Optimizing Resource Utilization in Wireless Multimedia Networks

Paramvir Bahl
Digital Equipment Corporation
Massachusetts, USA
bahl@acm.org

Imrich Chlamtac
University of Texas at Dallas
Texas, USA
chlamtac@utdallas.edu

András Faragó
Technical University of Budapest
Budapest, Hungary
farago@ttt-atm.bme.hu

Abstract

The task of supporting integrated multi-rate multimedia traffic in a bandwidth poor wireless environment poses a unique and challenging problem for network managers. 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 (QoS) guarantees for digital video while minimizing the maximum blocking probability for voice and data connections. At the connection level, optimum utilization of the reserved bandwidth is achieved through intra-frame statistical multiplexing, *while at the system-level, the delicate task of partitioning the bandwidth is accomplished by developing an efficient algorithm which uses traffic parameters consisting only of aggregate traffic load and the total available bandwidth. The algorithm built on non-trivial mathematical results, is simple, robust, and well suited for practical implementations.*

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 different kinds of multimedia services with a reasonable level of performance and without letting the demand from any one type shut-out other types of services. The challenge for the 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 connections as possible. This task of providing guaranteed quality of service 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.

Supporting real-time VBR video along with voice and data over bandwidth-constraint networks continues to be formidable problem. The difficulty arises because VBR video is unpredictably bursty and because it requires performance guarantees from the network. While resource reservation schemes work best for 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 if 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. This is the problem we address in this paper, that is, *can performance guarantees be provided to VBR video without significantly under-utilizing the bandwidth and can this be done in conjunction with minimizing the maximum blocking probability for voice and data connections?*

Our solution to this problem consists of three parts: (1) use a joint source-channel video codec, (2) provide connection lifetime reservation for real-time video connections with optimum bandwidth utilization, and (3) partition the available bandwidth in a manner that ensures that the maximum blocking probability for voice and data traffic is minimized.

From a connection's perspective, we advocate the use of a multi-resolution joint source-channel video codec Depending upon the application, this codec may either be a subband video codec or a region-based video codec [1], [2]. We then propose reserving the peak bandwidth for the primary subband (or region) while letting the secondary and tertiary subbands (regions) compete for bandwidth dynamically. A potential problem with reserving according to the peak requirement is that most of the time the actual amount used by the primary subband is far below the amount reserved. To avoid under-utilizing and wasting reserved bandwidth we have introduced the notion of *intra-frame statistical multiplexing* in [1]. The bandwidth left over after the primary subband (region) has been transmitted is used for transmitting the remaining subbands (or regions). Also packets received in error and whose retransmission has been requested by the receiver can be sent using this left-over bandwidth. In essence the idea combines statistical multiplexing

In the *Proceedings of the IEEE International Conference on Communications*, Montréal, Québec, Canada (June 1997*)*

at the system level with statistical multiplexing at the connection level to achieve optimum utilization.

From the system's perspective, we develop a simple algorithm that partitions the available bandwidth in a manner that minimizes the maximum blocking probability for voice and data connections while guaranteeing bandwidth for VBR video connections. It should be noted, that even when the distribution of the different traffic types is given, finding the optimal partitioning of bandwidth is a very difficult task, and for the general case can be modeled by an NP-complete graph coloring problem. The intractability of finding the optimum is present already in the simplest situation when the traffic consists of voice connections only and the statistics of the offered traffic are completely known. However the problem becomes even more difficult when the wireless network is carrying integrated non-homogeneous traffic, a situation occurring naturally in the case of wireless multimedia networks. In this case estimating the blocking probability of connections and its application in resource allocation strategies is further complicated for two fundamental reasons:

- Although there are methods for computing blocking probabilities for integrated systems under specific statistical assumptions [8] (e.g. multirate Poisson models), there are no simple closed formulas that can easily be applied to optimizing resource allocation.
- It is realistic to expect that traditional statistical assumptions 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 sections 2 and 3, we propose a solution for the allocation of transmission resources among different traffic classes under incompletely known conditions. When combined with our *intra-frame statistical multiplexing* proposal [1], our solution has the following main properties:

1. It provides guaranteed QoS for on-going real-time video sessions. This guarantee does not come at the expense of bandwidth, since all of the reserved bandwidth is utilized through intelligent 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 classes, but *not* on the 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 aggregate flow rates, as the detailed knowledge of the traffic mix would not significantly contribute to achieving smaller loss.

