Analysis of the Delay and Jitter of Voice Traffic Over the Internet

Mansour J. Karam, Fouad A. Tobagi

Abstract— **In the future, voice communication is expected to migrate from the Public Switched Telephone Network (PSTN) to the Internet. Because of the particular characteristics (low volume and burstiness) and stringent delay and loss requirements of voice traffic, it is important to separate voice traffic from other traffic in the network by providing it with a separate queue. In this study, we conduct a thorough assessment of voice delay in this context. We conclude that Priority Queuing is the most appropriate scheduling scheme for the handling of voice traffic, while preemption of non-voice packets is strongly recommended for sub-10 Mbit/s links. We also find that per-connection custom packetization is in most cases futile, i.e. one packet size allows a good compromise between an adequate end-to-end delay and an efficient bandwidth utilization for voice traffic.**

I. Introduction

The Internet will become a ubiquitous infrastructure, used by numerous applications having various requirements and that generate traffic that has different characteristics: in particular, web-based data applications, video applications and voice applications. Voice applications are expected to migrate from the Public Switch Telephone Network (PSTN) that services them today to the Internet. Owing to constant improvements over the years, traditional voice communication over the PSTN is today characterized by what is referred to as toll quality, that is low delay, high availability and adequate voice quality. For the Internet to compete with the PSTN, it should provide the same level of quality, which implies stringent delay, loss and reliability requirements on voice communication. In terms of traffic characteristics, voice streams have low data rates (in the order of tens of Kbit/s) and exhibit low burstiness. Because of these stringent requirements and particular characteristics, voice traffic should be treated differently than other traffic in the network. In fact, measurements on the Internet [28] as well as simulation studies [22] have shown that mixing voice traffic with both traditional TCP data traffic and UDP VBR video traffic can lead to either low average link utilization if delay requirements are met, or larger than desired delay for voice. Accordingly, allowing a mixture of voice with other traffic can lead to the need of complex admission control policies if the end-to-end delay requirements for voice were to be satisfied [22]. Hence, we assume in this study that voice traffic is separated from other traffic in the network by providing it with its own queue. That is, voice could be provided with a separate link, or a separate circuit using, for example Multi-protocol Lambda Switching [1]. If other traffic is flowing on the link, then voice traffic could be serviced using Priority Queuing (PQ), and given the highest priority over all other traffic. In LANs, such a treatment is suggested in the context of the IEEE 802.1p extension to the IEEE 802.1D standard [18]; at the Internet scale, high priority can be provided to voice traffic in the Differentiated Services framework [2], by means of mapping voice traffic to the Expedited Forwarding Per-hop Behavior (EF PHB) [19], and giving high priority to EF traffic relative to other traffic in the network. Finally, voice traffic could be given its fair share of the link using Weighted Round Robin (WRR) [24] or comparable schemes ([14], [30], [32]).

In this paper, we aim to demonstrate that in this context, the requirements of voice traffic can be attained using simple mechanisms, both in terms of scheduling and packetization. In this respect, we start with describing the particularity of voice traffic, in terms of characteristics, requirements and delay components. Singling out queuing delay as the only source of jitter, we present the methodology used to quantify it. We come up with appropriate models for voice delay, justify the intuition behind their choice and show their accuracy by comparing the resulting delay distributions to those obtained through network simulation. Using these models, we show that scheduling voice traffic using PQ and packetizing voice streams using a fixed size packetization scheme lead to an adequate handling for voice traffic in the Internet.

The paper is organized as follows: we start with a review of prior work (Section II). In Section III, we describe the particularity of voice traffic in terms of characteristics, requirements and delay components. In Section IV, we describe the models used for quantifying the network delay and jitter incurred by voice traffic in the Internet. We thereafter present the results of the analysis. In Section V, we compare the delay obtained for voice traffic using different scheduling schemes. By the same token, we also investigate the effect of various parameters on network delay and jitter. In Section VI, we comment on the choice of packet size for an adequate handling of voice traffic. Finally, we end in Section VII with a summary of the results and some concluding remarks.

