On the Efficiency of Voice over Integrated Services using Guaranteed Service

Maarten Büchli, Danny De Vleeschauwer, Jan Janssen,
Annelies Van Moffaert and Guido H. Petit

Alcatel Bell, Network Strategy Group
Francis Wellesplein 1
B-2018 Antwerp, Belgium
Tel.: +32 3 2407081 Fax: +32 3 2404888
E-mail: maarten.buchli@alcatel.be

***Abstract*—This paper presents an efficiency study of voice over Integrated Services (IntServ). In particular, the Guaranteed Service class is considered. This service class provides a deterministic upper bound on the end-to-end queuing delay. A method to calculate the optimal packetization delay, and hence, the optimal packet size, is presented. Choosing this packet size involves a trade-off between bandwidth efficiency and delay. Two scenarios are considered in this paper: an IP-phone-to-IP-phone and a gateway-to-gateway scenario. For the latter scenario two multiplexing approaches are evaluated and it is shown that they achieve approximately equal bandwidth efficiency. In addition our results demonstrate that with aggregated voice flows on one reserved bit pipe (gateway-to-gateway scenario) high bandwidth efficiency can be achieved.**

***Index terms*-- IntServ, Voice over IP, efficiency**

A. INTRODUCTION

1st. Background

Introducing telephony services on IP networks brings its own challenges with respect to voice quality, call set-up time and reliability. The performance of a VoIP network should be comparable to the current PSTN. Especially, the voice quality is of great concern. It depends on many parameters: on the application layer the type of codec, packetization and dejittering delay and on the transport layer the one-way delay, jitter and packet loss. The Quality of Service (QoS) of packet switched networks (e.g. IntServ) controls the transport parameters. The requirements for these parameters are requested by the voice application such that together with the application parameters a certain desired speech quality is obtained. By offering different classes of speech quality, an operator is able to offer telephony services at different prices, targeting different market segments.

2nd. Overview of previous work

Telephony has very stringent delay requirements. When perfect echo control is applied, the mouth-to-ear delay should not exceed 150 ms in order to obtain traditional PSTN quality [10]. Obtaining a small delay comes often at the expense of the bandwidth efficiency (i.e. ratio between the codec rate and the bit rate that has to be reserved). The size of the header of an IP packet is relatively large, often resulting in poor efficiency. The overhead of an IP packet consists of an IP/UDP/RTP header with a total size of 40 bytes. Header compression is not considered since it may only be used at a certain link on the path and not end-to-end. In [1], [2] the efficiency of telephony over packet networks with deterministic queuing delay guarantees was studied for the IP-phone-to-IP-phone scenario. It was shown that the bandwidth efficiency was quite poor. Therefore, it was suggested to make an overreservation to decrease the maximum queuing delay, thus allowing for an increase in packetization delay, and hence, resulting in relatively less header overhead. The excess of the reserved bandwidth can be consumed by best-effort traffic.

This paper extends these results with several contributions. First, for the calculations of the optimal packetization delay the (static) dejittering delay is also taken into account. Second, the bandwidth efficiency is studied for different types of codecs and third, the gateway-to-gateway scenario is considered. In other words, the scenario where a single reservation is made for an aggregate of voice flows.

In this paper we first consider the IP-phone-to-IP-phone scenario with regard to the optimal packet size when a static dejittering delay is used and the bandwidth efficiency when using different types of codecs. Then, we present a study of the gateway-to-gateway scenario. In this case one bit pipe is reserved between two gateways to transport multiple calls. Two methods are considered to multiplex voice flows into a bit pipe. One method is to multiplex IP packets from different flows and the second one is to multiplex voice frames from different voice flows into a single IP packet.

3rd. Contents

In Section B the different components of the mouth-to-ear delay are listed. How to specify the traffic parameters that describe a voice flow is shown in Section C. The IntServ architecture is discussed in Section D. The calculation of the optimal packetization delay for an IntServ network with Weighted Fair Queuing (WFQ) schedulers is described in Section E. The efficiency corresponding to this optimal packetization delay is evaluated in Section F. The paper concludes with Section G.

B. MOUTH-TO-EAR DELAY

An important parameter for interactive voice communications is the mouth-to-ear (M2E) delay. This is the time between the moment the sending party has spoken a word and the moment it is heard at the receiving party. The M2E delay consists of a deterministic and a stochastic part. The deterministic part consists of packetization (T_{pack}), serialization (T_{ser}), propagation (T_{prop}), dejittering ($T_{dejitter}$) and other (encoding, decoding etc.) delay (T_{oth}). The stochastic part consists of the queuing delay in each node. Although the queuing delay is a stochastic quantity, we are only interested in the maximum queuing delay, i.e. the delay of the slowest possible packet, because if this one arrives in time for play-out, all others do too. For this maximum queuing delay the absolute maximum or a reasonable quantile (for instance the 99% quantile) can be chosen. Taking all this into account, the mouth-to-ear delay can be written as

$$\hat{T}_{m2e} = \hat{T}_{queue} + T_{dejitter} + T_{pack} + T_{prop} + T_{ser} + T_{oth} \quad (1)$$

The Guaranteed Service [12] controls the maximum queuing delay and provides a deterministic upper bound for the traffic that fits within the traffic envelop that is specified in the traffic contract. Hence, the other delay components are not under the control of Guaranteed Service. The bound is worst-case, and therefore, not tight. The actual queuing delay observed in the network will usually be (much) smaller. An example of the delay distribution is shown in Figure 1. The tail of the delay is caused by the stochastic queuing delay.

