On Modeling VoIP Traffic in Broadband Networks

Ehsan Haghani, Swades De, and Nirwan Ansari
Advanced Networking Lab
Department of Electrical and Computer Engineering
New Jersey Institute of Technology
Newark, NJ 07102, USA
Email: Ehaghani@ieee.org, {Swadesd, Nirwan.Ansari}@njit.edu

***Abstract*— With the general trend towards ubiquitous access to networks, more users will prefer to make voice calls through the Internet. Voice over IP (VoIP) as the application which facilitates voice calls through the Internet will increasingly occupy more traffic. The growth of delay sensitive traffic that requires special quality of service from the network will impose new constraints on network designers who should wisely allocate the limited resources to users based on their required quality of service. An efficient resource allocation depends upon gaining accurate information about the traffic profile of user applications. In this paper, we have studied the access level traffic profile of VoIP applications and proposed a realistic distribution model for VoIP traffic. Based on our model, we have introduced an algorithm for resource allocation in networks. It is shown that using our algorithm will enhance the delay and utilization performance of the network.**

I. Introduction

Voice over IP (VoIP) is a rapidly growing service which is providing voice communications over packet-switched networks. With extensive growth of the Internet and growing demand for provisioning various applications and services, more service providers are trying to provide people with new applications and technologies to make their voice calls through the Internet.

With the advent of new technologies, people can access the Internet through different types of connections. These different technologies may impose various challenges on network design, but, in the terms of their functionalities, they all have to provide customers with network resources to run their applications. Regardless of the data link and physical layer protocols used in these technologies, the IP layer traffic generated by VoIP applications does exhibit similarities. These similarities become more visible when users are employing a broadband access, demanding services for their applications through a high speed connection. For example, a user who makes voice calls through a 3G wireless network expects the same voice quality as that of a DSL user. Towards the necessity of a thorough study of VoIP traffic, we try to investigate some of the key factors of VoIP traffic in this paper.

There are many IP based applications that generate VoIP packet traffic in the Internet, and we consider some of the main characteristics of the these traffics in this paper. Although there are various VoIP commercial applications that provide voice connections between PCs and phones, they all generate IP packets in the Internet. Some of the main commercial applications are Skype, MSN, Yahoo messenger, etc. These applications could be run on any PC or wireless device, and it is up to the service providers to provision their customers with enough network resources to make their voice calls.

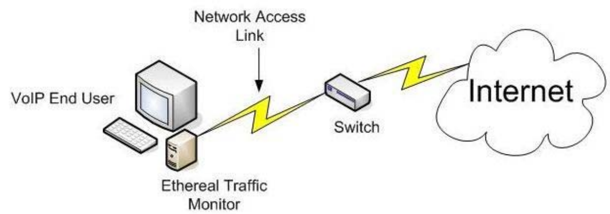

Fig. 1. Traffic Monitoring System

Knowledge of VoIP traffic characteristics becomes more crucial especially when the service providers are encountering scarcity in network resources and they are required to allocate their resources as efficiently as possible. This demand has motivated many researchers to study and model voice traffic. By reviewing the literatures, we can find many articles on modeling call arrival rate and call duration [1], [2]. In our research, we are interested in capturing traffic characteristics of a single VoIP connection at the end user. Understanding the main features of the VoIP traffic at the end user will help to anticipate the packet generation time which can be capitalized to improve the network efficiency.

There have been much research effort in modeling the VoIP traffic. The main traffic model adopted for voice traffic at the end user is the *ON-OFF* model [3]. This model is inspired by the nature of voice which is composed of periods of silence and sound. In this model, the source generates equal-size packets separated equally in time during the *ON* period and either does not generate any packet or generate smaller packets during the *OFF* period . Based on the voice characteristics, the duration of each period is assumed to be predetermined in this model [3].

Although the ON-OFF model has been used for studying the behavior of VoIP applications in networks [2], [4], the modeling of VoIP packets generated at end users requires further investigation. This is becoming more obvious when we note that VoIP applications do not necessarily perform the same procedure for generating voice packets [5]. It is also worth noting that the impact of transport layer protocols on VoIP packet streams is not considered in the ON-OFF model. Since VoIP applications might use different transport layer protocols such as TCP, UDP, or SCTP, the procedure of generating IP packets would be different, and the last-mile networks that deal with IP flows need further information to anticipate the behavior of VoIP traffic.

