Comparing the Efficiency of IP and ATM Telephony

Mario Baldi and Fulvio Risso
Dipartimento di Automatica e Informatica, Politecnico di Torino
Corso Duca degli Abruzzi, 24, 10129 Torino - Italy
phone +39 011 564 7067, fax +39 011 564 7099
{mbaldi, risso}@polito.it

Abstract— **Circuit switching, suited to providing real-time services due to the low and fixed switching delay, is not cost effective for building integrated services networks bursty data traffic because it is based on static allocation of resources which is not efficient with bursty data traffic. Moreover, since current circuit switching technologies handle flows at rates which are integer multiples of 64 Kb/s, low bit rate voice encodings cannot be taken advantage of without aggregating multiple phone calls on a single channel.**

This work explores the *real-time efficiency* of IP telephony, i.e. the volume of voice traffic with *deterministically* guaranteed quality related to the amount of network resources used. IP and ATM are taken into consideration as packet switching technology for carrying compressed voice and it is compared to circuit switching carrying PCM (64 Kb/s) encoded voice. ADPCM32 is the voice encoding scheme throughout most of the paper. The impact of several network parameters, among which the number of hops traversed by a call, on the real-time efficiency is studied.

I. Introduction

Circuit switching is particularly suitable to provide real-time services, like video and telephony, because of its low and fixed switching delays. However, it is based on static allocation of resources which is not cost effective for bursty data traffic. Moreover, current circuit switching technologies handle flows at rates which are integer multiples of 64 Kb/s; this prevents from taking advantage of low bit rate voice encodings, unless multiple phone calls are aggregated in a single flow significantly increasing the complexity of the network and of call handling.

Packet switching is appealing for carrying real-time traffic because it can benefit from (possibly variable bit rate) compression schemes and statistical multiplexing to more efficiently exploit network resources. This paper presents a study on the *efficiency* of packet switching in providing *toll quality* telephone services.

This work explores the *real-time efficiency* of IP telephony, i.e. the volume of voice traffic with *deterministically* guaranteed quality related to the amount of network resources used. IP and ATM are taken into consideration as packet switching technology for carrying compressed voice and it is compared to circuit switching carrying PCM (64 Kb/s) encoded voice. ADPCM32 is the voice encoding scheme throughout most of the paper.

Provision of Quality of Service (QoS) guarantees over packet switched networks requires two extremely important components:

1. Packet scheduling algorithms into nodes that are more effective in controlling buffering delay variation than First In First Out (FIFO) queueing;
2. Call Admission Control (CAC) in order to control the amount of real-time traffic having access to the network and to reserve resources to real-time flows.

These two components are strictly related since the amount of resources to be reserved to a real-time flow—thus the amount of real-time traffic acceptable on the network—depends on the scheduling algorithm deployed.

Whenever a new phone conversation is to be started, the needed QoS is required to the network through some sort of signalling protocol. On ATM networks signalling is performed through UNI signalling [1]; on IP networks an analogous signalling protocol could be based, for example, on the Resource ReserVation Protocol (RSVP) [2]. The QoS objective for a toll quality phone call is a deterministic bound of about 200 ms on the round-trip delay perceived by users.

This work is based on a call level simulator [3] which assumes that the Packet-by-Packet Generalised Processor Sharing (PGPS) [4], [5] scheduling algorithm is used into network nodes. Section II discusses how CAC is performed when PGPS is used to manage queues in network nodes. Indexes used throughout the paper to evaluate the efficiency in utilisation of network resources and the main factors affecting them are introduced in Section III. Section IV presents the problems related to the packetization process. Section V shows the behaviour of a packet switched network with an high number of network nodes. Section VI shows the results obtained with different resource allocation criteria. A brief discussion of the results is drawn in Section VII.