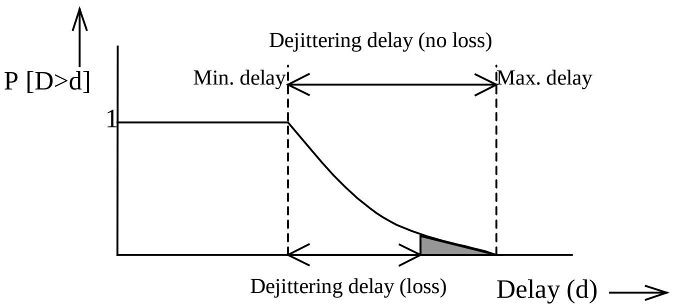

Figure 1: Example of delay distribution

Here the dejittering delay is chosen equal to the maximum queuing delay in order to prevent packets arriving too late for play-out. This is the delay denoted by the arrow with caption ‘Dejittering delay (no loss)’. Since the receiver knows this value it is easy to set the dejittering delay. When a certain packet loss is allowed the dejittering delay can be chosen smaller, for instance equal to the delay denoted by the arrow ‘Dejittering delay (loss)’. In this case, the packets with a delay within the gray colored area may be lost (depending on the queuing delay of the first packet) because they arrived too late for play-out. However, to estimate the resulting packet loss one has to know different quantiles of the queuing delay, i.e. the shape of the delay distribution. This information is not available in IntServ networks.

The analysis presented in this paper is worst case, i.e. the dejittering delay is chosen equal to the maximum queuing delay. Adaptive dejittering algorithms (see e.g. [11]) also exist, they estimate the queuing delay of the first packet. When perfect dejittering is used the queuing and dejittering delay of each packet is equal to the maximum queuing delay. However, it is questionable whether adaptive dejittering algorithms can converge fast enough during the typical length of a phone call.

C. TRAFFIC SPECIFICATION

In IntServ networks reservations can be made with the signaling protocol RSVP [5]. Such reservations are soft state, i.e. they have to be updated regularly otherwise the reservation expires. Other methods to establish reservations are also possible, for instance by SNMP [12]. When SNMP is used the reservation is static, it will only be torn down when it is explicitly instructed to do so.

When RSVP is used the sender sends a PATH message that includes (amongst others) a traffic specification, which consists of five parameters: peak rate (p), token rate (r), bucket depth (b), maximum packet size (M) and the minimum policed unit (m). For voice flows these parameters can be calculated as a function of the bit rate of the codec (R_{cod}), the packetization delay in seconds (T_{pack}) and the header size of a packet in bytes (S_{OH}). It is assumed that the voice source does not use voice activity detection (VAD), and hence, $r=p$, and that the voice source has a fixed bit rate codec and a fixed packetization delay.

In that case, the peak rate (and token rate) can be calculated as follows,

$$p = \frac{R_{cod}}{8} + \frac{S_{OH}}{T_{pack}} \qquad \text{[byte/s]} \qquad (2)$$

The maximum packet size M is calculated as

$$M = \frac{T_{pack} \cdot R_{cod}}{8} + S_{OH} \qquad \text{[byte]} \qquad (3)$$

Because the voice source sends packets of a fixed size with fixed inter-packet times (i.e. it is a non-bursty source), the maximum burst size is equal to the maximum packet size M, hence $b=M$. The minimum policed unit m can be chosen arbitrary as long as $m \le M$. The header size S_{OH} (IP/UDP/RTP) is 40 bytes. The codec bit rates R_{cod} together with the granularity are shown in Table 1 for different types of codecs. The codec granularity is the minimum time between two consecutive voice frames at the output of the coder. The packetization delay must always be a multiple of this.

CODEC	Bitrate [kb/s]	Granularity [ms]
G.711	64	0.125
G.726	16 / 24 / 32 / 40	0.125
G.728	12.8 / 16	0.625
G.729	6.4 / 8 / 11.8	10
G.723.1	5.3 / 6.3	30
GSM-FR	13	20

Table 1: Bit rate and granularity for different codecs

D. INTEGRATED SERVICES

The Internet Engineering Task Force (IETF) has proposed two architectures to provide QoS for IP networks: Integrated Services (IntServ) and Differentiated Service (DiffServ). In the DiffServ architecture, Per Hop Behaviors (PHB) are defined [8], [9]. Such a PHB defines the treatment to an aggregate of traffic. The Integrated Services [4] architecture is a per flow model. Three service classes exist in IntServ: Guaranteed Service [12], Controlled Load [15] and Best-Effort. In this paper we only consider the Guaranteed Service since it provides quantitative guarantees with respect to maximum queuing delay and bandwidth. We do not consider DiffServ.