II. Prior Work

Since the concept of a packet network that supports both voice and data traffic emerged in the early eighties, the work on packetized voice has been extensive. Queuing models were developed, used and compared to understand the queuing behavior of voice traffic in a packet network. In particular, the superposition of multiple periodic streams, which constitutes the most intuitive model for packetized voice was presented and analyzed analytically in [9]. The queuing model obtained is denoted $\sum D_i/D/1$. [1] [8] and [21] used $\sum D_i/D/1$ to derive the dis-

Mansour Karam is with Speedtrak Communications, San Mateo, CA 94401. Email: mans@speedtrak.com. Fouad Tobagi is with the Computer Systems Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA. Email: tobagi@stanford.edu

[1] In $\sum D_i/D/1$, the input process consists of the superposition of a fixed number of periodic streams, and each stream is characterized by a deterministic service time. The distribution of the queuing delay incurred by a packet in a stream is obtained, assuming different instances of the same experiment, where for each instance the phase of the various stream with respect to each other is chosen at random.

tribution of the queuing delay that is incurred by voice traffic, while [23] also derived the corresponding buffer distribution. [9], [21], and [29] compared the delay results obtained using $\sum D_i/D/1$ to those pertaining to the simpler $M/D/1$ queuing model, which was commonly used in practice to estimate the sizing of real systems. These studies found that $M/D/1$ significantly over-estimates the delays incurred on packetized voice when the utilization on the link is high and the number of streams multiplexed on the link low. On the other hand, [9], [21], and [29] found that in the case of lightly utilized high speed links, where the utilization is low and the number of streams multiplexed exceeds 100, $M/D/1$ and $\sum D_i/D/1$ yield similar results. [31] studied the effect of Speech Activity Detection (SAD) (that is, silence suppression) on voice delay: it was found that the increase in traffic variability that results from the inclusion of SAD in the encoding process hinders the advantage that is obtained from the reduction in average utilization[2]. Finally, [26] used the $\sum D_i/D/1$ and $M/D/1$ models to derive one-hop queuing delay, and convolution over many hops to derive multi-hop queuing delay for voice traffic.

Today, as the Differentiated Services architecture [2] has gained in popularity, new studies have emerged, aiming at specifying a service based on the EF PHB [19] that would be appropriate for the handling of voice traffic. PQ has been proposed by some as a simple and adequate scheduling scheme in this context [20], but refuted by others [7] because of the significant queuing delay variations (i.e., jitter) that could be incurred as a result of the residual transmission time of lower priority packets. In fact, [7] shows that the worst-case queuing delay incurred by voice traffic increases exponentially with the number of hops traveled, as a result of the increase in traffic burstiness with the number of hops. Conversely, [16] and [27] show that settling for a statistical guarantee (i.e., accepting a small portion of lost traffic) allows a significant reduction in delay, and suggest that PQ is indeed adequate for the handling of voice traffic. However, the results obtained are inconclusive, as the multi-hop scenario used in both studies underestimates the increase in traffic burstiness as the number of hops in the path of a voice stream increases.

III. Particularity of Voice Traffic

A. Voice Traffic Characteristics and Requirements

Voice Characteristics. The traditional voice encoder is G.711 (and its variants [10], [11]), which uses Pulse Code Modulation (PCM) to generate 8 bits samples per 125 microseconds, leading to a rate of 64 Kbit/s. In the last decade, new voice encoding schemes have been developed, which use Code Book Excited Linear Prediction (CELP) techniques, leading to drastic rate reductions at the expense of additional encoding delay: 8 Kbit/s for G.729A [13], 5.33 Kbit/s for G.723.1 [12]. Taking into account the headers that correspond to each of the protocol layers, the rate of the packetized voice stream remains in the order of tens of Kilobits per second, which is much lower than the data rates that correspond to typical video and data traffic. In addition, speech consists of an alternation of talk-spurts and silence periods. Silence suppression at the source takes advantage of this fact, leading to a substantial reduction of the average rate at the expense of increased variability.