2. Bandwidth Partitioning

Several bandwidth partitioning strategies that allocate bandwidth "fairly" for different traffic classes while attempting to achieve maximum network throughput have been proposed in literature [4], [5], [6]. Previous studies of these techniques in wireless networks have focused on the co-existence of data and voice traffic, while packet video has generally been ignored.

At the two extremes of such strategies is the Complete Sharing (CS) and Complete Partitioning (CP) (also called Mutually Restricted Access) strategies, and in between are the rest, generally referred to as hybrid 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 classes. 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 class 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 Borrowing (PB). A moving boundary exists between the bandwidth allocated for the various traffic classes and priority users (usually 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 both in micro-cellular and macro-cellular environments [5]. The reader is referred to [4] for a survey of such schemes.

2.1 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 wouldn't be too extreme to claim that the only practical way to guarantee quality of service is by providing bandwidth reservation for 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 [1] and [2] static reservation can be provided without wasting the precious bandwidth. With this technique we can build into a medium access protocol provisions for both static (lifetime) and dynamic bandwidth reservations. 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 additional bandwidth is available.

A natural question is to ask is, which is the best scheme (CS, CP, or PB) for bandwidth allocation from the point of view of providing guaranteeing quality of service for visual communications. Clearly complete sharing is not suitable. Complete Partitioning, on the other hand can deliver, but as noted earlier is wasteful of bandwidth. Priority Borrowing is thus the most viable candidate. We have extended Priority Borrowing to include static bandwidth reservation with a moving boundary. We call this new scheme *Priority Sharing with Restrictions* (PSR). Figure 1 illustrates this scheme.

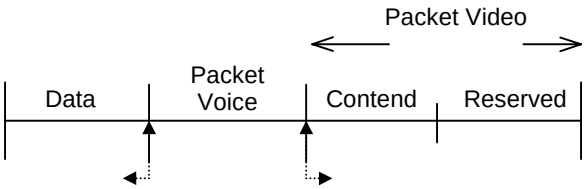

Figure 1: Priority Sharing with Restrictions

Briefly, in this scheme only real-time video connections are allowed to make lifetime (or static) reservations; voice and data connections make reservations dynamically. At connection establishment time bandwidth for video connections (for the main subband or region) is allocated from the *Reserved* portion the available spectrum.

	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, preemption possible; X - Borrowing not allowed; - - Don't care

Table 1: Rules for Priority Sharing with Restrictions

The amount of bandwidth reserved for such static reservations is determined by the network designers (see section 3). The remaining spectrum is divided among voice, data and video (for secondary, tertiary subbands/regions) users and is used for dynamic burst level reservation. In terms of priority, voice users have a highest priority, followed by video users, and data users in that order. Table 1 provides an example of who can borrow from whom..

Figure 2 shows a sample of the performance data for the PSR scheme as compared to the PB and CS schemes. Details of the experiment are provided next to the graph. The video codec used for the experiment was a region-based ITU's H.263 video codec. A segmentor preceded the H.263 codec dividing the image into 5 distinct regions before compression. This is shown Figure 3. The average voice and data traffic load was *30 Erlangs* and *10 Erlangs* respectively. The traffic load for video was increased by using the same compressed video stream multiple 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.