The bandwidth R that has to be reserved and the maximum queuing delay $\hat{T}_{queue}$ are related as follows for a certain traffic specification (p, r, b, M, m) [12],

$$\hat{T}_{queue} = \begin{cases} \frac{(b-M)(p-R)}{R(p-r)} + \frac{M}{R} + \frac{C_{tot}}{R} + D_{tot} & r < R < p \\ \frac{M}{R} + \frac{C_{tot}}{R} + D_{tot} & R \geq p \end{cases} \quad (4)$$

In the formula above, the C_{tot} and D_{tot} are referred to as the rate-dependent and rate-independent error-terms. They capture the difference of a real packet-based scheduler from the ideal (fluid flow) General Processor Sharing (GPS) scheduler [7]. Each IntServ node on the path that supports Guaranteed Service must update the C_{tot} and D_{tot} terms.

Rate-based schedulers, as e.g. WFQ, guarantee a certain bandwidth to a flow. In this case, delay and bandwidth are coupled. The virtual serialization effect is captured by the term C_{tot} and it is increased with M at each node. The fact that the scheduler is non-preemptive is captured by the term D_{tot}. At nodes that use WFQ it is increased with MTU / C_{link}, where C_{link} is the capacity of the link at that node and MTU the maximum transmission unit. For other rate-based schedulers the term D_{tot} can be different. In this paper we assume that WFQ scheduling is used at each node.

Schedulers, as for example Earliest Deadline First (EDF) [6], are delay-based schedulers. They have the property that bandwidth and delay is uncoupled. Therefore, the term C_{tot} is not increased in this case. However, the term D_{tot} is increased with the maximum queuing delay (i.e. the deadline) and the maximum service time of another packet.

E. OPTIMAL PACKETIZATION DELAY

In this section it is shown how to calculate the optimal packetization delay, and hence, the optimal packet size. In [1] it was already shown how to calculate this optimum. However, the dejittering delay was not taken into account. Therefore, the results are only valid when a perfect adaptive dejittering mechanism is used. In this paper we assume that static dejittering is used, i.e. the dejittering delay is assumed to be equal to the maximum queuing delay such that no packets arrive too late for play-out.

We divide the mouth-to-ear delay budget $\hat{T}_{m2e}$ into two parts. The first part consists of the packetization, maximum queuing and dejittering delay. All these terms depend on the packet size M. The queuing delay through eq. (4) and the dejittering delay because it is chosen equal to the maximum queuing delay. The packetization delay depends on M through eq. (3). The second part, which we denote as T_{min}, includes the serialization, propagation and all other fixed delays. When calculating the optimal packet size only the first part is considered (which is equal to $\hat{T}_{m2e} - T_{\min}$). This is done, because the delay components in the first part are related to each other when Guaranteed Service is used.

There exists an optimal packet size when Guaranteed Service is used due to a trade-off between bandwidth efficiency and delay. When high efficiency is desired, the packetization delay should be chosen large. However, from eq. (4) it turns out that the maximum queuing delay will increase with the packet size because $C_{tot}=N_{stag}\cdot M$ (N_{stag} is number of hops) in case of WFQ. On the other hand when a small delay is desired, a small packetization delay has to be chosen. This results in small packets with relative large headers (hence low efficiency) but also in a small maximum queuing delay.

To visualize this trade-off and to show the existence of an optimal packetization delay the bit rate R that has to be reserved is shown in Figure 2. As an example, this figure was made for a scenario with 10 hops, a codec of 32 kb/s and a $\hat{T}_{m2e} - T_{\min}$ of 100ms. The curve 'codec+header' shows the bandwidth of the voice flow including the header overhead as a function of the packetization delay (see eq. (2)). The curve denoted as $R_{\min}$ is the minimum amount of bandwidth that has to be reserved to obtain a maximum queuing delay regardless of the peak rate of the flow (see eq. 4) such that $T_{pack} + \hat{T}_{queue} + T_{dejitter} = \hat{T}_{m2e} - T_{\min}$. The curve 'R' shows the reservation that has to be made in order to reserve enough capacity while adhering to the delay constraint. The point in the figure where the bandwidth that has to be reserved in the network is minimal is the point of the optimal packetization delay. This optimum can be calculated with eq. (5) when WFQ schedulers are used in each node on the path. The optimal packet size can then be calculated with eq. (3). In this optimum the reserved bandwidth R is equal to the peak rate p. Note that

