribution has a skew coefficient that is a non-negative constant, implying that the shape of the distribution is fixed with a dominant right tail.

REFERENCES

1. P. Bahl and I. Chlamtac, 'H.263 based video codec for real-time visual communications over wireless radio networks', *Proc. IEEE Int. Conf. on Universal Personal Communications*, San Diego, CA, October 1997.
2. P. Bahl and W.-L. Hsu, 'Adaptive region based, multi-scale, motion compensated video coding for error prone, bandwidth limited communication channels', *Proc. Multimedia Networks and Applications, SPIE's Int. Symp. on Voice, Video, & Data Communications*, Dallas, TX, November 1997.
3. K. W. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*, Springer, New York, 1995.
4. Y. H. Kim and C. K. Un, 'Analysis of bandwidth allocation strategies with access restrictions in broadband ISDN', *IEEE Trans. Commn.* **COM-41**, 771–781 (1993).
5. B. Kraimeche and M. Schwartz, 'Bandwidth allocation strategies in wideband integrated networks', *IEEE J. Select. Areas Commun.* **SAC-4**, 869–878 (1986).
6. M. Schwartz, 'Network management and control issues in multimedia wireless networks', *IEEE Personal Commn.*, **2**, (5), 8–16 (1995).
7. P. Bahl, 'Real-time visual communications over narrowband wireless radio networks', *PhD Dissertation*, University of Massachusetts Amherst, 1997.
8. W. C. Y. Lee, *Mobile Cellular Telecommunications, Analog and Digital System*, McGraw-Hill, New York, 1995.
9. X. Qiu and V. O. K. Li, 'Dynamic Reservation Multiple Access (DRMA): a new ultiple access scheme for personal communication systems (PCS)', *ACM J. Wireless Netw.* **2**, (1996).
10. W. Hoeffding, 'Probability inequalities for sums of bounded random variables', *J. Am. Statist. Assoc*, **58**, 13–30 (1993).
11. P. Bahl and I. Chlamtac, 'Influence of available bandwidth on the statistical characterization of compressed video', *Tech. Rep. TR-96-CSE-7*, ECE Department, University of Massachusetts, 1996.
12. W. Weibull, 'A statistical theory of the strength of materials', *Proc. Royal Swedish Institute for Engineering Research, No. 151*, Stockholm, 1939.
13. P. Bahl, I. Chlamtac and A. Farago, 'Optimizing resource utilization in wireless multimedia networks', *Proc. IEEE Conf. on Computer Communications*, Montreal, June 1997.
14. E. J. Gumber, *Statistics of Extremes*, Columbia University Press, New York, 1988.
15. ITU-T, SG15, *ITU Recommendation H.263*, 'Video coding for narrow telecommunication channels at < 64 kbit/s', (1995).

Authors' biographies

Paramuir Bahl holds a Ph.D. in Computer Systems Engineering from the University of Massachusetts Amherst. He is currently with Microsoft Corporation where he is carrying out research in peripatetic computing, multi-hop multi-service ad hoc networks, mobility-aware self-configuring operating systems and real-time audio-visual communications. Prior to being with Microsoft, Dr Bahl spent 9 years at Digital Equipment Corporation. There he initiated, led, and contributed to several seminal multimedia projects, including the industry's first desktop hardware adapters for audio/video compression, world's fastest software implementation of ISO and ITU audio/video codes, and fast and scalable image and video rendering. In 1994 he received Digital's multi-year Engineering Fellowship Award for carrying out research in real-time visual communications over wireless networks.

Dr Bahl is the Vice Chair of ACM SIGMOBILE and founding Editor-in-Chief of *ACM Mobile Computing and Communications Review*. He has served as a guest editor for *IEEE Journal on Selected Areas in Communications, ACM Journal on Special Topics in Mobile Networking and Applications* and *IEEE Communications Magazine*. Dr Bahl has served on a number of organizing and technical program committees of IEEE and ACM including MobiCom, INFOCOM, PIMRC and MoMuC. He has organized sessions and panels on the topics of multimedia communications, wireless LANs and mobile computing. He has authored more than two dozen publications and is the author of one patent issued and several patents pending in the areas of digital signal processing, computer communications and wireless networks. Dr Bahl is a senior member of the IEEE, a member of ACM and a past president of the electrical engineering honour society Eta Kappa Nu (Zeta Pi Chapter).

Imrich Chlamtac holds a PhD in computer science from the University of Minnesota (1979) and BSc and MSc degrees in mathematics awarded with the Highest Distinction. He currently holds the Distinguished Chair in Telecommunications at the University of Texas at Dallas. He is also President of Boston Communications Networks, a company dealing with research and development of wireless sytems and high-speed networks. He is the author of over 200 papers in refereed journals and conferences, multiple book chapters and co-author of *Local Networks* (Lexington Books, 1981). Dr Chlamtac is the founding Editor-in-Chief of the *ACM–URSI–Baltzer Wireless Networks* (WINET) and *ACM–Baltzer Mobile Networking and Nomadic Applications* (NOMAD) journals and served on the editorial board of *IEEE Transactions on Communications* and several other journals. He served as the General Chair of several ACM and IEEE conferences and workshops and is the founder of ACM/IEEE MobiCom. Dr Chlamtac is an IEEE Fellow and ACM Fellow and in the past was a Fulbright Scholar and an IEEE, Northern Telecom and BNR Distinguished Lecturer.

Andras Farago received his MSc and PhD in electrical engineering from the Technical University of Budapest in 1976 and 1981 respectively, and in 1996 he obtained the distinguished title ‘Doctor of the Hungarian Academy of Sciences’. He is a member of the Department of Telecommunications and Telematics, Technical University of Budapest and obtained the distinguished Szchenyi Professorship from the Hungarian Ministry for Education. At the Technical University of Budapest he is co-founder of the High Speed Networks Laboratory, a telecom research laboratory supported by Ericsson. He worked as a visiting Senior Research Fellow at the University of Massachusetts Amherst in 1991–1992. He spent a sabbatical year at Boston University in 1996. Recently he obtained a professorship at the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas. He is member of IFIP Working Group 6.3 ‘Performance of Communication Systems’ and founder member of the Hungarian Chapter of ACM. His main research interest is in developing algorithms and applying mathematical methods for the design and optimization of various types of telecommunication networks and protocols.