[2]More specifically, [31] found that the queuing delay resulting from an $M/D/1$ model in which the rate of the incoming (Poisson) process is set equal to the average incoming rate of voice traffic heavily underestimates the delay incurred by voice traffic in the network.

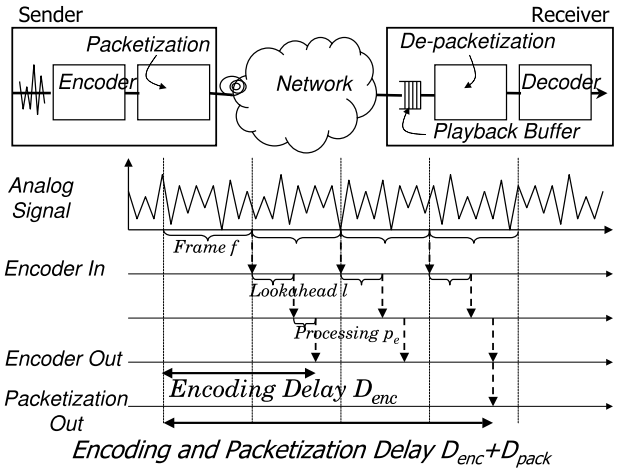

Fig. 1. End-to-end delay components for voice traffic.

Voice Requirements. Voice requirements are stringent: toll-quality real-time communication is needed, which limits the maximal tolerable round-trip delay to 200-300 ms; that is, one-way delay must be in the range of 100-150 ms for adequate performance. In addition, jitter should be small enough (i.e. 50 ms at most) so that playback at the receiver remains smooth. On the other hand, the tolerable packet loss L_{max} is small. In fact, since packet loss in the Internet is correlated [3], if packet loss were to occur, the number of contiguous packets which are lost is usually larger than one. Hence, the duration of the corresponding portion of voice bit-stream that is lost (which we refer to as a "clip") can easily exceed 60 ms even when a smaller packet formation time is used. [17] shows through subjective testing that clips exceeding 60 ms affect the intelligibility of the received speech. For this reason, for toll-quality communication, L_{max} must be set to a relatively low value (i.e., 10^{-5}) to insure that such clips occur infrequently. Taking advantage of L_{max} as opposed to conducting a worst case analysis is necessary, since the later would lead to delay results which are extremely pessimistic (e.g., [7]), and that do not apply to realistic situations. Consequently, in the remainder of this study, we measure delay by the $(1 - L_{max})^{th}$ percentile, where $L_{max} = 10^{-5}$.

B. End-to-end Delay Components of Voice Packets

End-to-end delay consists of the delay incurred by the voice signal from the instant it is produced by the speaker until it is heard by the listener at the destination (see Figure 1): the analog signal is first encoded, incurring an encoding delay D_{enc}, which in turn consists of the sum of the frame size f, the look-ahead delay l, and the processing delay p_e (see Table I). In general, the lower the rate of the encoded bit-stream r, the larger the frame size, look-ahead delay and processing delay, and consequently D_{enc}. The encoded bit-stream thus generated is then packetized, incurring a packetization delay D_{pack}, function of the number of frames k included in one packet, i.e. $D_{pack} = (k-1)\,f$. Voice packets are then transmitted on the network, incurring transmission delay T_h, queueing delay Q_h and propagation delay P_h at

TABLE I

FRAME SIZE f, LOOK-AHEAD DELAY l AND PROCESSING DELAY p_e FOR G.711, G.729A AND G.723.1.