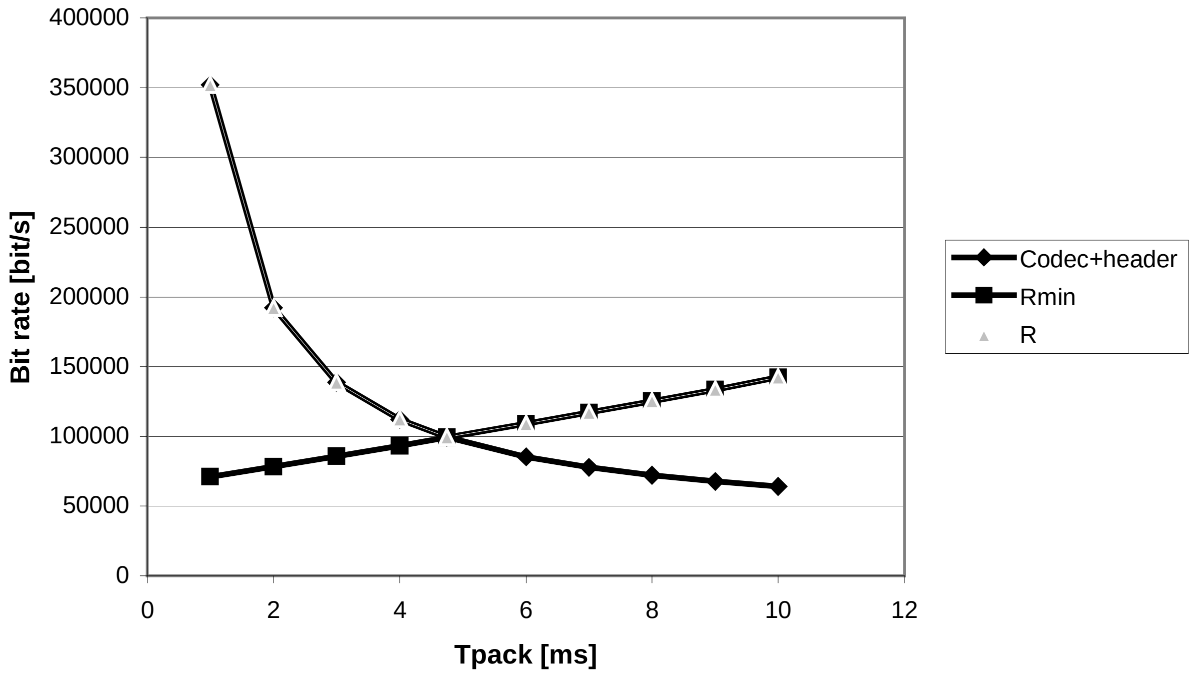

Figure 2: Different bit rates as a function of the packetization delay

the peak rate depends on the packetization delay due to the header overhead.

The optimal packetization delay is given by [2]

$$T_{pack}^{opt} = \frac{\hat{T}_{m2e} - T_{\min} - 2\sum_{i=1}^{N_{stag}} \frac{MTU^i}{C_{link}^i}}{1 + 2N_{stag}} \tag{5}$$

with:

N_{stag} Number of hops
MTU^i Maximum packet size at node *i*
C^i_{link} Link capacity at node *i*
$\hat{T}_{m2e}$ Mouth-to-ear delay budget
T_{min} Serialization, propagation and other delays

To be able to determine this optimal packetization delay the sender should know the values of MTU^i, N_{stag}, T_{min} and C^i_{link}. IntServ capable routers support a set of general control and characterization parameters [13]. These parameters can be obtained by using for instance RSVP. In Table 2 the relevant parameters are listed and explained. These parameters can be used to calculate the optimal packetization delay with eq. (5).

Parameter name	Symbol
NUMBER_OF_IS_HOPS	Nstag
MINIMUM_PATH_LATENCY	Tmin
PATH_MTU	min. MTU on the path
AVAILABLE_PATH_BANDWIDTH	min. Clink on the path

Table 2: General parameters supported by IntServ

F. EFFICIENCY OF GUARANTEED SERVICE

In this section we define the efficiency as the ratio between the codec bit rate R_{cod} and the bit rate R reserved in the network. The efficiency of the guaranteed service is studied for two scenarios. The first scenario is the IP-phone-to-IP-phone scenario, which is depicted in Figure 3. Such a scenario can occur for instance in a corporate network.

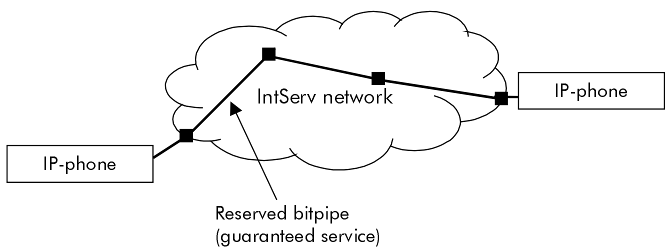

Figure 3: IP-phone-to-IP-phone scenario

Another scenario is an IntServ network where a telecom operator has deployed a number of VoIP gateways. Between all gateways a bit pipe is reserved and all the calls from one gateway to another are aggregated. When the bit pipe is in danger of getting saturated it is possible to either block new calls or to increase the capacity of the bit pipes at the next RSVP reservation update. This scenario is shown in Figure 4.