In order to model the traffic profile at the end user more accurately, we decided to run VoIP applications and monitor the packets traversed through networks. By monitoring the uplink traffic at each user, we could generate real traces for VoIP traffic, that incorporate the impact of all protocols above the IP layer. The resulted traces would help us gain better insight on the VoIP traffic profile which would guide us towards more detailed and accurate traffic modeling for VoIP applications.

This paper proposes a realistic model for VoIP traffic traversed in the uplink based on real traffic traces. In particular, the model describes the *inter-packet time* of VoIP traffic. The inter-packet time is the time between two consecutive packets sent to the network from the VoIP applications. The outline of the rest of the paper is as follows. In Section II, we explain the methodology for generating the traffic traces. Section III explains the characteristics of the generated traces. In Section IV, a VoIP traffic model is introduced, and its parameters are elaborated. Section V presents simulation results and performance comparison of conventional TDMA networks with that of networks that exploit the proposed model in resource allocation for VoIP traffic. Concluding remarks are given in Section VI.

II. Methodology for Trace Generation

As mentioned in Section I, in order to understand the traffic characteristics of VoIP connections at end users, we need to capture the VoIP packets. We ran the *ethereal* at each side of the VoIP connection. By using ethereal, we could capture the traffic at each layer, and the captured data would help us generate real-time traffic traces. An overview of the trace generating system is shown in Fig. 1.

Our method in generating traffic traces has also been used in other research works [6], [7]. Nevertheless, we will apply this method in deriving a legitimate model for VoIP traffic. For this reason, we have generated hundreds of traffic traces, and by studying them we could conceive the common characteristics of those traces. Additionally, we have monitored voice calls between different source-destination pairs with different VoIP applications

III. Trace Results

VoIP connections can be initiated by different applications, and either side of a connection might have access to different kinds of networks and use different devices. In order to generate legitimate traces, we established different kinds of VoIP connections via different applications. We also made voice calls to PCs and phones located in different parts of the world. As depicted in Fig. 1, we ran ethereal on a PC

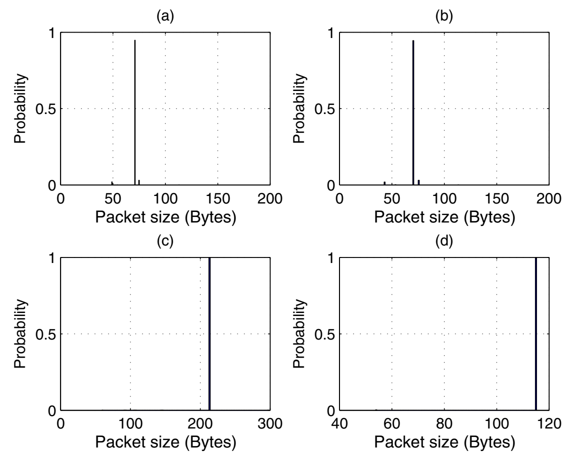

Fig. 2. PMF of VoIP packet size: a) Skype; b) another instance of Skype; c) Yahoo Messenger; d)MSN.

and connected it to the network via different technologies. We especially studied the traces generated from DSL and 100 Mbps LAN connections as they provide the end users with high speed network access. We also used Skype, MSN, and Yahoo Messenger as the applications for generating VoIP traffic. These applications can be easily installed on any PC or mobile device.

We have observed some particular characteristics of the VoIP traffic by studying the resulted traces. We examined the packet sizes and inter-packet time in the uplink for each voice connection. A summary of our results is described next.

A. Packet Size

By capturing the packets generated by VoIP applications we found that the packet sizes are not varying much during the time of a conversation. Although different voice connections might result in different packet sizes, each connection will bond to a relatively fixed size for the majority of its packets. The resulted probability mass function (PMF) of the VoIP packet size is shown in Fig. 2.

As mentioned in [8], G.711 and G.723.1 are two of the standard speech codecs used in VoIP applications. These standards generate equal size packets. The size of the packets is a function of the available bandwidth. On the other hand, transport layer protocols may change the size of the data segment, but the key factor resulted from the traces and standards is the fact that the majority of packets of a connection bond to a fixed size.