Encoding scheme (Nominal bit rate r)	G.711 (64 Kbit/s)	G.729A (8 Kbit/s)	G.723.1 (5.3/6.4 Kbit/s)
Frame size f	125μs	10ms	30ms
Look-ahead l	0	5ms	7.5ms
Processing delay p_e	Negligible	Less than 10ms	Less than 30ms

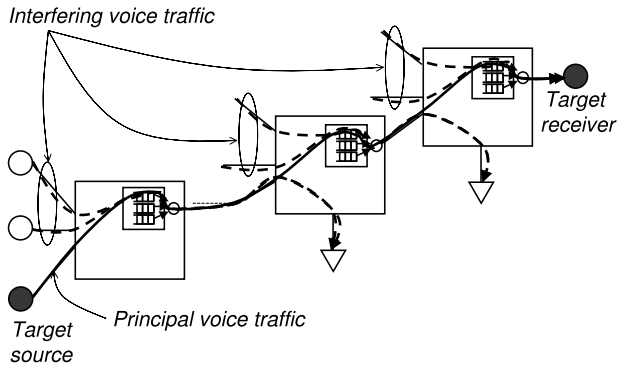

Fig. 2. General network Topology.

each hop h in the path from the source to the destination. Propagation depends on the distance between the source and the receiver[3]. At the receiver, packets are delayed in a playback buffer incurring a playback delay D_{play}. De-packetization is then performed, and the reconstructed encoded bit-stream is decoded at the destination incurring a decoding delay D_{dec}. We consider processing to be fast enough so that processing delay at both the source and the receiver (p_e and D_{dec}) are ignored. Therefore, denoting the formation time $T_f \doteq kf$, we get

$$D = T_f + l + \sum_{h \in Path} (T_h + Q_h + P_h) + D_{play} \qquad (1)$$

From Equation (1), it is clear that for a given voice connection, the only random component of voice delay (that is, the only source of jitter) consists of queuing delay in the network, $Q \doteq \sum_{h \in Path} Q_h$. The playback delay D_{play} insures that most of the packets transmitted are available the instant they have to be handed to the decoder. Assuming that Q_{max} represents the maximum queuing delay percentile incurred in the network, the receiver must delay the first packet of a voice stream by a full Q_{max}, [4] i.e. $D_{play} = Q_{max}$; if, in addition, that packet has already incurred in the network a queuing delay equal to Q_{max}, then the end-to-end delay budget equation becomes

$$T_f + l + \sum_{h \in Path} (T_h + P_h) + 2Q_{max} \leq D_{max} \qquad (2)$$

Equation (2) clearly shows the importance of queuing delay, formation time and propagation delay. In particular, it can be used to estimate the maximum jitter Q_{max} that can be tolerated for a given connection; for example, if G.729A is used to encode the voice stream, a formation time $T_f = 30$ ms is used to packetize the encoded bit stream, the total transmission and propagation delay on the path are equal to 5 and 40 ms, respectively, then for $100 \leq D_{max} \leq 150$ ms, $10 \leq Q_{max} \leq 35$ ms. Transmission time being negligible most of the time (except on very slow links), Equation (2) clearly shows the importance of queuing delay, formation time and propagation delay, which are the three components we focus on in this paper.

[3]For calls within a given local area, propagation delay is negligible. For intra-continental calls within the United States (e.g., San Francisco to Boston), the propagation delay is in the order of 30 ms whereas inter-continental calls result in propagation delays ranging from 50 ms (e.g., San Francisco to Paris) to 100 ms (e.g., San Francisco to Hong-Kong).

[4]Even though RTP provides a sequence number and time-stamp for each packet, source-receiver synchronization is not supported. Hence, the source and receiver are typically non-synchronized, so the receiver can't determine the amount of jitter already incurred by the first packet received.

IV. MODEL FOR QUEUING DELAY IN THE NETWORK

In this section, we aim to identify the sources of queuing delay incurred by voice traffic and derive models that captures the effects of interest. We validate our models using network simulation.

The general network topology considered is shown in Figure 2; it consists of a succession of hops in tandem, which represents the path followed by a voice stream from the source to the receiver. To be conservative, we consider the pessimistic scenario in which the traffic that is interfering with the tagged stream is injected at each hop, independently of the other hops in the path. Also, we assume the general case in which the interfering traffic can be generated by a source many hops away from the switch at which the interference with the target stream takes place (as opposed to the source of the interfering traffic being directly attached to the switch in question). Doing so, the variability of the interfering voice traffic gets larger as it travels through multiple hops before it is made to interfere with the target stream, leading, in turn to a larger queuing delay. This scenario is more reflective of realistic traffic conditions. We consider a number of link speeds, ranging from 384 Kbit/s to 45 Mbit/s, including T1, 10Base-T and T3 links.