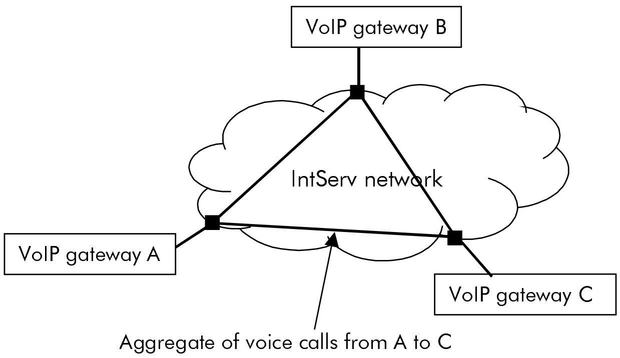

Figure 4: Gateway-to-gateway scenario

1st. IP-phone-to-IP-phone

With eq. (5) the optimal packetization delay can be calculated. In Figure 5 the optimal packetization delay is shown as a function of the number of traversed hops for several delay budgets. The delay budget (shown in the legend) is equal to $\hat{T}_{m2e} - T_{\min}$. Hence, to calculate the mouth-to-ear delay the propagation delay, serialization and other delays (i.e. encoding, etc.) should be added.

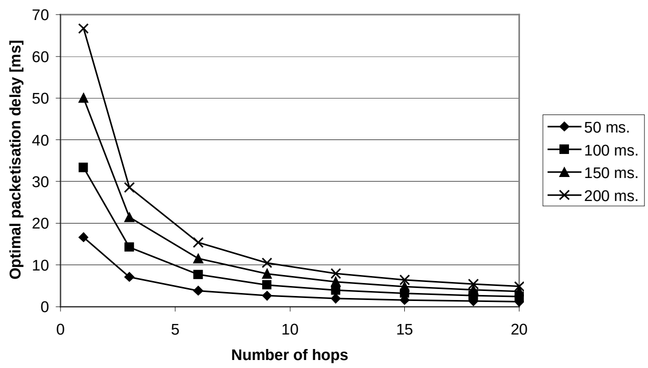

Figure 5: Optimal packetization delays

Figure 5 is only valid for codecs with infinitesimal small granularity. As explained before the packetization delay can only be a multiple of the codec granularity. This phenomenon becomes particularly important when the codec granularity is larger than the optimal packetization delay. In this case, the latter should be chosen equal to the codec granularity. This will result in an overreservation because the packetization delay cannot be chosen optimally. From Figure 5 it can be concluded that the optimal packetization delay is often smaller than the codec granularity (see Table 1). When the optimal packetization is in between two multiples of the codec granularity, the packetization delay should be chosen equal to the closest multiple of the codec granularity.

The peak rate of the voice flow can be calculated from the optimal packetization delay with eq. (2). Because the packetization delay is optimally chosen, the peak rate is equal to the bandwidth that has to be reserved. The efficiency as a function of the number of hops is shown in Figure 6 for a delay budget ($\hat{T}_{m2e} - T_{\min}$) of 100 ms. The bit rates shown in the legend are the codec bit rates R_{cod} (i.e. without IP/UDP/RTP overhead).

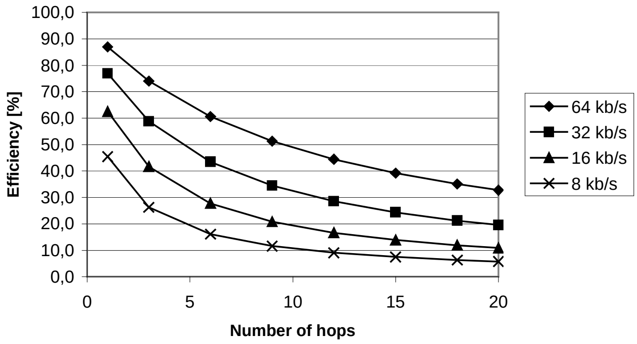

Figure 6: Efficiency for a delay budget of 100 ms

From Figure 6 several remarks can be made. First, a large number of hops on the path results in lower efficiency. This is because the maximum queuing delay accumulates at each hop. Second, the larger the bit rate of the codec, the better the efficiency (but the amount of bandwidth that has to be reserved increases with the bit rate of the codec of course). In other words, the smaller the bit rate of the codec the more dominant the header overhead becomes. Another observation can be made when the efficiency figures are made for different delay budgets (not shown in this paper). The tighter the delay budget the worse the efficiency is. This is because tighter delay budgets result in smaller packetization delays, and hence, smaller packets. Relative small packets result in more overhead due to header bytes.

The efficiency in the case above is quite low. This can be improved by allowing an overreservation in order to reduce the maximum queuing delay. In this case an assumption has to be made about the percentage of voice traffic on the network.

