University of Würzburg
Institute of Computer Science
Research Report Series

Carrying Wireless Traffic over IP Using Realtime Transport Protocol Multiplexing

Michael Menth

Report No. 247 January 2000

Department of Distributed Systems
Institute of Computer Science
University of Würzburg
Am Hubland, D-97074 Würzburg, Germany
menth@informatik.uni-wuerzburg.de

Carrying Wireless Traffic over IP Using Realtime Transport Protocol Multiplexing

Michael Menth
Department of Distributed Systems
Institute of Computer Science
University of Würzburg
Am Hubland, D-97074 Würzburg, Germany
menth@informatik.uni-wuerzburg.de

Abstract

The RTP/UDP/IP protocol suite is in the process of being standardized for multiplexing realtime flows. It aims at reducing header overhead, therefore, it is apt for transporting compressed voice in access links or in the core of mobile communication networks. To provide realtime quality of service, networks require traffic parameters which can be met using a spacer.

In this paper, an exact analysis for the RTP/UDP/IP multiplexing scheme including spacing is derived. Influences on the performance of the system are investigated and an optimum value for the multiplexing timer is found. A numerical comparison to the data tunneling method is also presented. Additionally, a leaky bucket description of the multiplexed IP packet output stream is given as an alternative to spacing. The affinity to AAL-2 multiplexing in ATM is pointed out and the fundamental differences with respect to performance are explained.

Keywords: IP, RTP, Multiplexing, UMTS, AAL-2, CAC, QoS, VoIP, Discrete-Time Analysis

1 Introduction

The success of the Internet Protocol (IP) has started the discussion in the standardization community of the 3rd generation of mobile communication systems (3GPP) to introduce IP as the transport technology in the wireline part of future mobile cellular communication systems [1]. Characteristics of compressed realtime voice data traffic are the small-sized packets and the strict quality of service (QoS) requirements, i.e., upper bounds on packet loss and delay. Both are problematic with IP technology.

A User Datagram Protocol (UDP) header is mandatory to carry information over IP networks and for voice data an additional Realtime Transmission Protocol (RTP) header is commonly used. Tunneling, i.e., carrying a single voice sample in one IP packet yields a low bandwidth exploitation due to header overhead. This can be overcome by multiplexing several voice samples into the payload of a single RTP/UDP/IP packet. A timer is used to limit the multiplexing delay. Therefore, the RTP multiplexing scheme [2] was recently discussed in the Internet Engineering Task Force (IETF).

So far, the Internet is without realtime capabilities, however, IP is currently being enhanced with realtime enabling techniques. To obtain hard realtime guarantees, a flow

has to declare and to comply with some traffic descriptors. Packets that are not conform are dropped by the network unless they are delayed by a spacer.

In a transmission system with RTP multiplexing and subsequent spacing or policing, the parameters like multiplexing timer value, spacer buffer size or leaky bucket parameters have to be set properly in order to maximize the number of supportable calls for which the QoS characteristics are met. A problem of similar nature occurs in ATM networks and is solved by using the ATM Adaptation Layer 2 (AAL-2) for multiplexing. This has been thoroughly investigated in [3, 4, 5, 6, 7].

In this paper RTP multiplexing is investigated. Case studies are made to set system parameters in order to maximize the performance. In Section 2 we develop a model for the transmission of wireless voice traffic over IP networks and explain the fundamental differences to AAL-2. In Section 3, the analysis is derived and the numerical results are presented in Section 4. Finally, Section 5 concludes the paper.

2 Model for Tunneling and Multiplexing Wireless Data

Traffic originating from a cellular mobile communication system is either tunneled or multiplexed into IP packets and undergoes a spacer before transportation through a real-time IP network. In the following, we concentrate on voice data to develop a model for transmission over IP technology.

2.1 Source Model

In today's cellular mobile communication systems (see Figure 1), voice data originating at a mobile handset is transmitted over the radio interface to a base station. It is transported over a wired access link to the core network and takes then the reverse order to reach the destination handset. We aim at modeling the traffic occurring at the access link.

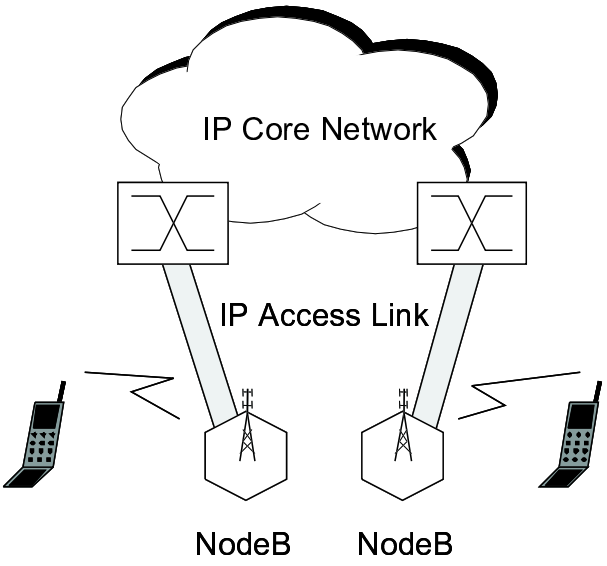

Figure 1: Mobile network structure

In the Universal Mobile Telecommunications System (UMTS) a handset transmits voice samples periodically every 20 msec. Thus, the arrival process is fully characterized by a single $F = 20$ msec time frame. The probability that a sender is associated with an arbitrary instant within that interval is evenly distributed. This entails an exponential distribution of the interarrival time A of consecutive voice samples if the periodic time structure is not taken into account. However, according to [8], the maximum allowed delay is 1 msec which is only $\frac{1}{20}$ of the frame period, so that the periodical structure of the arrival process is not supposed to have great influence on the performance, especially in the presence of many users. The discrete nature of digital communication systems proposes the geometric distribution – the discrete time counterpart of the negative-exponential distribution – to model the interarrival times of consecutively arriving voice samples.

