Carrying Wireless Traffic over IP
Using Realtime Transport Protocol Multiplexing

Michael Menth
University of Würzburg, Am Hubland, D-97074 Würzburg, Germany
Tel: (+49) 931-8886644, E-Mail: 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 analysis for the RTP/UDP/IP multiplexing scheme including spacing is derived. Influences on the performance of the system are investigated. An optimum value for the multiplexing timer is found and the effects of the load on the system is shown. A numerical comparison of various voice data multiplexing and tunneling alternatives is presented. The affinity to AAL-2 multiplexing in ATM is pointed out throughout the paper 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 realtime 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 , 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.

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 B gained from an IS-96 vocoder output [3, 9] ($\overline{B} = 18.348$). 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 mean size $\overline{B}$ of a compressed voice sample. Hence, there is unused space in the cell due to the fixed cell length which represents $\frac{53-\overline{B}}{\overline{B}} \cdot 100\% = 189\%$ overhead.

The IP packet size is variable and wasted payload does not exist. However, the IP header size is $H_{IP} = 20$ bytes for the old IP version 4 [10] and even 40 bytes for the new IP version 6[11]. The UDP header has $H_{UDP} = 8$ bytes [12] and the RTP header comprises $H_{RTP} = 12$ bytes [13]. In this work we use the old IP version 4. Hence, the header overhead for voice data transmission over RTP amounts to $\frac{H_{IP}+H_{UDP}+H_{RTP}}{\overline{B}} \cdot 100\% = 218\%$.

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 $H_{mini} = 2$ bytes mini header [2] and multiplexed into an RTP/UDP/IP packet. The delay must be controlled 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 – and by number of 256 allowable users. Note that only one octet is reserved for the channel identifier (CID) both in AAL-2 as in RTP multiplexing.

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.

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.

3 Analysis of RTP Multiplexing

A simulation run to obtain reliable probabilities in the range of 10^{-6} is very time consuming, while the analysis to be derived takes about 2 minutes on a Pentium III processor. This illustrates the advantage of the analytical approach over simulation.

In this section, a discrete-time 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 notations 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 from Section 2 can be divided into two parts. When there is no IP packet and a voice sample arrives, multiplexing is started. The spacer counter just before this time instant is denoted by S^0. The multiplexing time is limited by the timer value TCU (timer common usage). 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 $L(N)$. The final spacer counter is denoted by S^1 and this behaviour is described by Algorithm 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 S^1 is diminished by $I(N) \cdot C$ resulting S^0. This is expressed by Algorithm 2.

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, 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.

```
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

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

```
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

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, UDP, and RTP header per packet plus the mini header for each multiplexed voice sample. Consequently, the IP packet size is given by the conditional random variable

$$L(N = i) = H_{IP} + H_{UDP} + H_{RTP} + \sum_{j=1}^{i}(H_{mini} + B_j). \tag{2}$$

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

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

The loss probability $p_{loss}^{vs} = \overline{N_{lost}}/\overline{N}$ 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.

$$\begin{aligned} \overline{N_{lost}}(N=i,S^0=j) &= p_{loss}^{IP}(N=i,S^0=j)\cdot i \qquad (5)\\ \overline{N}(N=i) &= i \qquad (6) \end{aligned}$$

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 $(H_{IP} + H_{UDP} + H_{RTP} + H_{mini} \cdot i)$ and the voice sample payload size $(L(N = i) = k) - (H_{IP} + H_{UDP} + H_{RTP} + H_{mini} \cdot i)$ is the IP packet size without the protocol header.

$$\begin{aligned} \overline{U}(N=i,S^0=j) &= (1-p_{loss}^{IP}(N=i,S^0=j))\cdot\\ &\quad (H_{IP}+H_{UDP}+H_{RTP}+H_{mini}\cdot i) \qquad (7)\\ \overline{V}(N=i,S^0=j) &= \sum_{(L(N=i)=k)+\max((S^0=j)-TCU\cdot C,0)\leq S_{max}} l(N=i)[k]\cdot\\ &\quad ((L(N=i)=k)-(H_{IP}+H_{UDP}+H_{RTP}+H_{mini}\cdot i)) \qquad (8) \end{aligned}$$

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

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

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

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

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: $H(N)=\begin{cases}H_{IP}+H_{UDP} & \text{for UDP/IP tunneling}\\ H_{IP}+H_{UDP}+H_{RTP} & \text{for RTP/UDP/IP tunneling.}\end{cases}$
- The intermultiplex time equals the interarrival time: $I(N)=A$.
- A multiplexing time is not existent: $M(N)=0$ (constant).

4 Results

First, the QoS criteria are defined and a suited bandwidth is considered for multiplexing voice data. This leads to a slight modification of the multiplexing protocol. Then, parameter studies are conducted keeping the bandwidth fixed. An optimum timer value TCU is found as well as the corresponding required spacer size is evaluated. The behaviour of the QoS parameters is observed varying the load. Finally, a comparison among various types of voice data multiplexing and tunneling is made for different bandwidth.