We denote the target efficiency as ε (in eq. (5) ε=1 was used). Eq. (5) now becomes [1]

$$T_{pack}^{opt} = \frac{\hat{T}_{m2e} - T_{\min} - 2\sum_{i=1}^{N_{stag}} \frac{MTU_{\max}^{i}}{C_{link}^{i}}}{1 + 2N_{stag}\varepsilon} \qquad (6)$$

The target efficiency ε has to be chosen equal to the target percentage of voice traffic on the network. The reserved bandwidth that is not used can be consumed by best-effort traffic. In Figure 7 the optimal packetization delays are shown when the voice traffic is targeted at 10% of the network capacity, hence R=10·p. From this figure it can be concluded that by allowing an overreservation the efficiency will improve. This is due to the fact that the queuing delay is decreased by making an overreservation, leaving more delay budget for packetization. Larger packetization delays result in packets with relatively less overhead.

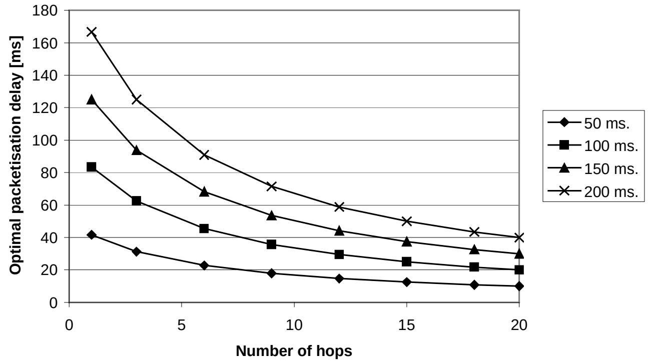

Figure 7: Optimal packetization delay in case of 10% voice load on the network

2nd. Gateway-to-gateway

This section discusses the gateway-to-gateway scenario. In this case several voice flows are aggregated into a single bit pipe. This pipe can be considered as a virtual leased line that is provided by the guaranteed service. When the number of calls is increasing, the gateway can either block new calls or increase the reservation at the next reservation update.

Due to aggregation the processing required for the set-up of reservations at each router will decrease since only one (large) reservation is made instead of a reservation for each individual flow. Also the amount of state information will reduce. Two methods for multiplexing voice flows on one bit pipe are discussed. First, the multiplexing of packets from different flows is discussed, and second, the multiplexing of voice frames into a single packet. Finally, both methods are compared.

1) Multiplexing IP packets

The first approach to multiplex multiple voice calls over one reserved bit pipe is to multiplex the packet flows of the different voice sources. This is shown in Figure 8.

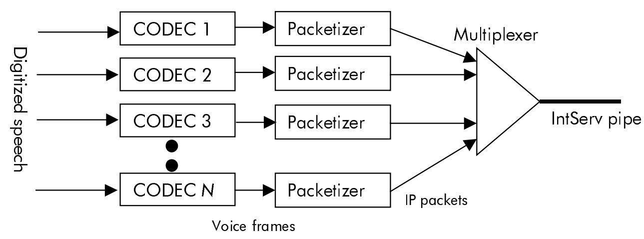

Figure 8: Multiplexing of different packet flows into one bit pipe

In this case it should be taken care of that the multiplexer does not introduce (too much) packet loss and delay. Packet loss and delay due to multiplexing can be avoided completely when the packetizers are clocked in such a way that the packets are produced one after the other. This way no queuing is needed in the multiplexer and no packet loss occurs. This timing is illustrated in Figure 9 when all voice sources use the same type of codec and packetization delay.

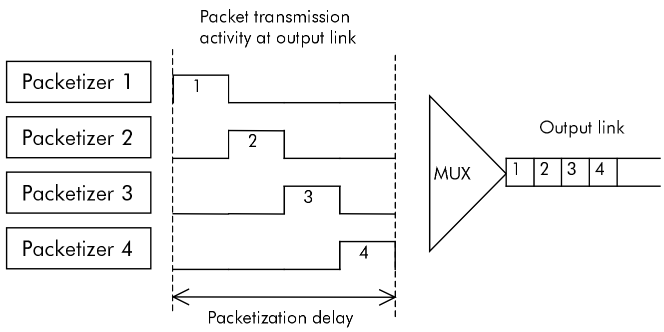

Figure 9: Timing of packetizers

To determine the optimal packetization delay (and hence, the packet size) for an aggregate of identical packet flows eq. (5) has to be modified slightly. Denote N as the maximum number of multiplexed voice flows. Note that taking N as the current number of flows would be more efficient. Notice that this will result in changing the packetization delay of each flow at the set up or release of a flow. Now, the optimal packetization delay is given by

$$T_{pack}^{opt} = \frac{\hat{T}_{m2e} - T_{\min} - T_{MUX} - 2\sum_{i=1}^{N_{stag}} \frac{MTU^i}{C_{link}^i}}{1 + (2N_{stag}\varepsilon / N)} \qquad (7)$$

When the packets are generated as shown in Figure 9, the T_{MUX} (multiplexing delay) term will be zero.

With this multiplexing approach each call uses a different destination port number. Hence, the port number at the receiver is used to associate the packet flow with a certain voice call.

2) Multiplexing voice frames

Another multiplexing method is to multiplex voice frames from different sources into a single IP packet. This is shown in Figure 10.