As described in Section I, we shield voice traffic from other traffic in the network by providing it with a separate queue. We consider that Silence Suppression is implemented by the voice encoder, yielding a decrease in the total voice load. Accordingly, the model used for voice traffic consists of an ON/OFF pattern, modeling the succession of talkspurts and silence periods generated by the voice encoder. As described in [8], talkspurts and silence periods alternate according to a two-state Markov chain. In our experiments, we consider average periods of talk-spurt and silence equal to 1.65 and 1.35 seconds, respectively ([8]).

We identify two sources of voice delay jitter (see Figure 3):

1. Queueing delay behind voice packets in the same queue: this component depends on the variability of the voice traffic pattern which, at the source originates from silence suppression. In case the link (or circuit) on which voice travels is not shared with other traffic, then a given voice stream is not affected as it travels through consecutive hops from the source to the receiver, that is its variability remains the same at each hop in the path. In

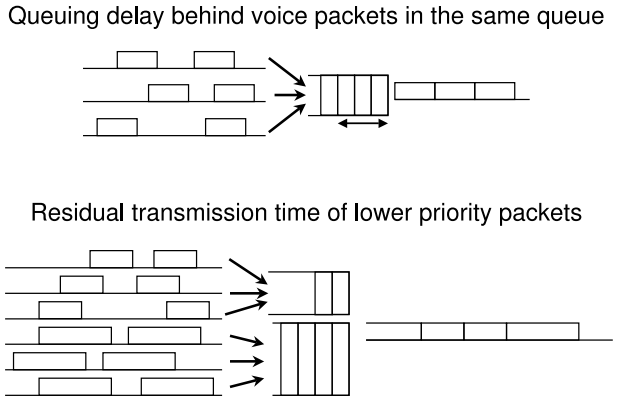

Fig. 3. The two source of jitter for voice traffic when provided with a separate queue.

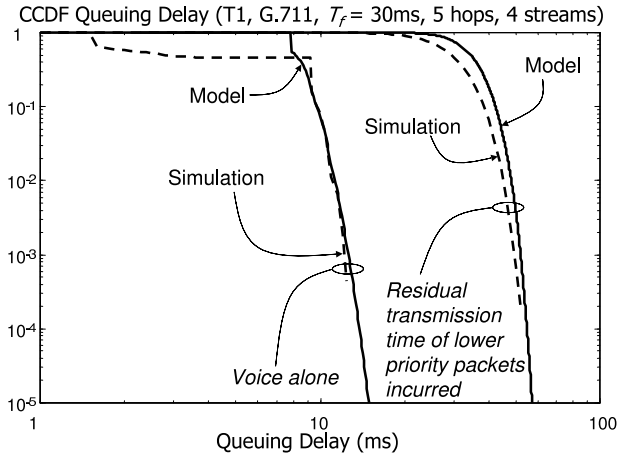

Fig. 4. Model versus simulation results.

this context, the interfering traffic at a given hop can be considered to be generated by sources directly attached to the switch in question.

2. In case voice traffic shares the link with other traffic, it then incurs the residual transmission time of non-voice packets. For example, in case non-preemptive PQ is used, then voice packets that arrive to the queue during the transmission of a lower priority packet will be delayed until the end of the ongoing transmission of that packet. Also, as voice packets corresponding to a given stream incur such a delay, their inter-arrival time is modified: as a result, voice traffic variability increases with the number of hops, in turn leading to an increase in the queuing delay incurred in the voice queue.

Accordingly, the jitter experienced clearly depends on whether voice traffic shares the link (or circuit) with other traffic. In the following section, we study each case independently.