4.1 QoS Criteria and Time Scale

We define the loss probability of at most 10^{-6} for a voice sample a QoS criterion for voice transmission. According to [8] the delay must not be larger than $D=1$ msec, however, this value is not fixed yet. Since this is very restrictive, we define a second QoS criterion: the probability of a voice sample waiting longer than the delay budget must be at most 10^{-4}. These are the constraints for all data computed in the following section.

The net traffic bandwidth is denoted by $C^* = n\cdot\frac{\overline{B}}{F}$ 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. The number of supportable calls can be computed by $n=\frac{\rho\cdot C\cdot F}{\overline{B}}$.

For the interarrival time (given in TU) 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 $c_{var} = \sqrt{\frac{\overline{A}-1}{\overline{A}}}$. It is about 1 for most of the considered values of ρ. Therefore, parameter studies with $c_{var} = 1$ are relevant for the scenario depicted in Section 2. To investigate the sensitivity of the results, different coefficients of variations must be tested. Therefore, we decide to use a modified negative binomial distribution since its mean and coefficient of variation can be easily controlled.

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 the numerical results depends on the time scale. We set the value for $D = 128$ TU, hence, 128 TU correspond to 1 msec. If we decide one msec to be only 32 TU, the following results must be interpreted with a quarter of the indicated bandwidth and a delay budget of $D = 4$ msec.

4.2 Bandwidth

The first question is which bandwidth is apt to carry information using the suggested multiplexing protocol. One Mbps corresponds to a link speed of 128 $\frac{bytes}{msec}$. It is too little to carry every millisecond a multiplexed IP packet coming from a voice traffic stream with a coefficient of variation of $c_{var} = 1.0$ since the delay criterion can not be met. Two Mbps are not sufficient for $c_{var} = 2.0$. On the other side, with 3 Mbps 259 users could be carried if $c_{var} = 0.5$, however, the protocol allows only for 256 users. Hence, the proposed protocol is not fit for a wide field of applications. Therefore, we propose to reserve 2 octets for the CID value. The following investigations are performed using this proposal.

4.3 Optimum Timer Value

We investigate the influence of the timer value TCU on the multiplexing performance using 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 (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 a required minimum that still meets the delay constraint. For more variable interarrival times a larger spacer size is needed than for small ones (Figure 1) although the critical load, i. e., the number of allowable users or the amount of transmitted voice samples, is smaller as can be seen in Figure 2. 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. 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.

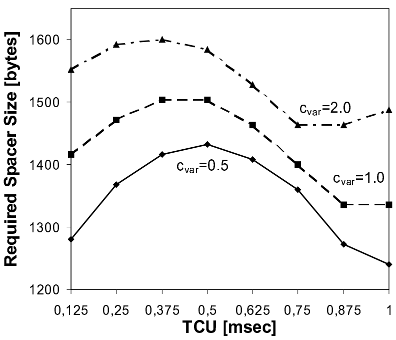

Figure 1: Impact of timer setting on the required spacer size

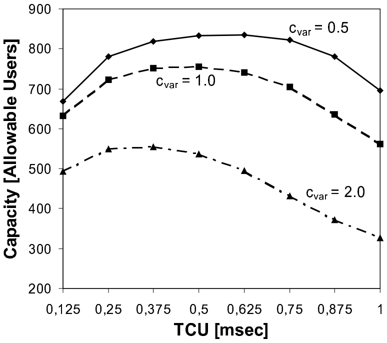

Figure 2: Impact of timer setting on the capacity

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 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 is not sensitive to the chosen bandwidth nor to the variance of a realistic voice sample size distribution ($0.35 \leq c_{var} \leq 0.5$). However, for a coefficient of variation of $c_{var} = 2.0$, which is unrealistically high 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 and AAL-2 multiplexing. Under full load the timer limits the AAL-2 multiplexing time only rarely because the small ATM cells are filled before a timeout occurs [3, 7]. 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.

4.4 QoS Behavior

Again, we consider an 8 Mbps link with a minimum spacer size of 1504 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. Therefore, the waiting time in a very low loaded system is longer than in a fully loaded system. Figure 3 also shows the quantile for the delay budget of the waiting time distribution as well as the loss probability. Their growth is of exponential order related to the number of supported users. Both exceed the QoS criterion at the critical load of around 0.676 since the spacer buffer has been minimized.