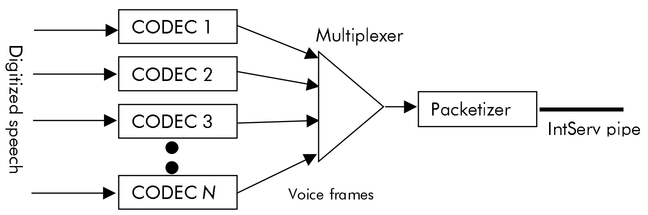

Figure 10: Multiplexing of voice frames into a single packet

For this approach eq. (5) can still be used. The different codecs are modeled as a single source with a rate of $N \cdot R_{cod.}$ The packet bit rate is equal to the codec bit rate plus multiplexing overhead times the number of flows plus the IP/UDP header.

This multiplexing approach needs some additional overhead. This is because the receiving gateway must know which voice frames belong to a certain voice flow/call. Several multiplexing schemes [3], [14] have been proposed. For most schemes an additional overhead of 3 or 4 bytes is introduced per multiplexed flow. However, no standardized solution exists to multiplex RTP flows into one IP/UDP packet yet.

3) Comparison

In this section the efficiency is investigated for the two multiplexing approaches described in the previous two sections. For purposes of comparison it is assumed that ten homogeneous voice flows are multiplexed into one bit pipe. These results are compared with the efficiency of ten separate per-flow reservations.

As an example, we consider the case where no overreservation is allowed, i.e. $\varepsilon = 1$ in eq. (6). The delay budget for packetization, queuing and dejittering is 100 ms. Each voice flow is assumed to use a 32 kb/s codec. Hence, the aggregate flow without overhead has a rate of 320 kb/s. In case of the voice frame multiplexing approach it is assumed that each packet has an IP/UDP header (28 bytes) and 4 bytes of multiplexing overhead per flow. In Figure 11 the efficiency is shown for the two multiplexing approaches and the per flow reservation approach.

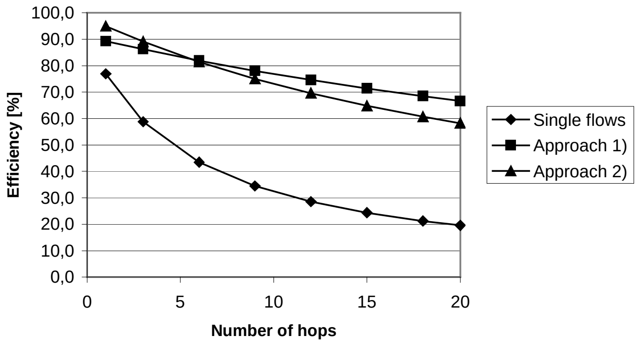

Figure 11: Efficiency for both scenarios

From the results of Figure 11 it is clear that the per-flow reservation is the least efficient and requires the largest bandwidth reservation per voice flow. This is because the maximum queuing delay depends on the packet size and the amount of reserved bandwidth. The bandwidth of a single voice flow is small, and hence, the packet size has to be chosen small in order to meet the maximum queuing delay requirements. When the payload of an IP packet is small, the header overhead is relatively large.

From Figure 11 it turns out that both multiplexing approaches are close with respect to efficiency (and hence, bandwidth reservations). The first approach has IP/UDP/RTP header overhead for each flow. Due to the aggregate reservation the maximum queuing delay of the bit pipe is relatively small such that more delay budget is left for packetization. This results in larger packets, and hence, relatively less overhead. In the second approach all flows are aggregated into a single IP packet. Due to multiplexing schemes some extra per-flow overhead is introduced in each packet for timing and sequencing. However, due to the high aggregate bit rate the packet size becomes large enough to achieve a high efficiency.

In case of approach 1 no multiplexing schemes are needed. The receiver only needs to use the port number to associate the different flows to a voice call. The multiplexing schemes for the second approach are not standardized yet and introduce more complexity.

The optimal packetization delays for the three different approaches are shown in Table 3. Due to the aggregate reservation the maximum queuing delay is small such that a large part of the delay budget is left for packetization in case of approach 1. Because of this large packetization delay the packet size does not become too small (the packet is only filled by a single source). In case of approach 2 the packetization delay is very small but the packet is filled with the aggregate of voice flows.

	Optimal packetization delay [ms]	
Number of hops	Per-flow approach Approach 2	Approach 1
1	33	83
3	14	63
6	8	45
9	5	36
12	4	29
15	3	25
18	3	22
20	2	20

Table 3: Optimal packetization delay

Summarizing, it can be concluded that the per-flow reservations are the least efficient. Therefore, all flows going from one gateway to the other should be aggregated. The efficiency of both multiplexing approaches is close to each other. However, approach 1 has the advantage that individual flows can easily be distinguished by the destination port number as opposed to approach 2 where a multiplexing scheme is needed.

G. CONCLUSIONS

In this paper the Guaranteed Service of IntServ has been studied for two VoIP scenarios: an IP-phone-to-IP-phone and a gateway-to-gateway scenario.