II. Call Admission Control

PGPS is derived from the Generalised Processor Sharing (GPS) algorithm which assumes the *fluid flow* model of traffic: each active flow feeds a separate buffer and all the backlogged buffers are served concurrently. A GPS scheduler guarantees to each flow i a minimum service rate

$$g_i = \frac{\phi_i}{\sum_j \phi_j} \cdot r, \tag{1}$$

where r is the output rate, usually the output link capacity, and $\frac{\phi_i}{\sum_j \phi_j}$ represents the fraction of the link capacity *reserved* to flow i.

Provided that a flow is compliant with the traffic exiting a leaky bucket with an output rate $B_i < g_i$ and depth σ_i, GPS guarantees an upper bound on the queueing delay of each flow i $Q_i^{GPS} = \sigma_i / g_i$.

PGPS, also developed by Demer, Keshav and Shenkar under the name of *Weighted Fair Queueing* [6], extends GPS in order to handle packet-based flows. The basic idea behind PGPS is quite straightforward: incoming packets are scheduled for transmission according to their equivalent GPS service time, i.e. the instant of time in which the last bit of a packet would be sent by GPS.

Assuming that a packet flow is compliant with the above leaky bucket (i.e. leak rate B_i and bucket depth σ_i), the queueing delay bound is (Equation 12.1 in [7])

$$D_i = \frac{\sigma_i}{g_i} + \frac{(h_i - 1) \cdot L_i}{g_i} + \sum_{m=1}^{h_i} \frac{L_{max}}{r_m} \quad (2)$$

where h_i is the number of hops on the path of flow i, r_m is the service rate of the m^{th} node (usually the capacity of link m), L_i is the maximum packet size for flow i and L_{max} is the maximum packet size allowed in the network.

The delay bound provided by Equation 2 is basically proportional to the burstiness of the source σ_i, the number of traversed nodes h_i, the maximum packet length of the session itself (L_i) and of the network (L_{max}). It is inversely proportional to the weight ϕ_i associated with that source and the links bandwidth r_m.

The queueing delay, i.e. Equation 2, is only a component of the overall end-to-end delay. The CAC is provided with a delay requirement D_{req} which is the network delay budget for the call obtained by subtracting from the delay acceptable by the user both the time needed for application level processing (i.e. audio or video compression) and the protocol processing time, not including the delay introduced by the packetization process. The CAC uses the following inequality to determine the amount of network resources needed to guarantee the required QoS to a flow and decide whether to accept it or not:

$$D_{req} \geq Dpack + Dprop_0 + \frac{\sigma_i + (h_i - 1) \cdot L_i}{g_i} + \\ + \sum_{m=1}^{h_i} \left(\frac{L_{max}}{r_m} + Dprop_m\right) \quad (3)$$

The inequality takes into consideration the propagation delay $Dprop_m$ on the m^{th} link of the path and the packetization delay $Dpack$.

The CAC checks whether each link on the call path has an amount of available (i.e. not yet reserved) bandwidth larger than $\max(\rho_i, g_i^*)$, where ρ_i is the bandwidth required for the transmission of the i^{th} flow and g_i^* is the minimum g_i value that satisfies Inequality 3. If enough bandwidth is available, the appropriate amount is reserved to the call on every link traversed. When the amount of bandwidth g_i^* needed to meet the QoS requirement of a flow is larger than the amount ρ_i required to transmit the flow i including protocol overheads, we say that *bandwidth over-allocation* is performed. When a call is torn down, the bandwidth previously reserved to it is released.

Bandwidth over-allocation, effective with the PGPS scheduler, can be less useful with other scheduler mechanisms. Particularly, there exists schedulers in which effectiveness depends on other parameters (e.g., packet size is crucial with Weighted Round Robin and Class Based Queueing), others that are specifically designed to provide guaranteed delay (e.g., Stop and Go queueing [8] or Jitter-EDD) and do not require over-allocation at all, and others that decouple bandwidth allocation and delay (Hierarchical Fair Service Curve [9], for example).

III. Efficiency of Guaranteed Services over Packet Networks

This study uses the following set of three efficiency indexes [3] that can be used to compare the efficiency of packet switching and circuit switching.