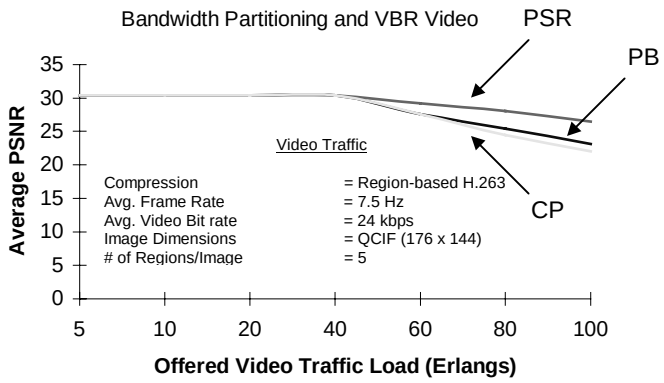

Figure 2: Comparisons of Bandwidth Partitioning Schemes

3. Minimizing Maximum Blocking Probability

Let us consider a wireless multimedia network supporting N traffic types (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 class T_i, is known ahead of time.

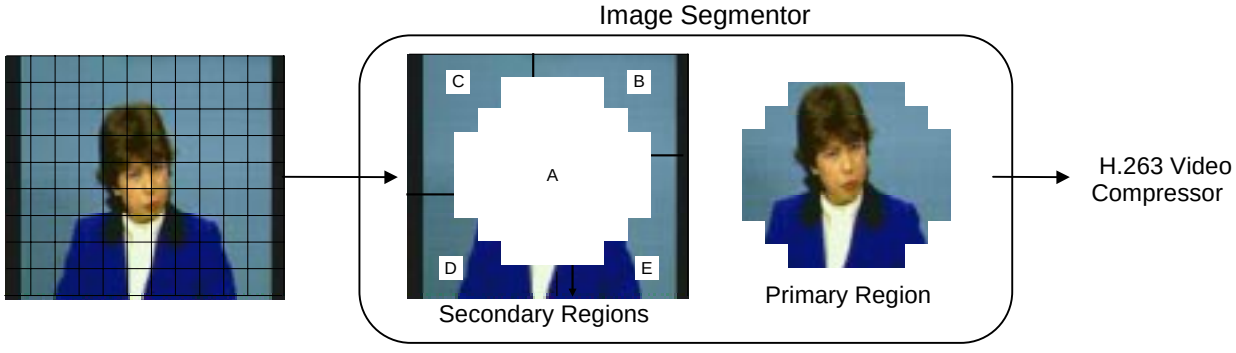

Figure 3: Region based H.263 codec

Now, given this traffic demand, we wish to determine a nominal allocation of the bandwidth B to be given 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 (BBU's). 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 one BBU.)

Let $d_i(t)$ be a random variable describing the actual instantaneous aggregated demand by traffic class T_i. Assuming stationarity, the average demand D_i is the expected value of $d_i(t)$ independent of t, that is, $D_i = E\{d_i(t)\}$. We measure the Grade of Service (GoS) for traffic class 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 class 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[1]:

$$\Theta = \max_i \Theta_i = \max_i P(d_i(t) \geq B_i) \qquad (1)$$

Thus, our optimization task can be stated as follows:

Given the aggregate load D_i for each traffic class T_i and the total system bandwidth B, determine the allocated transmission capacity B_i for each traffic type such that

$$\Theta = \max_i P(d_i(t) \geq B_i) \quad \textit{is minimized,}$$

$$\textit{subject to} \quad \sum_{i=1}^{N} B_i = B \qquad (2)$$

The *Smart Allocate* Algorithm

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 can one 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 probabilities, known from the theory of large deviations, typically require much more information, such as 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.

[1] For example, $\Theta = 0.01$ represents at most a 1% probability that a given request will be blocked due to unavailability of sufficient bandwidth.

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:

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

Furthermore, this estimation is the best possible in the following sense: For any fixed $\varepsilon > 0$ and for any fixed D and C with $C \geq D$ there exist infinitely many counterexamples for which

$$P(X \geq C) > \left(\frac{D}{C}\right)^C e^{C-D-\varepsilon n} \textit{ holds} \qquad (4)$$

The proof of Theorem 1 is based on a bound due to Hoeffding [7], which is a powerful generalization of the well known Chernoff bound on the tail of the binomial distribution. Note that Chernoff's bound alone would not be enough for our purpose. For space limitations we omit the proof, it can be found in [3]

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 GoS estimation:

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

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 class T_i and the total system transmission bandwidth B, determine the allocated transmission capacity B_i for each traffic class, such that

$$\tilde{\Theta} = \max_i \left\{ \left(\frac{D_i}{B_i}\right)^{B_i} e^{B_i - D_i} \right\} \quad \textit{is minimum,} \qquad (6)$$

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

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 $B_1, \ldots, B_N$ with $\sum_{i=1}^{N} B_i = B$ is an optimal solution for (6) if and only if:

$$\left(\frac{D_1}{B_1}\right)^{B_1} e^{B_1 - D_1} = \ldots = \left(\frac{D_N}{B_N}\right)^{B_N} e^{B_N - D_N} \quad \textit{holds.} \quad (7)$$

For the proof of Theorem 2 see [3].

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 $0 < \sigma \le 1$, let $B_i(\sigma)$ be the unique solution for B_i of the equation:

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

The unique solvability of equation (8) follows from the facts that for $B_i = D_i$ the left-hand side is 1 and otherwise it is a strictly decreasing continuous function of B_i that tends to 0 as B_i grows. (For a proof see [3]). It follows from this 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 (6) such that we successively iterate a value $0 < \sigma \le 1$ for which $\sum_i B_i = B$ holds within a given error bound $\varepsilon > 0$. This algorithm that we call *Smart Allocate* is shown in Figure 4.

The correctness of the algorithm and the rate of convergence are stated in the following theorem (for the proof see [3]).

Theorem 3 *Algorithm Smart Allocate converges to an optimal solution of equation (8) at geometric rate, i.e. the error decreases exponentially.*

It should be noted 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 common sense 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 *eq. (7)*, which is the asymptotically optimal GoS estimation, is obtained if and only if

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

holds, which is different in general.

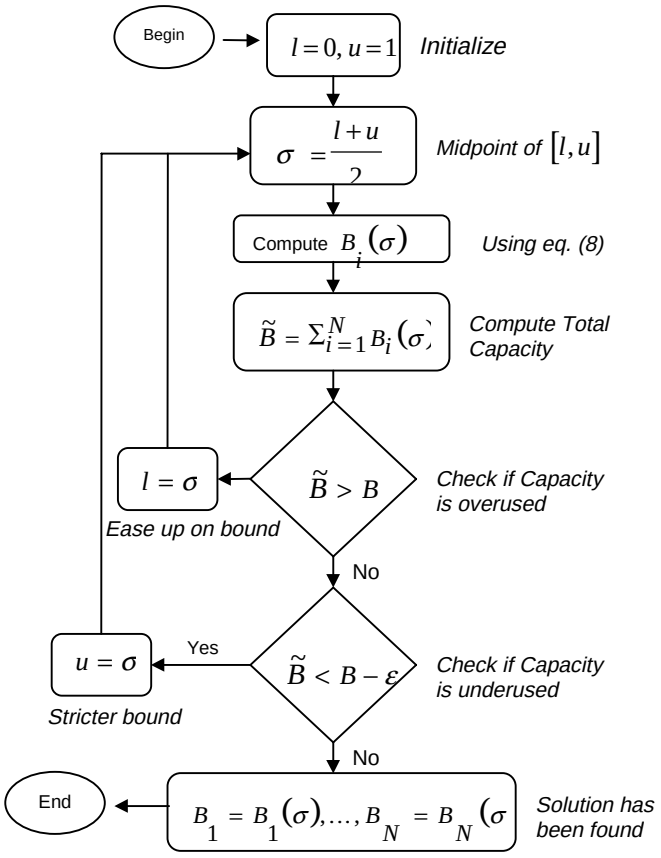

Figure 4: Algorithm *Smart Allocate*

Calculating D_i for Digital Video, Voice and Data

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:

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

In (10) it is assumed that there are finitely many possible b values and they are normalized such that $0 \le b \le 1$ 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 don't 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 examining the density functions of the primary subbands (region) of several similar video sequences and then deriving the distribution for

the maximum (peak) values. From this derived distribution the desire mean can be computed [10].