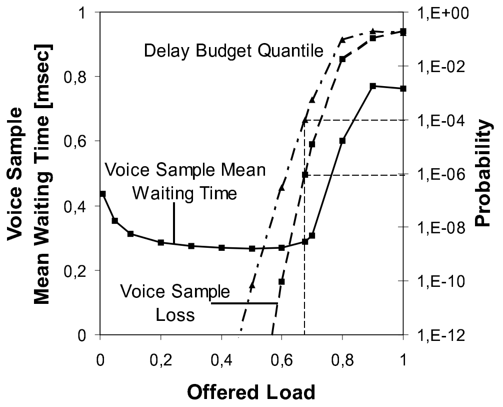

Figure 3: Impact of the offered load on the voice sample mean waiting time, the delay budget quantile, and the loss probability

4.5 Comparison of Multiplexing and Tunneling Alternatives

As mentioned before, tunneling can be performed by using RTP but also by the mere UDP/IP protocol suite. On the other side, for multiplexing in IP networks, one can think of different options. The existing draft with the 2 bytes large mini header (Mux-2-12) has several drawbacks. The timestamp and the sequence number can only be guaranteed if they are carried in the mini header (Mux-12-12) blowing it up to 12 bytes. The other drawback is the small CID which can be overcome by reserving one more octet (Mux-3-12). Finally, on can think of a multiplexing protocol consisting only of the mini headers (Mux-3-0) since the information from the RTP header is not reliable anyway on a lossy link.

The performance of a multiplexing protocol highly depends on the traffic to be carried. The higher the bandwidth and the offered load, the smaller are the interarrival times. This increases the number of multiplexed voice samples in an IP packet and reduces the proportion of the overhead as can be seen in Figure 5. Hence, the critical load increases more than by the raw multiplexing gain (Figure 4), while for tunneling the overhead stays the same.

Both Mux-3-0 and Mux-3-12 support an offered load of up to 0.77 whereas (RTP/)UDP/IP tunneling only sustains 0.28 and 0.37, 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 Mux-3-0 and Mux-3-12 a bandwidth of only 6.7 Mbps is required. The existent multiplexing protocol can only be used for a bandwidth up to 3.5 Mbps (MUX-3-12) and 5 Mbps (MUX-12-12), since the number of multiplexed users exceeds 256, otherwise. Mux-12-12 does not reveal a great advantage over tunneling.

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 multiplexing in IP networks at a bandwidth of 6 Mbps, and for 16 Mbps the overhead is only 20%.

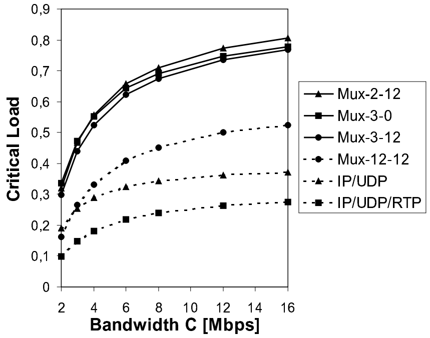

Figure 4: Impact of the bandwidth on the critical load

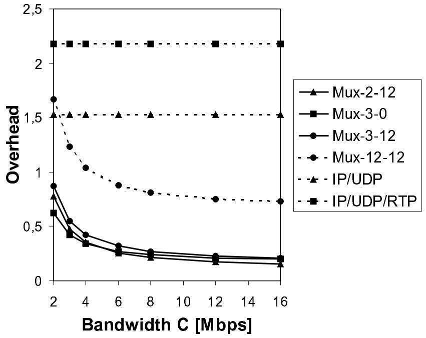

Figure 5: Impact of the bandwidth on the protocol overhead

5 Conclusion

We established a model for the transmission of compressed voice samples in UMTS using RTP multiplexing. The affinity to AAL-2 was shown and differences regarding the performance behavior were 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. First, we modified the existing protocol draft [2] with a larger connection ID to allow more than 256 calls for multiplexing. This seems to be necessary given the tight QoS requirements in [8]. A performance tradeoff regarding the timer value could be shown and explained. This resulted into an optimum timer value 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. Hence, thorough source modeling is crucial for a reasonable parameter setting in an RTP multiplexing system. Furthermore, the influence of the offered load on the QoS parameters was shown. A performance comparison of various multiplexing and tunneling alternatives was presented. For multiplexing only half or even less the bandwidth was needed to carry the same number of connections as with tunneling. For very high link speeds, the overhead is only two thirds compared to AAL-2.

Further studies should address the traffic transport in UMTS networks in the presence of voice, circuit and packet switched data. It is not clear yet, how the different QoS criteria can be met.

Acknowledgements

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

References

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

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

[3] N. Gerlich and M. Ritter, "Carrying CDMA traffic over ATM using AAL–2: A performance study," Tech. Report 188, Univ. of Wuerzburg, Inst. of Comp. Science, Sep. 1997.

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

[5] N. Gerlich, *Transporting Wireless Network Traffic on Wired Networks - A Performance Study*. PhD thesis, University of Wuerzburg, Faculty of Computer Science, Apr. 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. 163–172, Sep. 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, Sep. 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, Aug. 1981.

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

[12] J. Postel, "RFC768: User datagram protocol.", Sep. 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, Jan. 1996.

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

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

[16] J. Wroclawski, "RFC2210: The use of RSVP with IETF integrated services.", Sep. 1997.

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

[18] S. Blake, D. Black, M. Carlson, S. Davies, Z. Wang, and W. Weiss, "RFC2475: An architecture for differentiated services.", Dec. 1998.

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