1. The *effective load* is the data rate at the application level and gives an idea of the amount of real-time traffic carried by the network. The effective load does not account for the protocol overhead, i.e. it is the capacity that should be required to send the data on a circuit switched network.
2. The *real load* is the raw link capacity used by user data; it corresponds to the effective load augmented by the overhead introduced by the various protocol layers.
3. The *apparent load* is the bandwidth reserved to the phone calls (more in general to the real-time sessions) in order to meet their QoS requirements[1].

These indexes provide a measure of how effectively calls with real-time guarantees can be carried by the network. For example, the lower the apparent bandwidth of a call, the higher is the amount of such calls the network can carry; the larger the real bandwidth, the higher is the amount of raw transmission capacity required.

Considering a given amount of network resources, *efficiency* can be viewed from two different perspectives:

1. *Real-time efficiency* takes into account the amount of real-time traffic carried by the network and it is relevant when the network is intended to carry mainly real-time traffic, like a commercial telephone network. The real-time efficiency can be measured in comparison with circuit switching as B_{CS}/B_{app}, where B_{app} is the apparent bandwidth of a call and B_{CS} is the bandwidth required to transmit a voice call over a circuit switched network, i.e. 64 Kb/s [2].
2. *Transport efficiency* refers to the overall amount of traffic carried by the network and is relevant when a significant part of the traffic has to be best effort and the provision of the corresponding service is not a marginal issue. The transport efficiency can be measured as

[1] When referring to a single call instead of the overall network occupancy, the term “bandwidth” is used instead of “load”.

[2] 64 Kb/s is the capacity granularity for channel allocation on a circuit switched network.

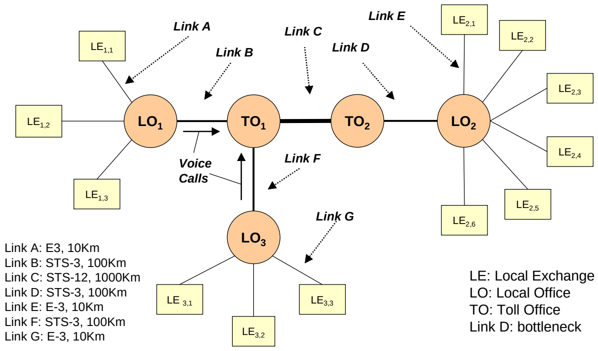

Fig. 1. Network topology used in the simulation.

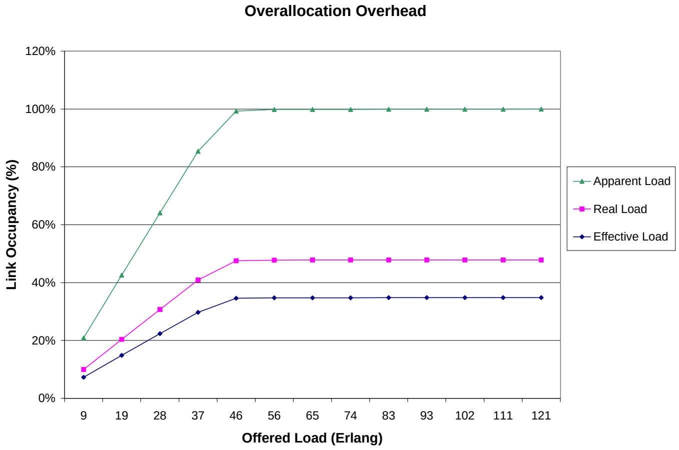

Fig. 2. Voice over IP: efficiency indexes on link D.

B_{CS}/B_{real}, where B_{real} is the real bandwidth of a call.

This paper reports the results of a simulation study on the network shown in Fig. 1; the topology has been designed after the one of a domestic telephone network.

Fig. 2 shows the effective, real and apparent load on link D as a percentage of the link capacity[3]. Voice samples are carried into IP packets with the usual encapsulation for real-time traffic which features UDP and RTP at the higher layers. IP packets are encapsulated in PPP frames trasmitted directly over SONET/SDH links. The packet payload size has been chosen to be 128 bytes, which leads to a packetization delay of 32 ms.