A. Voice Alone on the Link or Circuit

Without silence suppression, the encoder produces frames at regular intervals f, while packets are generated at regular intervals T_f. That is, each voice stream is periodic with period T_f. Also, since packets are of equal size, then the service time of packets is deterministic, equal to $\overline{x} = rT_f + 8H$, where H denotes the header size in bytes. The queuing model that is characterized by an input process that consists of a superposition of n independent periodic sources, and by a deterministic service time is $\sum D_i/D/1$, and has received extensive attention in the past[5]. (See Section II.) In our case, silence suppression is implemented, so the voice traffic load is reduced and its variability increased. We have simulated a number of different scenarios using the topology shown in Figure 2. In Figure 4, we show the complementary queuing delay distribution and compare it to the complementary queuing delay distribution that results from a $\sum D_i/D/1$ model. The plot shows that the delay distribution obtained using the $\sum D_i/D/1$ model has a fast decaying tail, which is expected given the low level of burstiness in the traffic. Clearly, when silence suppression is implemented, then the resulting complementary delay distribution is bounded by that obtained with $\sum D_i/D/1$ for low delay values. However, the tail of the distribution is mainly affected by the periods in which all voice streams are in their talk-spurt. Thus, the tail of both distributions match very closely, and both models lead to interchangeable delay results for the 10^{-5} tolerable loss rate considered in this study. This phenomenon agrees with previous findings showing that the variability that entails from silence suppression increases the incurred delay percentiles, thus reducing the advantage obtained through an increase in the multiplexing gain [29] [31]. (See Section II.) Having observed this result for a large range of scenarios, we conclude that the $\sum D_i/D/1$ model approximates well the delay results obtained in case voice is transmitted alone on the link. Since hops are independent, then the total delay incurred by voice packets traveling through a given number of hops N is distributed according to the N-fold convolution of the one hop delay distribution.

[5]Note that in $\sum D_i/D/1$, the only source of randomness originates from the random phase of each of the periodic streams within a period T_f.

B. Voice Sharing the Link with Other Traffic

As described above, in case voice shares the link with other traffic, it incurs the residual transmission time of packets transmitted over lower priority queues. We treat the case of PQ first; we then extend the results to WRR. We first characterize the added variability that results from the perturbations on the high priority queue from packets belonging to the lower priority queues. We show that the resulting variability is bounded by that inherent to a Poisson process. We then consider the voice input process to be Poisson, and find the resulting distribution of queuing delay incurred by voice traffic.

For a given hop, we have

$$t_{k+1}^{'} - t_k^{'} = t_{k+1} - t_k + (\tau_{k+1} - \tau_k)$$

where $t_{k+1} - t_k$, $t_{k+1}^{'} - t_k^{'}$ represent the time interval separating two consecutive packets k and $k+1$ belonging to the same stream before their admittance to the queue, and after their transmission on the link, respectively; τ_k represents the perturbation introduced by lower priority packets over one hop, and is essentially equal to their residual transmission time. Assuming that lower priority packets are all full size MTU packets, τ_k is uniformly distributed, ranging from 0 to T_{MTU}, where T_{MTU} denotes the transmission time of a full size MTU packet. Thus, the perturbation $\xi_k = \tau_{k+1} - \tau_k$ is distributed according to a symmetric, triangular distribution ranging from $-T_{MTU}$