In [10] it is shown that the distribution for the peak values of video frame sizes is determined to be:

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

where u and $\alpha\ (>0)$ are the location and the scale parameters of the distribution, and Y is the random variable representing the peak values of the primary region in various video sequences. Then it is trivial to show that the mean of Y is given as:

$$\eta_y = u + \left(\frac{0.577}{\alpha}\right)$$

finally, $D_{video} = M \times \eta_y$, where M is the estimated number of video connections to be supported; Assuming CBR for voice connections with a nominal rate of R, $D_{voice} = N \times R$, where N is the estimated number of voice connections; Knowing D_{video} and $D_{voice,}$ we get $D_{data} = B - D_{video} - D_{voice}$.

3.1. Illustrative Example

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). Lets consider a TDM-based 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 types 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 $B_2 = 66$ BBUs. Then our upper bound on the largest blocking probability gives the value 0.57%. An improvement of 43% over the load proportional approach.

Implicit in the calculations above is the assumption that one type of users do not borrow BBUs from other type of users. Thus, for the case of *Partial Sharing with Restrictions*, the largest blocking probability calculated by the *Smart Allocate* algorithm is a conservative estimate. If we take into account the rules outlined in Table 1 and incorporate borrowing of BBUs, then with the assumption of independence, the maximum blocking probability calculated will be smaller than the one computed above.

4. Conclusion

We have presented a bandwidth reservation, utilization, and partitioning strategy for supporting multirate multimedia traffic in resource constrained wireless networks. Through adept combination of connection -level improvements — using a low-bit rate multi-resolution VBR video codec with intelligent bandwidth reservation, and system-level improvements — partitioning available bandwidth, a wireless network can simultaneously support digital video (with QoS guarantees), voice, and data communications. Our method for partitioning and allocating transmission capacities to different traffic classes is useful as it does not require detailed prior knowledge of the underlying traffic. The partitioning algorithm (the *Smart Allocate* algorithm) is simple, easy to implement, and the geometric rate of convergence ensures that the result is found quickly. These properties make it well suited for practical application, even in the case when the aggregate load values change and the bandwidth allocation has to be re-computed from time to time.

References

[1] P. Bahl and I. Chlamtac, "Bandwidth Allocation in Wireless Networks for Multiresolution VBR Video traffic," *Proceedings of the IEEE Computer Communications Workshop*, Reston, Virginia (Sept. 1996)

[2] P. Bahl and I. Chlamtac, "Strategies for Transmission of Compressed Video Over Error Prone Radio Channels," *Proceedings of Workshop on Mobile Multimedia Communications,* Princeton, New Jersey (Sept. 1996)

[3] P. Bahl, I. Chlamtac, and A. Faragó, "A Novel Approach to Bandwidth Partitioning in Wireless ATM Networks," *University of Massachusetts, ECE Dept.* TR-96-CSE-10

[4] Y. H. Kim, and C. K. Un, "Analysis of Bandwidth Allocation Strategies with Access Restrictions in Broadband ISDN," *IEEE Transactions on Communications*, Vol. 41, No. 5 (May 1993): 771-781

[5] M. Schwartz, "Network Management and Control Issues in Multimedia Wireless Networks," *IEEE Persoanl Communications*, Vol. 2, No. 5, (June 1995): 8-16

[6] B. Kraimeche and M. Schwartz, "Bandwidth Allocation Strategies in Wideband Integrated Networks," *IEEE Selected Areas in Communication*s, Vol. SAC-4 (Sept. 1986): 869-878

[7] W. Hoeffding, "Probability Inequalities for Sums of Bounded Random Variables," *Journal of American Statistics Association*, Vol. 58 (19963): 13-30

[8] K. W. Ross, Multiservice Loss Models for Broadband Telecommunication Networks, Springer, 1995

[9] E. J. Gumber, *Statistics of Extremes*, Columbia University Press, New York, 1988

[10] P. Bahl and I. Chlamtac, "Influence of Available Bandwidth on the Statistical Characterization of Compressed Video," University of Massachusetts, ECE Dept. technical Report TR-96-CSE-7