When per-flow reservations are used, the efficiency of small-bandwidth voice flows with stringent delay requirements is low. This results in a network transporting voice traffic that consists for a large percentage of header overhead.

It was also observed that the smaller the bit rate of the codec the more dominant the header overhead becomes. The efficiency can be improved by making an overreservation to achieve a small maximum queuing delay such that the packetization delay can be increased. A larger packetization delay results in larger packets, and hence, relatively less overhead. The overreservation will result in reserved but unused bandwidth. However, this excess bandwidth can be used by best-effort traffic.

Another method to improve the efficiency is to reserve high bit rate IntServ pipes. This is only possible in the gateway-to-gateway scenario. Because of the high bandwidth the maximum queuing delay in the WFQ nodes on the path is small. Therefore, more delay budget is left for packetization. Such a large IntServ pipe can be used to multiplex multiple voice flows. Two multiplexing approaches were evaluated. One approach is to multiplex packets from different voice sources and the second approach is to multiplex voice frames from different sources on a bit pipe. Both multiplexing methods achieve approximately the same efficiency. However, when the first approach is used the voice flows can easily be distinguished using the destination port number. For the second approach a multiplexing scheme is needed. Such a scheme is not yet standardized and it introduces additional complexity. Therefore, multiplexing of packets (approach 1) is preferred.

In our future work we will include DiffServ networks into the calculation. From the viewpoint of an end-to-end IntServ connection a DiffServ cloud is considered as a single IntServ hop that also has to update the C_{tot} and D_{tot} terms appropriately. Hence eq. (5) can be modified to take DiffServ networks into account, if the updating mechanism for C_{tot} and D_{tot} is known.

H. ACKNOWLEDGEMENTS

This work was carried out within the framework of the project LIMSON, sponsored by the Flemish institute for the promotion of scientific and technological research in the industry (IWT).

I. REFERENCES

[1] Mario Baldi, Davide Bergamasco and Fulvio Risso, “On the Efficiency of Packet Telephony”, 7th IFIP Int’l Conference Telecommunication Systems, March 1999.

[2] Mario Baldi and Fulvio Risso, “Efficiency of Packet Voice with Deterministic Delay”, IEEE Communications Magazine, pp. 170-177, May 2000.

[3] Mooi C. Chuah, Enrique J. Hernandez-Valencia, “A LightWeight IP Encapsulation (LIPE) Scheme”, IETF draft <draft-chuah-avt-lipe-00.txt>, June 2000, work in progress.

[4] R. Braden, D. Clark and S. Shenker: IETF Request for Comment 1633, http://www.ietf.org/rfc/rfc1633.txt, “Integrated Services in the Internet Architecture: an Overview”, June 1994.

[5] R. Braden, L. Zhang, S. Berson, S. Herzog and S. Jamin: IETF Request for Comment 2205, http://www.ietf.org/rfc/rfc2205.txt, “Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification”, September 1997.

[6] Fabio M. Chiussi and Vijay Sivaraman, “Achieving High Utilization in Guaranteed Services Networks Using Early-Deadline-First Scheduling”, IWQOS ’98, pp. 209-217, Napa, California, May 1998.

[7] R. Guérin and V. Peris, “Quality-of-Service in packet networks: basic mechanisms and directions”, Computer Networks 31, pp. 169-189, 1999.

[8] J. Heinanen, F. Baker, W. Weiss and J. Wroclawski: IETF Request for Comment 2597, http://www.ietf.org/rfc/rfc2597.txt, “Assured Forwarding PHB group”, June 1999.

[9] V. Jacobson, K. Nichols and K. Poduri: IETF Request for Comment 2598, http://www.ietf.org/rfc/rfc2598.txt, “An Expedited Forwarding PHB”, June 1999.

[10] J. Janssen, D. De Vleeschauwer, G.H. Petit, “Delay and Distortion Bounds for Packetized Voice Calls of Traditional PSTN Quality”, Proceedings of the 1st IPTEL workshop, GMD report 95, pp. 105-110, Berlin, April 2000.

[11] S.B. Moon, J. Kurose, D. Towsley, “Packet audio play-out adjustment: performance bounds and algorithms”, ACM/Springer Multimedia Systems, Vol. 6, pp. 17-28, January 1998.

[12] S. Shenker, C. Partridge and R. Guerin: IETF Request for Comment 2212, http://www.ietf.org/rfc/rfc2212.txt, “Specification of Guaranteed Quality of Service”, September 1997.

[13] S. Shenker, J. Wroclawski: IETF Request for Comment 2215, http://www.ietf.org/rfc/rfc2215.txt, “General Characterization Parameters for Integrated Service Network Elements”, September 1997.

[14] Keiko Tanigawa, Tohru Hoshi, Koji Tsukada: “Simple RTP Multiplexing Transfer Methods for VoIP”, IETF draft <draft-tanigawa-rtp-multiplex-01.txt>, November 1998, (Expired).

[15] J. Wroclawski: IETF Request for Comment 2211, http://www.ietf.org/rfc/rfc2211.txt, “Specification of the Controlled-Load Network Element Service”, September 1997.