In the leftmost part of the plot, the three loads increase linearly as the traffic offered to the network increases. This means that all the calls are accepted. When the offered traffic becomes large enough to saturate the bottleneck link (i.e. the apparent load reaches 100% of the bottleneck link capacity), the three load curves flatten, indicating that part of the incoming calls are rejected by the CAC. The flat part of the curves represents the maximum link utilisation achievable in this scenario.

The difference between the apparent load and the real load curves is the *bandwidth over-allocation* performed by the CAC. However this over-allocated bandwidth is not really wasted since it can be used to transmit best effort traffic which has not delay requirements.

The effective load represents the fraction of link bandwidth that circuit switching would require to carry the same number of phone calls accepted by the packet switched network. Thus, effective load enables the comparison between the packet switched telephone network and the circuit switched one from the efficiency standpoint.

The difference between the real load and the effective load curves represents the amount of bandwidth wasted to carry the protocol overhead, i.e., packet headers. This waste is unavoidable and can be considered as the fee to be paid in order to benefit from the advantages of packet switching (voice compression, real-time and best effort traffic multiplexing, and deployment of "inexpensive" packet switching equipment in place of "costly" circuit switching devices).

The difference between the apparent load and the effective load curves shows how the resource allocation relates to the amount of information (voice samples) carried. For example, Fig. 2 shows that approximately only the 35% of the network capacity is actually used to carry ADPCM32 voice calls.

The bandwidth over-allocation plays a key role since, as shown by Fig. 2, it can have a significantly stronger impact on real-time efficiency than protocol overhead. Bandwidth over-allocation and protocol overhead are tightly coupled, as shown in the next section.

IV. Header and Packet Size

Packet size affects real-time efficiency since a large packet requires long packetization delay and more bandwidth might to be allocated in order to meet the QoS requirement on the end-to-end delay.

The header size depends on the protocol architecture deployed in the network. This section studies the effect of varying the packetization delay with different protocol architectures.

Fig. 3 shows a plot of the real bandwidth required by an ADPCM32 phone call versus the size of the packet payload (i.e., the packetization delay) for different network technologies. The real bandwidth required on a circuit switched network by both an ADPCM32 (that is equal to the effective bandwidth of the packet switching call) and a PCM call is plotted as well[4].

The real bandwidth on an IP network decreases as the packetization delay (and thus the payload size) increases, because of the fixed IP header size. Since ATM uses fixed size cells, Fig. 3 shows the real bandwidth of an ATM phone call for voice payload sizes only up the the size of an ATM payload. The real bandwidth increases for decreasing voice

[3] Throughout the paper we often refer to the load on link D as the load on the network. This is motivated by the fact that being D the potential bottleneck link of the considered topology, its utilisation is a good representative of the overall load on the network.

[4] In a circuit switched network, the real, apparent and effective bandwidth are the same.

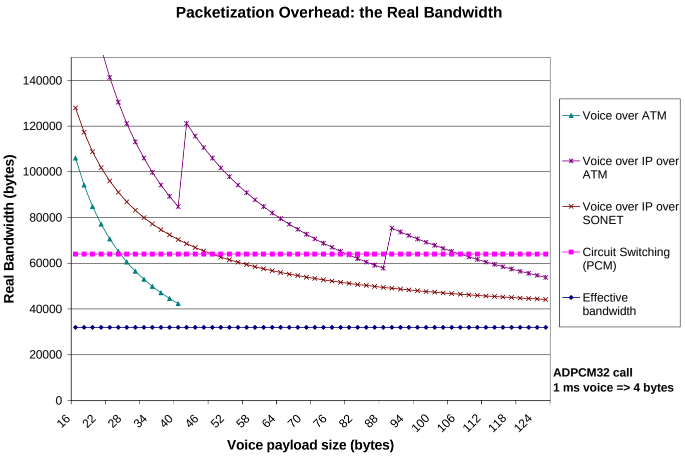

Fig. 3. Impact of packetization delay over the real bandwidth of a phone call with various technologies.