B. Inter-Packet Time

In order to gain real insight about the behavior of the VoIP packets, it is crucial to know their variations in a timely fashion. We monitored the uplink packets made by VoIP applications in different scenarios and measured the time between subsequent packets that we call *inter-packet* time. We measured the inter-packet time of VoIP packets for

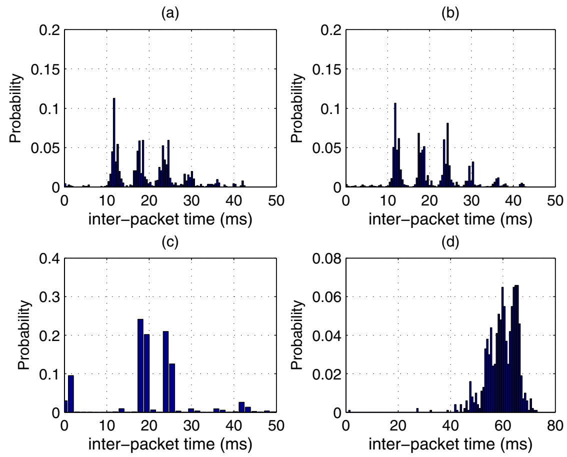

Fig. 3. PMF of VoIP inter-packet time from real traces: a) Skype; b) another instance of Skype; c) Yahoo Messenger; d) MSN.

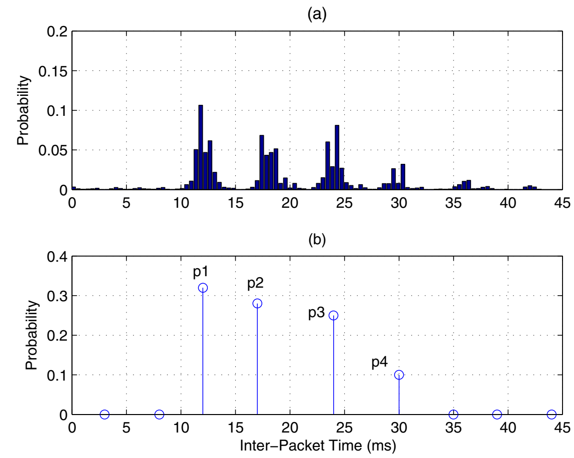

Fig. 4. Distribution of inter-packet time: (a) real distribution from the trace; (b) resulted model for distribution.

different destinations with different voice applications. The measurement tests were run during the call duration which is in the order of a few minutes. Some of the resulted distributions for inter-packet time of VoIP packets are shown in Fig. 3.

The inter-packet time of VoIP traffic has been measured in some other research works as well [5], [7], [8]. However, these works presented similar patterns for the traffic profile, but they had not neither proposed a distribution model for inter-packet time nor used this distribution in resource allocation.

As shown in Fig. 3, the inter-packet time for different calls will result in different distributions. Nevertheless, all the distributions can be accurately modeled, which will be discuss next.

IV. Distribution Model

Regarding the resulted traces shown in Section III, we want to introduce a model which captures the behavior of VoIP traffic at the uplink of end users. Considering the resulted traces, we found that more than 95% of generated packets will have the same packet size. Based on the route bandwidth, any VoIP application may deploy various voice coding standards. Thus, the size of packets generated by a VoIP application might not be the same for different voice calls [8].

We use our captured traffic traces to model the inter-packet time in the VoIP packets. As depicted in Fig. 3, for any voice connection, the inter-packet time is mainly located close to a few distinct values which are referred to as *taps* in this paper. Inspired by this observation, we define an inter-packet distribution as shown in Fig. 4. The distribution of the inter-packet time of uplink packets can be written as (1).

$$P_{\Delta}(\Delta) = \sum_{i=1}^{N} p_i(\Delta_i) \tag{1}$$

In (1), $P_{\Delta}(\Delta)$ is the Probability Mass Function (PMF) of the inter-packet time, N is the number of taps available in the distribution model, and $p_i(\Delta_i)$ is the probability of having a inter-packet time equal to Δ_i or $Pr\{\Delta = \Delta_i\}$. Therefore, the distribution of inter-packet time for any VoIP connection can be modeled with two matrices:

- The inter-packet time matrix Δ, which is $1 \times N$, shows the location of taps. $\Delta = [\Delta_1, \Delta_2, ..., \Delta_N]$.
- The probability matrix P, which is $1 \times N$, shows the value of PMF for each tap. $P = [p_1, p_2, ..., p_N]$.