Table 1: Packet length distribution of 8k vocoder

Packet Length [bytes]	12	15	20	32
Probability	0.598	0.072	0.039	0.291

In the considered wireless network a variable bitrate vocoder is used. During an off-phase of a conversation the information can be better compressed than during an on-phase resulting in voice samples of different size. Therefore, a sample trace of a single vocoder is clearly positively correlated. But simulations have shown that for superposition of several users the sizes of consecutively arriving voice samples can be assumed to be sufficiently uncorrelated. Table 1 shows a typical distribution of the sample size gained from an IS-96 vocoder output [3, 9]. So we model the voice sample size B by an independently and identically distributed (iid) random variable according to the given histogram.

2.2 Tunneling and Multiplexing

Carrying very short data packets by tunneling through an ATM or IP network yields a low bandwidth exploitation due to protocol overhead. However, the cause for the overhead is different in both systems.

ATM's cell payload size easily doubles the size of a compressed voice sample which is about 20 bytes. Hence, there is unused space in the cell due to the fixed cell length which represents $\frac{53-20}{20} \cdot 100\% = 165\%$ overhead.

The IP packet size is variable and wasted payload does not exist. However, the IP header size is 20 bytes for the old IP version 4 [10] and even 40 bytes for the new IP version 6[11]. The UDP header has 8 bytes [12] and the RTP header comprises 12 bytes [13]. Hence, the header overhead for voice data transmission over RTP amounts to $\frac{20+8+12}{20} \cdot 100\% = 200\%$.

Multiplexing can be used in both cases to reduce the overhead. In ATM networks, the voice samples are equipped with a 3 bytes header and then transmitted as a stream in the CPS-PDU payload (47 bytes) over the network [14]. A timer controls the multiplexing delay, i.e., if a voice sample waits more than a specified time for cell completion, the cell is sent regardless of whether it is filled or not. Thus, AAL-2 prevents wasted payload. Once an ATM cell is completely filled, the overhead can not be further reduced.

In IP networks, the voice samples are supplied with a 2 bytes mini header [2] and multiplexed into an RTP/UDP/IP packet. The delay must be controlled here by a timer, too. This multiplexing scheme reduces the fraction of the header size with respect to carried payload. The overhead minimization is only limited by the maximum transfer unit of the underlying transport mechanism.

If the offered load in ATM networks is sufficiently high such that most of the cells are completed before the timer stops multiplexing, the timer has nearly no impact [3, 7]. In contrast, the IP packet payload size is variable and, therefore, multiplexing is only limited by the timer. This is the essential difference to AAL-2 multiplexing.

2.3 IP Realtime Transport Parameters

After an IP packet or an ATM cell is filled by multiplexing, it is sent through a realtime network. Realtime transportation requires the network to dedicate enough bandwidth to the desired virtual leased line. In return, the access must be controlled to shelter the QoS from an – intentionally or not – misbehaved source. To achieve that aim, traffic parameters that describe the flow are necessary to make appropriate resource reservations. In ATM's Constant Bit Rate class [15] a peak cell rate must not be exceeded. For Integrated Services' [16] Guaranteed Quality of Service class [17] the data stream must be leaky bucket conform. Whether Differentiated Services [18] will support hard realtime constraints is not clear yet.

If a packet is found not to be conform to the traffic contract at the policer of the network, it will be discarded either immediately at the ingress or will be marked and dropped later within the network if congestion occurs. Therefore, packets must be spaced, i.e., deferred until the traffic contract is met. Spacing introduces additional delay which might have a considerable impact on the overall performance of the system. Another possibility is to describe the IP stream by the leaky bucket parameters transmission rate C and bucket size S_{max}. Using these parameters for policing, there is hardly waiting time for IP packets but still a loss probability due to policing by the network.

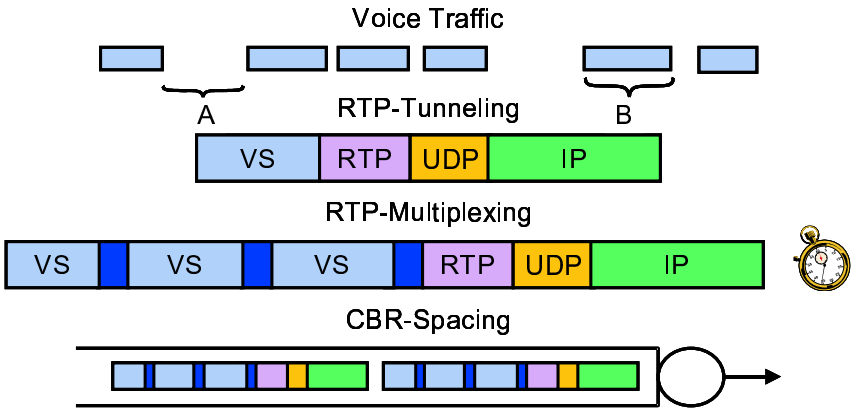

Figure 2: Multiplexing and tunneling of voice samples with spacing

The spacer that we consider works as follows. It has a byte counter S that shows the virtual occupancy of its queue. It is decreased linearly by the link rate C over time but does not fall short of zero. When an IP packet of size B arrives, it is accepted if the counter S plus the new packet's size do not exceed the queue limit S_{max}. In this case the counter is increased by B bytes and the packet will be sent after $\frac{S}{C}$ time. Otherwise, the IP packet is discarded.