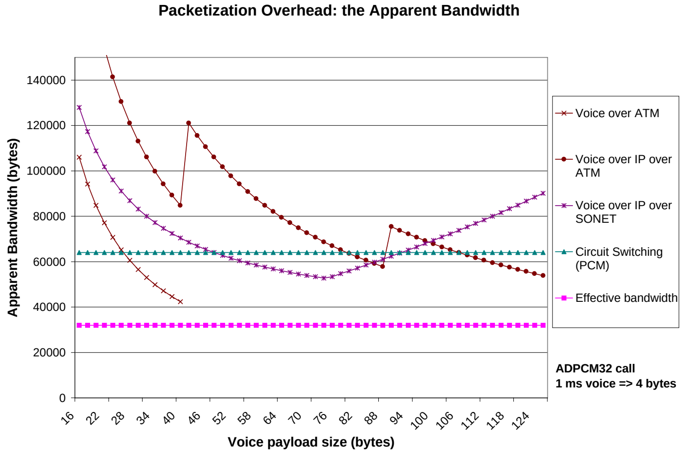

Fig. 4. Impact of packetization delay over the apparent bandwidth of a phone call with various technologies.

payloads because fixed size cells are sent more frequently. It is worth noting that encoding schemes rather than ADPCM32, such as GSM, generate voice samples larger than 48 bytes; such cases feature voice payload sizes larger than 48 bytes and longer packetization delays.

When IP packets are encapsulated into ATM cells, the real bandwidth tends to decrease, but discontinuously. This is due to the fact that when the IP payload size is increased, the IP packet size sometimes exceeds the size of an integral number of cell payloads, so a new cell is needed to carry a fragment of the packet. The real bandwidth is anyway larger than when IP packets are encapsulated in PPP frames transmitted directly over SDH/SONET links.

Fig. 3 shows that if the packet size is chosen in such way that the overhead introduced by the header is small enough, the transmission of a phone call through a packet network requires less bandwidth than through a circuit switched network where a whole 64 Kb/s channel is reserved. This means that the transport efficiency in a packet telephone network can be higher than in traditional telephone networks.

Fig. 4 depicts the apparent bandwidth needed to meet the

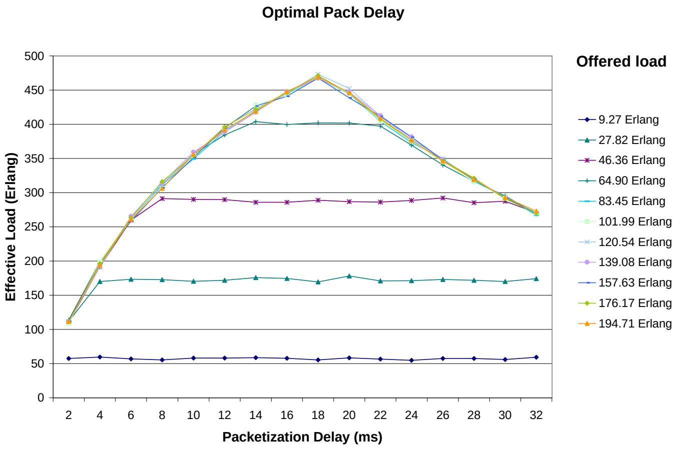

Fig. 5. Optimal packetization delay.

200 ms round-trip delay requirement versus various packetization delays using different technologies. If the delay requirement is not too tight, the real-time efficiency can be higher than 1, i.e. the packet switched network can carry more phone calls than the circuit switched one.

Fig. 3 and Fig. 4 show that the apparent bandwidth differs from the real bandwidth only on the IP network. Thus, in the considered scenario, IP is the only technology which requires bandwidth over-allocation. In fact, as the packetization delay increases, the delay budget left to queueing shrinks and over-allocation is required in order to keep the end-to-end delay below the QoS requirement. Thus, there exists an optimal packet size which, by providing minimum apparent bandwidth for a call, maximises the efficiency of IP telephony. The optimal packet size can be devised analytically [3] and intuitively seen in Fig. 5.