Based on the PMF model matrices, the probability of sending the next packet in Δ_n^{ms} is

$$\Pr\{\Delta < \Delta_n\} = \sum_{i=1}^{n-1} p_i \tag{2}$$

Thus, the probability of having an inter-packet time longer than Δ_N is

$$\Pr\{\Delta > \Delta_N\} = 1 - \sum_{i=1}^{N} p_i \tag{3}$$

As an example, for a voice call trace and its deduced model shown in Fig. 4, with $N = 4$, the inter-packet time matrix Δ and the probability matrix P are as follows.

$$\Delta = [12^{ms}, 17^{ms}, 24^{ms}, 30^{ms}]$$

$$P = [0.32, 0.28, 0.25, 0.10]$$

For any voice call, the number of taps, N, and the value of PMF at each tap might be different but as we will show later, finding the parameters of the model, will be quite fast as compared to the call duration. It allows the resource allocators to make optimum decisions in the minimal time. As mentioned in Section III, in all of the generated traces, the packet sizes and inter-packet time values are close to some fixed values that are functions of the network bandwidth. Since these are the parameters of our model, revealing accurate model parameters in the least time will significantly improve the performance of the network. In the case of a change in network status as the codecs will update the packeting procedure, the model parameters will also be updated within a short period of time.

Fig. 5 shows the resulted PMF of inter-packet time calculated for the same trace at different times. As depicted in Fig. 5 the location of taps can be computed even in a short period of time like $1sec$.

V. SYSTEM SIMULATIONS

In this section, we propose an algorithm for resource allocation and simulate the impact of the proposed algorithm on the network performance. Since VoIP traffic is delay sensitive, we will especially consider the delay efficiency and bandwidth utility. For this purpose, we assumed a user running a VoIP application and requiring network access to transmit its packets in the uplink towards the destination. We supposed that parameters of the distribution model have converged and the network connection is in the steady state mode. We used the real traces explained in Section II for simulating the VoIP packet stream. As mentioned in Section III, the traces were generated in broadband networks with available uplink bit rates of more than $1Mbps$. Thus, the resulted inter-packet times shown in Fig. 3 were functions of the VoIP application rather than media access control (MAC) protocol. Therefore, the generated traces could be used for simulating the packet stream in any broadband network.

We considered the MAC layer of a broadband TDMA system consisting of frames, each with the length equal to Γ. Each frame is composed of a constant number of time slots, each with a length of τ. In our simulations, we assumed that τ is equal to the time that a user needs to transmit a VoIP packet. As VoIP traffic is delay sensitive, it is desired that the user can transmit the VoIP packet with the minimum delay. Therefore, the scheduling algorithm has to assign a time slot to the user as soon as it has a packet to send. Regarding the distribution of inter-packet time discussed in Section III, the time difference between two consecutive packets is a random variable. Furthermore, as mentioned in Section III, the size of VoIP packets is relatively small, and it will incur a waste of bandwidth and introduce additional delay if the user requests a time slot for any single packet. Accordingly, we assumed that the user cannot send requests for bandwidth; however, it was permitted to piggyback the bandwidth request with the data packets. Hence, if the user has more than one packet in its queue or the packet size is bigger than normal, it will piggyback the request for extra time slots in the data packet. Thus, the out of order or big packets will not waste any bandwidth as the scheduler will know their existence prior to any resource allocations.