Figure 2 summarizes the model for multiplexing or tunneling voice samples with additional spacing. Voice samples arrive with geometrically distributed interarrival times A. Their size B follows a given histogram. Then they are tunneled via the (RTP/)UDP/IP protocol or multiplexed using the RTP/UDP/IP protocol. The resulting IP packets are treated according to the specification of the above described spacer.

3 Analysis of RTP Multiplexing

In this section, an analysis for the RTP multiplexing model is derived. First, a discrete-time Markov model of the system is set up to derive the stationary state distribution. Based on this the loss probability, the waiting time distribution, as well as the average protocol overhead for the voice samples are computed. Adjusting the input parameters, the same analysis can be applied to (RTP/)UDP/IP tunneling.

For the sake of simplicity some notational conventions are introduced regarding an arbitrary random variable X:

$X_{min}, \quad X_{max}$	minimum and maximum value of X,
$X(Y=i)$	random variable X conditioned on $Y=i$,
$x, \quad x[i]$	distribution of X and probability $\Pr(X=i)$,
$x(Y=i)$	distribution conditioned on $Y=i$,
$\overline{X} = \sum_{i=X_{min}}^{X_{max}} x[i] \cdot i$	mean of X,
$\sum_{i=X_{min}}^{X_{max}} p(X=i) \cdot x[i]$	conditonal probability $p(X=i)$ is unconditoned by X.

3.1 Stationary Distribution

To find the stationary state distribution of the model a numerical framework for solving discrete and finite Markov models [6] is applied which is basically a generalized formalization of the method used in [19]. Then, the input distributions for the analysis are specified.

3.1.1 Applying the Framework

The framework is extended by the use of conditional distributions and the generation of a start vector by a short Markov chain simulation. Only a description of the Markov model is needed from which the numerical program can be syntactically deduced. The description comprises renewal points, state variables, factors influencing the system and a state transition function describing the behavior of the system.

The model is as follows. At the arrival instant of the first voice sample multiplexing into an IP packet is started and the spacer counter is denoted by S^0. The multiplexing time is limited by the timer value TCU. After multiplexing, the IP packet contains N voice samples resulting in an IP packet of size $L(N)$ that depends on the number of multiplexed

voice samples. The spacer occupancy is reduced by $TCU \cdot C$ yielding S'. If the updated spacer occupancy plus the size of the new IP packet does not exceed the spacer size S_{max}, the IP packet is accepted for transmission by the spacer increasing its counter by $L(N)$. Then, the spacer counter is denoted by S^1. The time until the next voice sample arrives is called intermultiplex time and denoted by $I(N)$. It is dependent on how many samples are multiplexed into the IP packet. Within that time the counter is diminished by $I(N) \cdot C$. When the next voice sample arrives, the procedure starts again.

The model must be formalized to be applied to the framework. To reduce the computational complexity, we identify two renewal points. The first is at multiplexing start, the second one just after discarding or accepting the multiplexed IP packet for transmission. The factors that influence the state transition from the first renewal point to the second are the number N of voice samples in the multiplexed IP packet and the resulting IP packet size $L(N)$. The next transition is depending on the intermultiplex time $I(N)$. The state of the first renewal point is given by the spacer counter S^0 while the description of the second renewal point comprises both the counter S^1 as well as the number N^1 of multiplexed voice samples in the last IP packet.

The transition function f describes the state evolution between renewal points of the first type. It can be decomposed into $f = f^1 \circ f^0$, where $\circ$ denotes the composition operator. Starting with an initial vector s_0^0, the successor distributions s_i^0 are computed using f and s_{i-1}^0 and the distributions of the factors N, $L(N)$, and $I(N)$. The limit of their average eventually yields the stationary distribution s^0.

Function f^0 (see Algorithm 1) characterizes the system during the multiplexing time. At the end of the multiplexing interval of duration TCU, the spacer occupancy is reduced by $TCU \cdot C$. Within that time, N voice samples have arrived and are multiplexed into an IP packet of size $L(N)$. If the counter plus the packet size exceed the spacer's size S_{max}, the IP packet is accepted for transmission increasing the occupancy, otherwise, the IP packet is discarded.

Input: state (S^0), multiplexed voice samples N, and IP packet size $L(N)$
  $S' := \max(S^0 - TCU \cdot C, 0)$
  if $(S' + L(N) \leq S_{max})$ **then** {IP packet sent}
    $S^1 := S' + L(N)$
  else {IP packet lost}
    $S^1 := S'$
  end if
  $N^1 := N$
Output: state (S^1, N^1)

Algorithm 1: Function f^0 - Multiplexing

Function f^1 (see Algorithm 2) corresponds to the state transition from the second renewal point to the first. It describes the system during the intermultiplex time $I(N)$. The spacer occupancy is reduced linearly by the link rate C.

Input: state (S^1, N^1) and intermultiplex time $I(N^1)$
$S^0 := \max(S^1 - I(N^1) \cdot C, 0)$
Output: state (S^0)

Algorithm 2: Function f^1 - Intermultiplex Time

3.1.2 Specification of Input Distributions

The distributions of N, $L(N)$, and $I(N)$ must be computed to determine the input for the analysis. They are derived from a given voice sample interarrival time distribution a and a given voice sample size distribution b.