IP over ATM provides lower apparent bandwidth (higher number of calls carried) than IP over SONET/SDH for long packetization delays. This stems from the fact that even though the IP payload size, large enough to generate a low real load, introduces a large packetization delay, no bandwidth over-allocation is required. In fact, due to the small cell size, the queueing delay experienced by ATM cells in the network is short enough to comply with the delay budget left to queueing.

Among the various packet technologies, ATM is the one characterised by the smallest apparent bandwidth because (1) no over-allocation is required due to the low queueing (and packetization) delay and (2) the real load is low due to the small cell header.

In general, all the packet technologies require bandwidth over-allocation if due to a tight delay bound or to a large propagation delay, the delay budget left to queueing is too small to be satisfied with a service rate equal to the real bandwidth. This is detailed in the next section which deals with the number of hops on the path of a call.

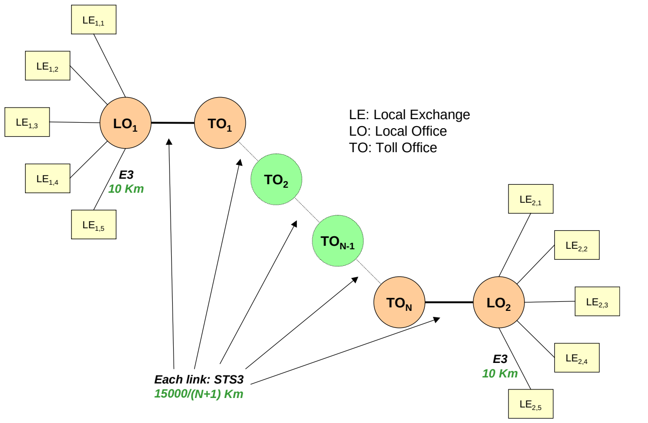

Fig. 6. Network topology used in simulations on long distance paths with increasing number of intermediate nodes.

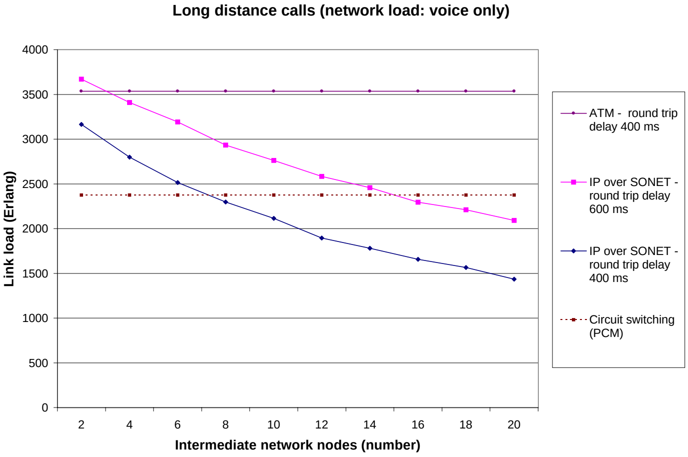

Fig. 7. Maximum long distance call load accepted by the network carrying only voice traffic.

V. Hops

The network topology shown in Fig. 6 with a variable number of toll offices is used to evaluate the impact of the number of nodes traversed by calls. Simulations are run with two QoS requirements: 400 ms and a 600 ms round-trip delay[5]. The IP packet size is chosen to maximise the number of voice calls accepted on the network; ATM cells are filled completely.

Figures 7 and 8 show the maximum call load accepted by the network versus the number of nodes on the path of the calls. The graphs show two different working conditions:

1. the network is used to carry mainly voice calls (Fig. 7) and
2. the network is deployed to carry the same amount of voice and data (Fig. 8), i.e. 50% voice, 50% data.

The first case requires the maximisation of real-time efficiency, while in the second one the focus must be on transport efficiency.