There is a probability that the scheduler reserves a time slot for the user but the user does not have any packet to send, and as a result that reserved time slot would be wasted. This probability is higher if reservations are made frequently. On the other hand, if the scheduler assigns time slots less frequently, the transmission delay of the VoIP packet will increase. Therefore, the scheduler has to optimize the

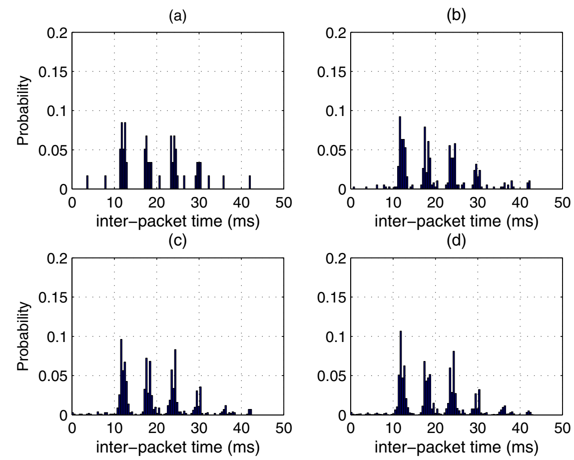

Fig. 5. Convergence of PMF for inter-packet time from traces with duration of T: a) T=1 sec; b) T=7 sec; c) T=20 sec; d) T=60 sec.

reservation in order to reduce both the packet delay and unused time slots. We compared two different scheduling algorithms: a conventional TDMA access method, and a novel method based on the traffic model discussed in Section IV.

- *Conventional Method*
  In this method, the user will have periodical access to network for sending its uplink VoIP packets towards the destination. A variable U is defined as the period of access in terms of frames. For example, $U = 1$ means the user will have access to the network in each consecutive frame, and $U = 2$ means it will gain access in every second frame. In our simulations, we examined the impact of $1 \leq U \leq 4$
- *Novel Method*
  In this method, we assume that the scheduler knows the parameters defined in Section IV. Therefore, the scheduler has enough knowledge to model and estimate the inter-packet time matrix elements Δ_i for $i = 1, 2, ..., N$. Based on this information, the scheduler wisely reserves a time slot for the user to transmit its data at the time of Δ_1. We assumed that if $\Delta_1 < \Gamma$,then the packet would be sent in the very next frame, and in other cases the packet would be sent in the frame, located Δ_1 away from the pervious packet. If the packet has not been generated till Δ_1, the scheduler will reserve another access in the frame located Δ_2 away from the previous packet. In this case, the first reservation would be wasted as it could have been assigned to other users. The scheduler will continually reserve time slots for the user at each Δ_i which is an element of the inter-packet time matrix Δ unless the source has a packet to send or the time elapses Δ_N from the previous packet. As shown in Fig. 3, more than 95% of the packets will be generated within Δ_N of the previous packet, but for those few remainders the scheduler will reserve a time slot for the VoIP user in every consecutive frame located further than Δ_N till the

1930-529X/07/$25.00 © 2007 IEEE
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2007 proceedings.

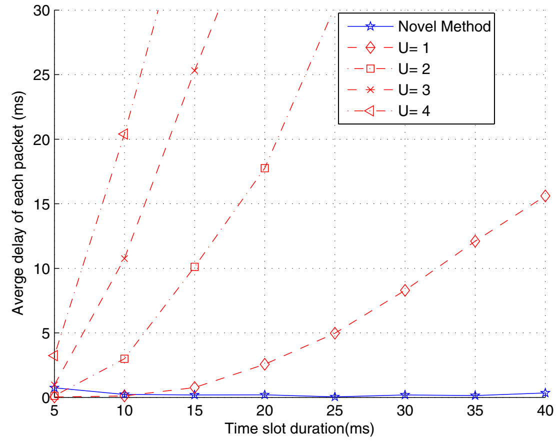

Fig. 6. Delay comparison

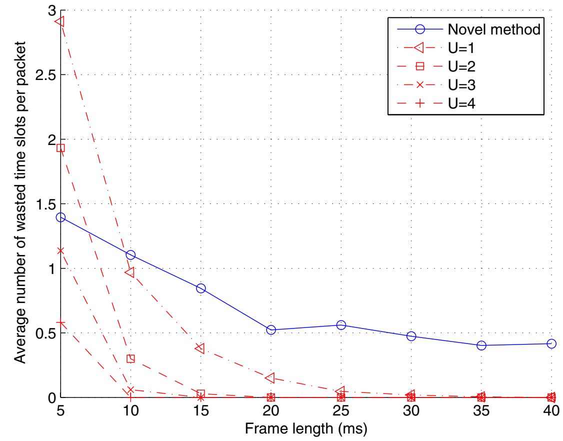

Fig. 7. Average wasted bandwidth

user sends its packet or the connection becomes timed out.