During TCU time $N = i$ packets arrive. This happens if the sum of $i - 1$ interarrival times $\sum_{j=1}^{i-1} A_j$ is shorter than TCU and the sum of i interarrival times $\sum_{j=1}^{i} A_j$ exceeds it. Hence, the distribution of N is defined by the condition

$$\Big(\sum_{j=1}^{i-1} A_j \leq TCU < \sum_{j=1}^{i} A_j\Big) \wedge \Big(N = i\Big). \tag{1}$$

The size of an IP packet carrying N voice packets consists of the size of the contained voice samples and protocol headers. The protocol overhead comprises the IP (version 4), UDP, and RTP header of 40 bytes plus 2 bytes mini header for each multiplexed voice sample. Consequently, the IP packet size is given by the conditional random variable

$$L(N = i) = \sum_{j=1}^{i} B_j + 2 \cdot i + 40. \tag{2}$$

The intermultiplex time $I(N)$ is the duration from a multiplexing timeout instant until the next voice sample arrives. Given that there are N samples in the last multiplexed IP packet, the intermultiplex time is the remainder of the n-th interarrival time after multiplexing. Therefore, the distribution of $I(N = i)$ is determined by the condition

$$\Big(\sum_{j=1}^{i-1} A_j \leq TCU < \sum_{j=1}^{i} A_j\Big) \wedge \Big(I(N = i) = \sum_{j=1}^{i} A_j - TCU\Big) \tag{3}$$

In case of a geometrical interarrival time $I(N) = A$ holds.

3.2 Performance Measures

The loss probability p_{loss}^{vs}, the waiting time distribution w, and the overhead o for a voice sample are obtained by using the stationary distribution s^0.

3.2.1 Voice Sample Loss Probability

We consider the multiplexing start. The next IP packet is ready for spacing after TCU time. In the meantime the spacer counter is diminished to $S' = \max(S^0 - TCU \cdot C, 0)$ bytes. If the newly arrived IP packet with $L(N)$ bytes plus the updated counter value S' exceed the spacer's capacity S_{max}, the IP packet is discarded. The loss probability

$p^{IP}_{loss}(N = i, S^0 = j)$ of an IP packet depends on the number N of contained voice samples and on the spacer counter S^0 at multiplexing start

$$p^{IP}_{loss}(N = i, S^0 = j) \quad = \sum_{(L(N=i)=k)+\max((S^0=j)-TCU\cdot C,0)>S_{max}} l(N = i)[k]. \tag{4}$$

The loss probability $p^{vs}_{loss} = \overline{N_{lost}}/\overline{N}$ of a voice sample is the average number $\overline{N_{lost}}$ of lost voice samples in an IP packet divided by the average number $\overline{N}$ of voice samples in an IP packet.

$$\overline{N_{lost}}(N = i, S^0 = j) \quad = \quad p^{IP}_{loss}(N = i, S^0 = j) \cdot i \tag{5}$$

$$\overline{N}(N = i) \quad = \quad i \tag{6}$$

Unconditioning them by N and S^0 yields $\overline{N_{lost}}$ and $\overline{N}$.

3.2.2 Voice Sample Overhead

The overhead $o = \overline{U}/\overline{V}$ is defined by the quotient of the mean sent protocol header size $\overline{U}$ of a transmitted IP packet and the mean of the sent payload size $\overline{V}$. Both are depending on $N = i$ and $S^0 = j$. The protocol header size of an IP packet is then $i \cdot 2 + 40$ and the voice sample payload size $(L(N = i) = k) - i \cdot 2 - 40$ is the IP packet size without the protocol header.

$$\overline{U}(N = i, S^0 = j) \quad = \quad (1 - p^{IP}_{loss}(N = i, S^0 = j)) \cdot (2 \cdot i + 40) \tag{7}$$

$$\overline{V}(N = i, S^0 = j) \quad = \sum_{(L(N=i)=k)+\max((S^0=j)-TCU\cdot C,0)\leq S_{max}} l(N = i)[k] \cdot (((L(N = i) = k) - 2 \cdot i - 40) \tag{8}$$

Again, unconditioning by N and S^0 yields $\overline{U}$ and $\overline{V}$.

3.2.3 Voice Sample Waiting Time

The voice sample waiting time $W = M + Q$ consists of the multiplexing delay M and the queuing time Q in the spacer. It can only be computed for packets that are not lost.

The multiplexing delay M that a voice sample encounters is the time from its arrival instant until the end of multiplexing. It is dependent on the number of multiplexed voice samples in the IP packet: if there is only one voice sample, it is clear that the multiplexing time is TCU; if there are more than one, it is more likely that it is shorter. The distribution for $M(N)$ is determined by

$$\Big(\sum_{j=1}^{i-1} A_j \leq TCU < \sum_{j=1}^{i} A_j\Big) \wedge \Big(0 \leq k < i\Big) \wedge \Big(M(N = i) \quad = \quad TCU - \sum_{j=1}^{k} A_j\Big). \tag{9}$$

At the beginning of multiplexing, the spacer counter is S^0 and it is reduced to $S' = \max(S^0 - TCU \cdot C, 0)$ after multiplexing. At this time the IP packet gets possibly accepted for transmission and encounters a conditional waiting time of

$$Q(S^0) \quad = \quad \max(S^0 - TCU \cdot C, 0)/C \tag{10}$$

The probability of waiting time $W = h$ is the quotient $w[h] = \overline{N_W}(W = h)/\overline{N_{sent}}$ of the average number $\overline{N_W}(W = h)$ of voice samples in an IP packet that have waiting time $W = h$ and the average number $\overline{N_{sent}}$ of sent voice samples in an IP packet.