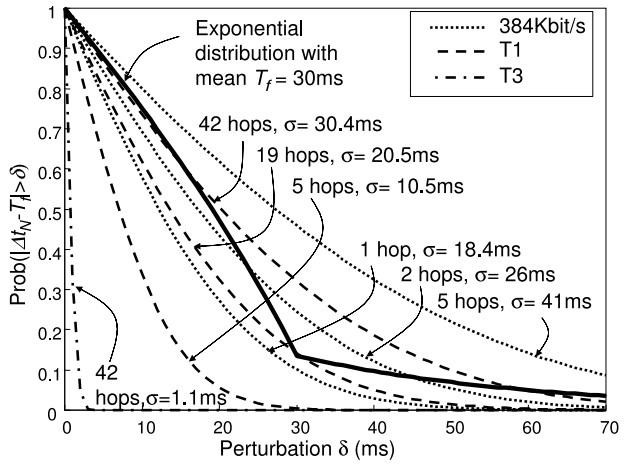

Fig. 5. $Prob\left(\Delta T - T_f > \delta\right)$ for $NORM\left(0, T_{MTU}\sqrt{\frac{N}{3}}\right)$ and Ξ (exponential, with mean T_f) for various number of hops and link speeds.

to T_{MTU}, with variance $\sigma^2 = \frac{T_{MTU}^2}{6}$. Furthermore, the perturbation incurred by the voice packets belonging to a stream traveling a given number of hops N is the sum of the perturbations incurred at each hop h, ξ_h. Since hops are independent, then the distribution corresponding to the total perturbation $\Gamma_N = \sum_{h=1}^{N} \xi_h$ is distributed according to the N-fold convolution of the distribution of ξ_h. From the central limit theorem, we know that $\Gamma_N \rightarrow \Omega_N$ as N increases, where $\Omega_N = NORM\left(0, T_{MTU}\sqrt{\frac{N}{6}}\right)$ represents a normal distribution with mean 0 and standard deviation $\sigma_N = T_{MTU}\sqrt{\frac{N}{6}}$. The convergence of Γ_N to $NORM\left(0, T_{MTU}\sqrt{\frac{N}{6}}\right)$ can be shown to be fast (within 10 hops), whereas the tail of the complementary distribution function of the limiting normal distribution always bounds that of Γ_N. Therefore, $NORM\left(0, T_{MTU}\sqrt{\frac{N}{6}}\right)$ constitutes a close approximation to Γ_N. Using Γ_N to compute the inter-arrival distribution of packet sizes $\Delta t_N = t_{k+1}^N - t_k^N$ for a voice stream traveling through N hops, we get $f_{\Delta t_N}(\alpha) = f_{\Omega_N^1 - \Omega_N^2}(\alpha - T_f)$, i.e. Δt_N is normal with mean T_f and standard deviation $T_{MTU}\sqrt{\frac{N}{3}}$. [6]

Note that the standard deviation of the inter-arrival process grows with the square root of the number of hops in the path of the voice stream; that is, increasing the number of hops only results in a contained increase in variability. Also, the tail of the distribution of Δt_N decays with $e^{-\alpha^2}$; intuitively, this shows that the resulting burstiness in the input process is lower than the burstiness of a Poisson arrival process, which distribution tail decays with $e^{-\alpha}$. To quantify this observation, we compare the distribution of Δt_N to that of an exponential distribution Ξ with mean T_f, (i.e $f_\Xi(\alpha) = T_f^{-1} e^{-T_f^{-1}\alpha}$). In Figure 5, we plot the probability that the difference between the inter-arrival period and T_f exceeds a given perturbation δ, a good measure of the stream variability. As can be seen from the figure, the variability of Δt_N only exceeds that of Ξ when links are extremely slow (384Kbit/s). For T1 links, the variability of Δt_N is the lowest for as much as a 19 hops path (that is, for all practical cases). Hence, if PQ is used, with voice traffic given the highest priority over other traffic in the link, we can use a Poisson inter-arrival process of mean T_f to bound the variability that results from the effect of lower priority non-voice packets. One side benefit in so doing is to render the distribution of inter-arrival time independent of the hop number in the path.