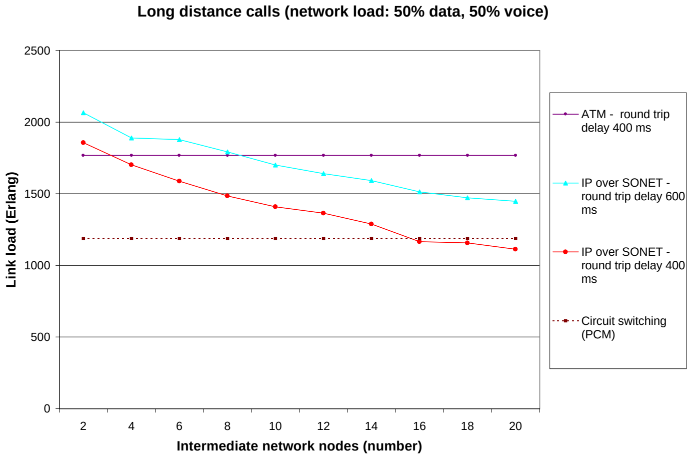

Fig. 8. Maximum long distance call load accepted by the network carrying 50% voice and 50% data traffic.

The load accepted on both a circuit switched and an ATM network is not affected by the number of hops traversed by a call because the contribution of each switch to the delay is small. Due to the small size of cells, ATM requires over-allocation only for large numbers of nodes, as shown in Figure 9.

Instead, an increase in the number of hops traversed significantly decreases the volume of calls accepted on an IP network. Unfortunately the Internet usually features a large number of routers on long distance paths. The topology of an IP network intended to carry telephony must be designed with this goal in mind in order to keep small the number of hops on any path.

If the number of phone calls accepted on the IP network with respect to the number accepted on the ATM or circuit switched one is considered as a performance index, the IP technology shows better performance in Fig. 8 than in Fig. 7. This means that IP technology can be more suitable for scenarios where the real-time traffic is not a relevant part of the overall network traffic. The simulations of Fig. 8 feature large IP packets in order to reduce the real load, thus increasing the transport efficiency. This requires over-allocation, however the performance of the solution is high because the over-allocation does not exceed the amount of best effort traffic, thus not limiting the amount of real-time traffic to be accepted. Rather, the amount of real-time traffic accepted on the network is limited by its real bandwidth.

VI. Call Admission Control and Resource Allocation

The results presented so far refer to a CAC based on *flat allocation* which reserves the same amount of bandwidth on all the nodes on the path of a call. Two other allocation schemes has been implemented in the simulator. The *capacity allocation* one implies that the allocation on each link is proportional to the link capacity. According to the *available allocation* method, the allocation on each link is proportional to the bandwidth available on the link.

The left axis of Fig. 10 shows the call load accepted by

[5] The provider could be willing to offer a low cost long distance service for which the user is required to tolerate higher round trip delays.

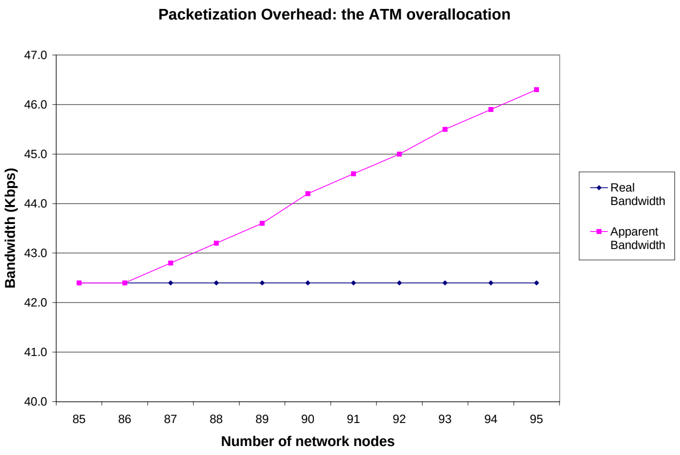

Fig. 9. Over-allocation with ATM.