The average number $\overline{N_W}(W = h, N = i, S^0 = j)$ of voice samples that wait $W = h$ time in an IP packet with $N = i$ voice samples and with a counter of $S^0 = j$ at multiplexing start is

$$\begin{aligned}\overline{N_W}(W = h, N = i, S^0 = j) &= (1 - p_{loss}^{IP}(N = i, S^0 = j)) \cdot i \cdot \\ &\quad \sum_{k=0}^{h} m(N = i)[k] \cdot q(S^0 = j)[h - k].\end{aligned} \tag{11}$$

Unconditioning of N and S^0 yields $\overline{N_W}(W = h)$. For the numerical program, advantage can be taken of the fact that the distribution of $Q(S^0)$ takes only the values 0 and 1.

The average number $\overline{N_{sent}} = \overline{N} - \overline{N_{lost}}$ is the average number of multiplexed voice samples without the lost voice samples.

3.3 Analysis of (RTP/)UDP/IP Tunneling

The analysis of (RTP/)UDP/IP tunneling is a special case of RTP/UDP/IP multiplexing.

- Only one voice sample is transported by an IP packet: $N = 1$ (constant).
- The header overhead of an IP packet tunneling a voice sample only depends on the protocol suite: $H(N) = \begin{cases} 8 + 20 & \text{for UDP/IP tunneling} \\ 12 + 8 + 20 & \text{for RTP/UDP/IP tunneling.} \end{cases}$
- The intermultiplex time equals the interarrival time: $I(N) = A$.
- A multiplexing time is not existent: $M(N) = 0$ (constant).

4 Results

In the previously derived analysis the spacer counter S and the packet size B are measured in bytes while time is measured in discretized time units (TU). The interpretation of numerical results depends on this time scale. If we assume that one TU is $\frac{1}{32}$ msec, the delay budget D of 128 TU corresponds to 4 msec as it is assumed in [3]. Assuming the TU to be $\frac{1}{128}$ msec, the delay budget is 1 msec as it is proposed in [8]. This consideration shows that the analysis is not restricted to this specific parameter set. In the following, we use the latter of the two proposed interpretations.

The QoS requirements for a voice connection are a loss probability of at most 10^{-6} and a quantile of the waiting time distribution for the delay budget 1 msec less than 10^{-4}, i.e., $\Pr(W > 1 \text{ msec}) \leq 10^{-4}$. This is a constraint for all data computed in the following sections. Using simulation, probabilities in the range of 10^{-6} are very time consuming to compute, which illustrates the advantages of an analytical approach over simulation. The analysis output was tested against simulation output and does not differ.

We define the net traffic bandwidth $C^* = n \cdot \frac{\overline{B}}{F}$ where n is the number of users, $\overline{B}$ the mean voice sample size and F the frame length in UMTS. The offered load $\rho = \frac{C^*}{C}$ is the

fraction of the net traffic bandwidth C^* divided by the link bandwidth C. Since the QoS requirements must be met, the offered load plus additional overhead must be smaller than 1. The number of supportable calls can be computed by

$$n = \frac{\rho \cdot C \cdot F}{\overline{B}}. \tag{12}$$

First, the optimum timer value and the mean waiting time for an 8 Mbps link are considered, the required spacer size is evaluated and an equivalent leaky bucket description is given. Then experiments varying on bandwidth are conducted.

4.1 Optimum Timer Value

We investigate the influence of the timer value on the performance of the system. In the following we consider an 8 Mbps link. Initially, the spacer size is set to 2048 bytes, which corresponds to a delay of 2 msec. The consequence is that if a loss occurs, the delay budget QoS criterion (1 msec) has been exceeded before. Thus, the maximum or critical offered load can be found for a given delay constraint. Then, the spacer size is minimized to find the minimum required spacer size.

For the interarrival time of consecutively arriving voice samples the geometric distribution seems to be appropriate for the voice transmission model on the access link. It can be scaled by $\overline{A} = \frac{\overline{B}}{C \cdot \rho}$. The coefficient of variation is $c_{var} = \sqrt{\frac{\overline{A}-1}{\overline{A}}}$ and ranges for the considered values between 0.5 and 1.0. To make investigations more comparable, a modified negative binomial distribution is used since its mean and coefficient of variation can be easily controlled. We make parameter studies for coefficients of variation of 0.5, 1.0 and 2.0 to observe the influence of the variability of the carried traffic.

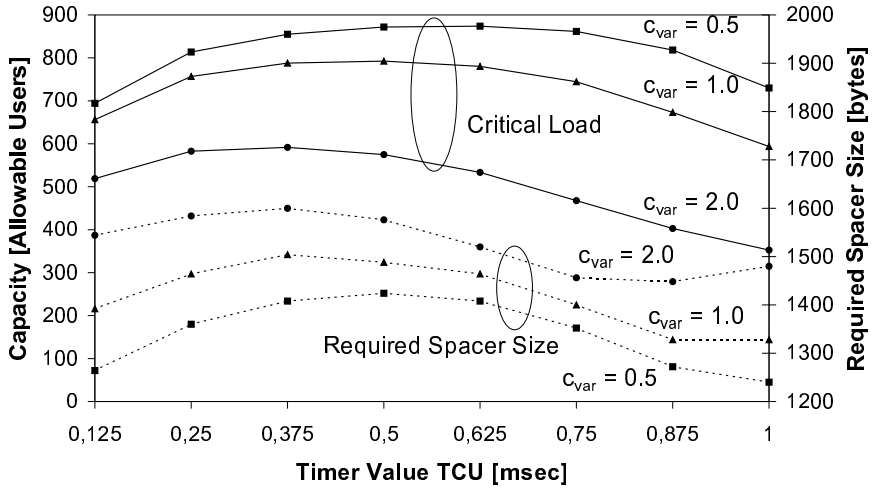

Figure 3: Impact of timer setting on the capacity and the required spacer size

For more variable interarrival times a larger spacer size is needed although the amount of transmitted voice samples is less. The required spacer size is very sensitive to variance in interarrival time. The maximum number of connections can be admitted for a timer

value of 0.375, 0.5 and 0.625 msec when the coefficient of variation is 2.0, 1.0 and 0.5, respectively, as can be seen in Figure 3. The optimum timer value depends on the variability of the interarrival time, however, a timer value of 0.5 msec yields a good performance in all cases. Therefore, all experiments are conducted in the following with the timer value set to 0.5 msec.