We measured the average number of frames a packet should wait till the user secures a permission for transmission. We also computed the average number of wasted reserved time slots due to the false estimation of the next packet generation time. The number of wasted time slots can be used as an indicator for the efficiency of the network bandwidth utilization. We changed the frame duration from $5ms$ to $40ms$, and observed the impact of U in the conventional method.

Fig. 6 shows the average delay time that a packet waits before transmission versus frame duration Γ for the novel method and conventional method with reservation frequency $1 \leq U \leq 4$. Fig. 7 demonstrates the average number of time slots wasted for each packet before it grants the reservation. As depicted in Figs. 6 and 7, we observe that although the conventional method will waste slightly less time slots at longer frame lengths, its delay performance is far worse which is not acceptable for delay sensitive applications such as VoIP. It is also worth noting that for any frame length, the difference between the average number of wasted reservations for the conventional and novel methods would be less than 1 time slot which is negligible. Thus, the effectiveness of our proposed model is demonstrated.

VI. Conclusion

In this paper, we have studied the VoIP traffic behavior at end users. VoIP is an important application for next generation broadband access networks. Thus, understanding the characteristics of VoIP traffic is crucial for designing efficient networks. In order to determine an accurate model for VoIP traffic transmitted in the uplink, we have captured the uplink VoIP traffic generated by different applications at the end user, and we have defined an accurate model based on the resulted traces.

In this work, we have modeled the packet size and inter-packet time of VoIP traffic. We found that the packet sizes do not vary much during the conversation time. We have also proposed a multi-tap model to capture the features of the inter-packet time. Based on this model, we have proposed an algorithm for resource allocation in TDMA networks. It was observed that with the accurate anticipation of the packet generation time in our novel method, the average number of missed bandwidth reservations will be less than that of conventional methods for shorter frame lengths, and comparable to that of conventional methods for longer frame lengths. It was also shown that as the frame duration increases, the average delay that a packet waits at the source before transmission in the uplink using this algorithm will be less than that of conventional resource allocation methods.

References

[1] J. Daigle and J. Langford, “Models for analysis of packet voice communications systems,” *IEEE Journal on Selected Areas in Communications*, vol. 4, pp. 847 – 855, 1986.

[2] T. D. Dang, B. Sonkoly, and S. Molnar, “Fractal analysis and modeling of VoIP traffic,” in *Proceedings 11th International Telecommunications Network Strategy and Planning Symposium*, Jun. 2004, pp. 123 – 130.

[3] P. Pragtong, T. J. Erke, and K. M. Ahmed, “Analysis and modeling of VoIP conversation traffic in the real network,” in *Proceedings Fifth International Conference on Information, Communications and Signal Processing*, Dec. 2005, pp. 388 – 392.

[4] W. Jiang and H. Schulzrinne, “Analysis of on-off patterns in VoIP and their effect on voice traffic aggregation,” in *Proceedings Ninth International Conference on Computer Communications and Networks*, Oct. 2000, pp. 82 – 87.

[5] W.-H. Chiang, W.-C. Xiao, and C.-F. Chou, “A performance study of VoIP applications: MSN vs. Skype,” in *Proceedings The First Multimedia Communications Workshop (MULTICOMM)*, Jun. 2006, pp. 13 – 18.

[6] H. Toral-Cruz and D. Torres-Roman, “Traffic analysis for IP telephony,” in *Proceeding 2nd International Conference on Electrical and Electronics Engineering*, Sept. 2005, pp. 136 – 139.

[7] A. Sukhov, P. Calyam, W. Daly, and A. Illin, “Towards an analytical model for characterizing behavior of high-speed VVoIP applications,” in *TERENA Networking Conference (TNC)*, Jun. 2005.

[8] S. Sharafeddine, A. Riedl, J. Glasmann, and J. Totzke, “On traffic characteristics and bandwidth requirements of voice over IP applications,” in *Proceedings Eighth IEEE International Symposium on Computers and Communication (ISCC)*, Jul. 2003, pp. 1324 – 1330.