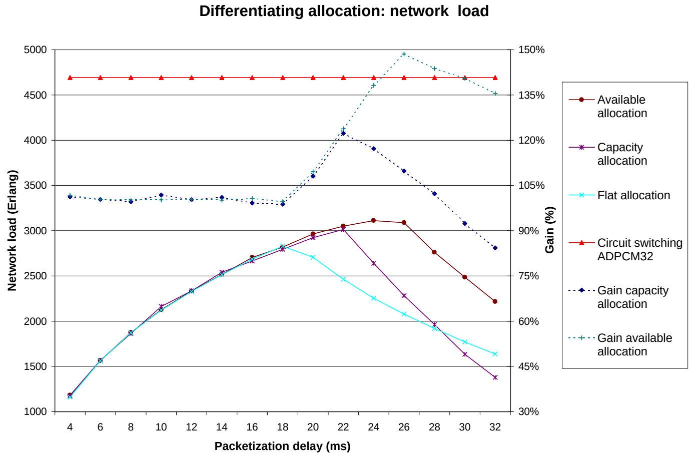

Fig. 10. Different allocation criteria: call load accepted by the networks and, on the right, the gain with respect to the flat resource allocation.

the network with the various over-allocation methods versus the packetization delay. The right axis shows the gain of the capacity allocation and available allocation methods with respect to the flat one. The available allocation method performs significantly better than the others.

VII. Discussion

In this paper we analyse the efficiency of packet telephony based on IP and ATM. Packet telephony features advantages over traditional circuit switched telephony: both data traffic and voice traffic are carried on the same network, cheap packet switches are deployed in place of circuit switches, and high performance codecs are exploited to produce voice flows at a very low bit rate.

However, in order to be able to keep low the user perceived delay, bandwidth might be *over-allocated* to voice calls, especially if IP is the packet technology deployed. Over-allocation reduces the maximum amount of voice traffic the network is able to carry, i.e., the *real-time efficiency* of the network. The higher the number of nodes on the path of voice calls, the larger the over-allocation required. ATM based packet telephony features high real-time efficiency, even when the number of nodes involved is high; also IP over ATM provides higher real-time efficiency than IP itself.

Bandwidth over-allocation is not an issue when the network is intended to carry a large amount of best effort traffic and IP can be a better choice. In this scenario in fact the network administrator should be concerned with maximising the transport efficiency, rather than the real-time one. This suggests that according to the ratio between the amount of real-time traffic and best effort traffic, the focus should switch on either of the two efficiencies. The point at which this switch of focus should be performed will be the subject of our future research.

Our further work is aimed at studying the real-time efficiency of packet telephony with *statistical* guarantees. More effective voice codings, like the ones based on silence suppression, will also be taken into consideration.

Acknowledgements

This work has been partially supported by Hewlett Packard. The authors wish to thank Vincenzo Frappietro for his work on the implementation of the simulator.

References

[1] The ATM Forum. *ATM User-Network Interface Specification - Version 3.1*. The ATM Forum, September 1994.

[2] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin. Resource ReSerVation protocol (RSVP) - version 1 functional specification. Standard Track RFC 2205, Internet Engineering Task Force, September 1997.

[3] M. Baldi, D. Bergamasco, and F. Risso. On the efficiency of packet telephony. In 7^{th} *IFIP International Conference on Telecommunication Systems*, March 1999.

[4] A. K. Parekh and R. G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The siglenode case. *IEEE/ACM Transactions on Networking*, 1(3):344–357, June 1993.

[5] A. K. Parekh and R. G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The multiple node case. *IEEE/ACM Transactions on Networking*, 2(2):137–150, April 1994.

[6] A. Demers, S. Keshav, and S. Shenker. Analysis and simulation of a fair queuing algorithm. *ACM Computer Communication Review (SIGCOMM'89)*, pages 3–12, 1989.

[7] C. Partridge. *Gigabit Networking*. Addison Wesley, October 1993.

[8] S. Golestani. A stop-and-go queuing framework for congestion management. In *ACM SIGCOMM '90*, pages 8–18, September 1990.

[9] Ion Stoica, Hui Zhang, and T. S. Eugene Ng. A hierarchical fair service curve algorithm for link-sharing, real-time and priority service. In *ACM SIGCOMM'97*, 1997.