The phenomenon that an optimum timer value exists can be explained as follows. On the one side, the overhead decreases with an increasing timer value and the total number of bytes arriving at the spacer is less, thus, reducing the queuing time in the spacer. On the other side, an increased timeout value means a longer multiplexing delay. This denotes a tradeoff for the waiting time $W = M + Q$ which is the sum of multiplexing and spacing delay.

The variance of the interarrival time of the voice samples is transformed into variance of IP packet size. This entails a broader distribution of the spacing time for a more variable interarrival time. The probability that the spacing time exceeds a critical value is higher. This is also confirmed by the larger required spacer size. If the spacing takes longer, the multiplexing time must be shorter to meet the maximum delay D. Hence, a more variable IP packet size requires a shorter multiplexing timer value.

Parameter studies have shown that the optimum timer value does neither depend on the bandwidth nor on the variability of the voice sample size. The coefficient of variation of the histogram is 0.49, values of 0.35 and 0.65 were also tested. However, for a coefficient of variation of 2.0, which is unrealistic for voice sample sizes, this result is not expected to hold.

The existence of a tradeoff regarding the timer value is one of the major differences between RTP multiplexing and AAL-2 multiplexing. The timer limits the AAL-2 multiplexing time only rarely because the small ATM cells are filled before a timeout occurs [3]. It has hardly any effect provided that it is set sufficiently large. Once a cell is filled, the overhead can not be further reduced and, therefore, the above tradeoff does not exist.

4.2 QoS Behavior

Again, we consider an 8 Mbps link with a minimum spacer size of 1488 bytes. The waiting time of a voice sample consists of the multiplexing delay M and the queuing time Q in the spacer. The more traffic arrives the shorter is the average multiplexing delay but the longer is the queuing time in the spacer. Figure 4 quantifies this tradeoff.

The figure also shows the quantile for the delay budget of the waiting time distribution. Its growth is roughly exponentially related to the number of supported users and exceeds the delay budget at the critical load of around 0.71. Note that the waiting time in a very low loaded system is larger than in a fully loaded system. The loss probability depends on the provided buffer size which is set such that its QoS criterion is violated only if the delay criterion is not met any more. It rises exponentially, too.

4.3 Comparison with Tunneling

A small offered load and small bandwidth yield large interarrival times reducing the number of multiplexed voice samples in an IP packet. Therefore, the comparison of the multiplexing and the tunneling scheme must be made depending on the bandwidth and the critical load.

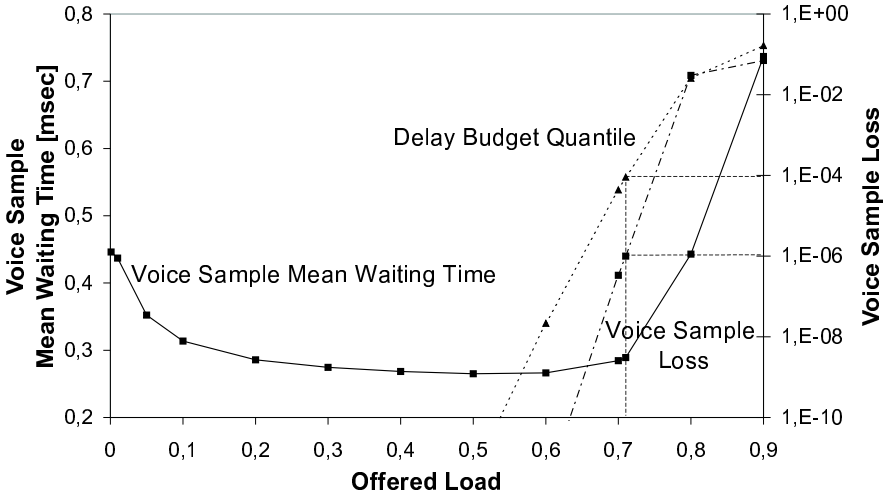

Figure 4: Impact of the offered load on the voice sample mean waiting time, the delay budget quantile, and the loss probability

For higher bandwidth the interarrival time is smaller for the same offered load and more voice samples can be multiplexed. This reduces the proportion of the protocol overhead drastically while for tunneling the overhead stays the same. This is illustrated in Figure 5. Hence, multiplexing allows a higher offered load than tunneling since less protocol overhead is transmitted. It supports an offered load of up to 0.808 whereas (RTP/)UDP/IP tunneling only sustains 0.277 and 0.372, respectively. In other words, to carry 618 calls, a bandwidth of 16 Mbps and 12.5 Mbps is needed for RTP/UDP/IP and UDP/IP tunneling while for multiplexing a bandwidth of only 6.5 Mbps is required.

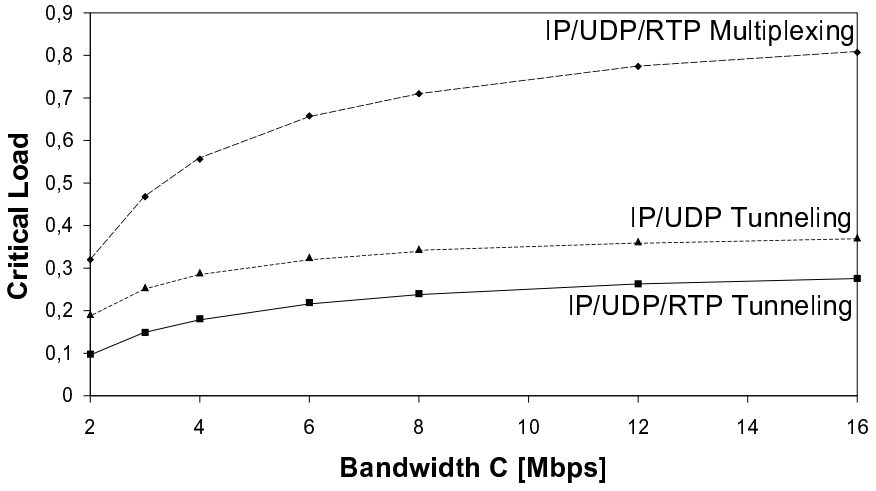

Figure 5: Impact of the bandwidth on the critical load

The protocol overhead is depicted in Figure 6. With tunneling the protocol overhead

is constant while with RTP multiplexing the overhead can be substantially reduced. For the AAL-2 multiplexing scheme this is different. Provided that all cells can be filled, the protocol overhead amounts to $\frac{6/(47/(\overline{B}+3))+3}{\overline{B}} = 0.312$ (ATM header: 5 bytes, AAL-2 overhead: 1 byte per cell and 3 bytes per sample) and can not be further reduced. The same protocol overhead reduction can be reached for RTP multiplexing at a bandwidth of 6 Mbps and for 16 Mbps the overhead is only half of the AAL-2 overhead.

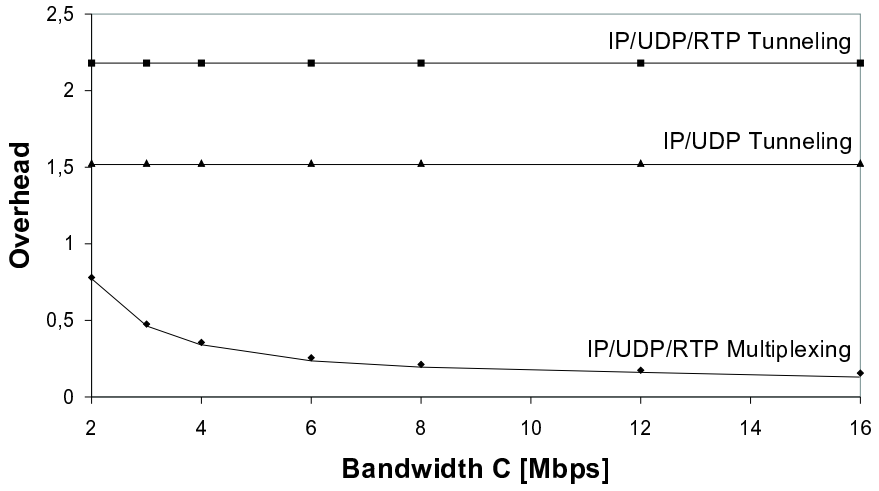

Figure 6: Impact of the bandwidth on the protocol overhead

4.4 Leaky Bucket Description

If the underlying network has considerably more bandwidth than the virtual leased line, IP packets can be sent with variable bit rate immediately after multiplexing without additional spacing and suffer no delay. Then, the timeout value can be set to the maximum allowed delay $T_{max} = D = 1$ msec, thus, increasing the capacity. However, the traffic stream has to be declared by a leaky bucket description. The leaky bucket parameters must be set carefully to avoid loss due to policing by the network.

Using the analysis we can determine the required spacer buffer to carry a certain number of calls over a given bandwidth respecting only the loss QoS criterion. We set the timer value to 1 msec and assume 400 calls to be transmitted over the wired link. In Figure 7 the tradeoff between peak rate and bucket size of a leaky bucket description is illustrated.

The required bucket size is highly dependent on the variance of the interarrival time of the voice packet stream. A minimum bucket size of 800, 1040 and 1550 bytes is required since this is the maximum IP packet size. The throughput is higher for larger bucket sizes since the same traffic can be transported over a lower average bandwidth. 400 calls ($c_{var} = 1$) can be transported over a 4 Mbps link while with spacing a 5 Mbps link is required. However, a bigger bucket size is probably more costly than a smaller one. Therefore, a cost function should be applied to these curves to find the optimum for pricing.

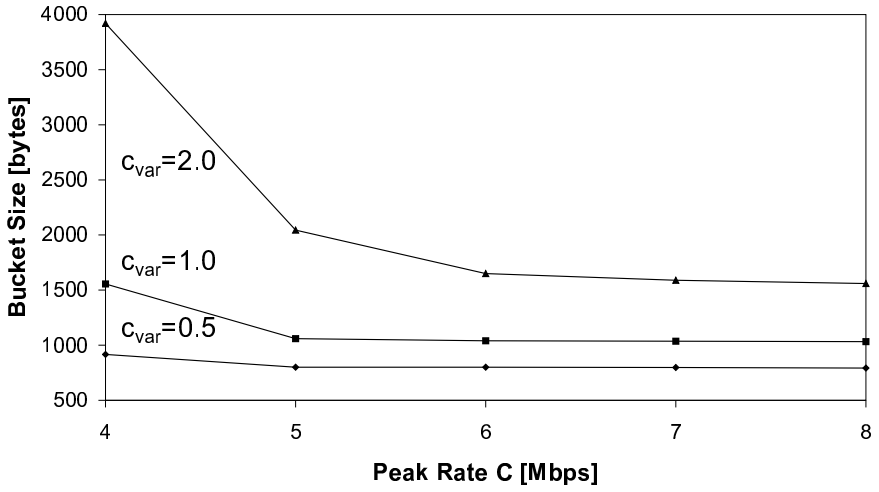

Figure 7: Impact of the peak rate on the leaky bucket size

5 Conclusion

We have investigated the RTP multiplexing protocol suite as it is suggested in [2]. We established a model for transmission of compressed voice data. The affinity to AAL-2 was shown and differences regarding the performance behavior are explained throughout the paper. For accuracy and computability reasons, we developed a discrete time analysis using a framework for solving discrete and finite Markov chains with some new extensions. The loss probability, the overhead, and the waiting time distribution for voice samples were computed.

Using the analysis, several case studies were made to investigate the behavior of RTP multiplexing. A performance tradeoff regarding the timer value could be shown and an explanation was given. A value was found which yielded good performance for various parameter studies. The link capacity, in terms of supportable users, and the required spacer sizes were very sensitive to variance in the interarrival times of the transported voice samples. The influence of the offered load on QoS measures of voice packets was shown. Using RTP multiplexing instead of (RTP/)UDP/IP tunneling only half the bandwidth or even less was needed to carry the same number of connections. For higher link speeds the overhead is only half compared to AAL-2. Finally, a leaky bucket description was presented for the multiplexed traffic. Variable bit rate transmission using a leaky bucket contract yielded a better bandwidth exploitation than constant bit rate transmission using a spacer since there was no spacing time.

From the sensitivity of the results on the variance of interarrival times of voice samples we learned that thorough source modeling is crucial for a reasonable dimensioning of system parameters in an RTP multiplexing system.

Acknowledgements

The author would like to thank M. Dümmler, K. Leibnitz and Prof. Dr.-Ing. P. Tran-Gia for their valuable discussions.

References

[1] 3GPP, "3G TR23.922 version 1.0.0: Architecture for an all IP network," October 1999.

[2] K. El-Khathib, G. Luo, G. Bochmann, and F. Pinjiang, "Multiplexing scheme for RTP flows between access routers <draft-ietf-avt-multiplexing-rtp-00.txt>." http://www.ietf.org/internet-drafts/draft-ietf-avt-multiplexing-rtp-01.txt, October 1999.

[3] N. Gerlich and M. Ritter, "Carrying CDMA traffic over ATM using AAL–2: A performance study," Technical Report, No. 188, University of Wuerzburg, Institut of Computer Science, September 1997.

[4] N. Gerlich and M. Menth, "The performance of AAL-2 carrying CDMA voice traffic," in *11th ITC Specialist Seminar*, (Yokohama, Japan), October 1998.

[5] N. Gerlich, *Transporting Wireless Network Traffic on Wired Networks - A Performance Study*. PhD thesis, University of Wuerzburg, Faculty of Computer Science, Am Hubland, April 1999.

[6] M. Menth and N. Gerlich, "A numerical framework for solving discrete finite markov models applied to the AAL-2 protocol," in *MMB '99, 10th GI/ITG Special Interest Conference*, (Trier), pp. 0163–0172, September 1999.

[7] B. Subbiah, S. Dixit, and N. R. Center, "Low-bit-rate voice and telephony over atm in cellular/mobile networks," *IEEE Personal Communications*, pp. 37–43, Dec 1999.

[8] 3GPP, "TSGR3#7(99)c05: Study item (ARC/3) overall delay budget within the access stratum." status report, September 1999.

[9] TIA/EIA/IS-96A, "Speech service option standard for wideband spread spectrum digital cellular system." Telecommunications Industry Association, 1995.

[10] J. Postel, "RFC791: Internet protocol." http://www.ietf.org/rfc/rfc0768.txt, August 1981.

[11] S. Deering and R. Hinden, "RFC2460: Internet protocol version 6 (IPv6) specification." ftp://ftp.isi.edu/in-notes/rfc2460.txt, December 1998.

[12] J. Postel, "RFC768: User datagram protocol." http://www.ietf.org/rfc/rfc0791.txt, September 1980.

[13] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RFC1889: RTP - a transport protocol for real-time applications." ftp://ftp.isi.edu/in-notes/rfc1889.txt, January 1996.

[14] International Telecommunication Union, *ITU-T Recommendation I.363.2. B-ISDN ATM Adaptation Layer Type 2 specification*, February 1997.

[15] The ATM Forum, *Traffic Management Specification, Version 4.0*, April 1996.

[16] J. Wroclawski, "RFC2210: The use of RSVP with IETF integrated services." ftp://ftp.isi.edu/in-notes/rfc2210.txt, September 1997.

[17] S. Shenker, C. Partridge, and R. Gueria, "RFC2212: Specification of guaranteed quality of service." ftp://ftp.isi.edu/in-notes/rfc2212.txt, September 1997.

[18] S. Blake, D. Black, M. Carlson, S. Davies, Z. Wang, and W. Weiss, "RFC2475: An architecture for differentiated services." ftp://ftp.isi.edu/in-notes/rfc2475.txt, December 1998.

[19] P. Tran-Gia, "Discrete-time analysis technique and application to usage parameter control modeling in ATM systems," in *8th Australian Teletraffic Research Seminar*, (Melbourne), December 1993.