We now find the distribution of the queuing delay incurred by a voice stream as it traverses a given hop; we use in that end the treatment of priority functions found in [25]: first, we derive the Laplace transform for the waiting time for voice packet in a two priority system where the high (voice) and the low (non-voice) queues are both fed by a Poisson input process, and where the service time for packets in both queues is deterministic, equal to $\overline{x_1}$ and $\overline{x_2}$, respectively. We get

$$W^*(s) = \left[(1-\rho) + \rho_2 \widetilde{X}_2^*(s)\right] \sum_{n=0}^{\infty} \rho_1^n \left(\widetilde{X}_1^*(s)\right)^n \qquad (3)$$

where ρ_1, ρ_2 represent the average utilization of the high priority voice queue and lower priority non-voice queues, respectively, $\rho = \rho_1 + \rho_2$ represent the total average utilization of the link, and $\widetilde{X}_i^*(s) = \frac{1 - e^{-s\overline{x_i}}}{s\overline{x_i}}$, $i = 1, 2$ is the Laplace transform of the residual life of the service time for packets in each queue. To be conservative, we assume that the traffic in the lower priority queue uses up all of the bandwidth that is unused by voice traffic; that is, we set ρ to 1; thus, Equation (3) becomes

$$W^*(s) = \left[\sum_{n=0}^{\infty} (1-\rho_1)\rho_1^n \left(\widetilde{X}_1^*(s)\right)^n\right] \widetilde{X}_2^*(s)$$

that is, $w = w_{M/D/1} + \widetilde{x}_2$, where $w_{M/D/1}$ represents the waiting time obtained from an $M/D/1$ queuing system, whereas $\widetilde{x}_2$ represents the residual transmission time of packets in the lower priority queue. Thus the delay incurred by voice packets is simply the sum of an $M/D/1$ derived waiting time and the residual transmission time of lower priority packets. Also, since the model is independent of the hop number, and since hops are independent, then the distribution of the queuing delay incurred by a voice stream that travels through N hops is simply the N-fold convolution of the queueing delay incurred by a voice stream over one individual hop.

In order to verify the proposed model, we have simulated a number of scenarios that are based on the multi-hop topology described above. In particular, we consider a hierarchical topology in which the interfering traffic at each of the hops visited by the target voice stream is assumed to have traveled the same number of hops as the target voice stream. In Figure 4, we plot for one such scenario the complementary distribution of the queuing delay incurred by voice packets belonging to the target stream as obtained from simulation, and compare it to that obtained by the model presented above. Consistent with our assumption, we consider in the simulations that the lower priority packets are all MTU-sized, and that the lower-priority queue never empties. As can be seen from the plots, the distribution obtained from the model closely bounds the distribution

[6] Since Ω_N^1 and $-\Omega_N^2$ are both normal distributions with mean 0 and variance σ_N^2, then $\Omega_N^1 - \Omega_N^2$ is a normal distribution with mean 0 and variance $\left(\sigma_N'\right)^2 = 2\sigma_N^2$.

obtained from the simulations; this fact has been observed for a number of different scenarios based on the general topology.

When WRR is used instead of PQ, two additional effects come into play: on one hand, given that the voice traffic is given a share W of the total bandwidth, then the actual bandwidth that is available for voice traffic, c_{voice} could be as low as W (that is, in case the remaining bandwidth is fully utilized by other traffic). In addition, because scheduling is done at the packet granularity, c_{voice} varies around W, which increases traffic variability and consequently delay percentiles. On the other hand, a given voice packet could incur the transmission time of more than one non-voice packet, depending on the number of queues that contend for the link. We intend to obtain delay results that correspond to a lower bound on the maximum delay incurred by voice traffic when serviced by a WRR scheduler. Hence, we assume (1) that the available bandwidth for voice traffic is equal to W at all times, and (2) that voice packets incur a residual transmission time corresponding to one non-voice packet.

V. Choice of Scheduling Scheme

Using the models developed in Section IV for voice delay, we now propose to compare different scheduling schemes that can be used to service the voice queue. We first investigate in Section V-A in detail the effect of the residual transmission time of non-voice packets on voice queueing delay . By the same token, we consider the effect of various network parameters (link utilization, number of hops and link bandwidth) on voice delay. We then compare in Section V-B PQ to competing schemes, and establish its adequacy in support of voice traffic.

A. Residual Transmission Time of Non-Voice packets

In case voice traffic is provided with a separate link (or circuit), it is completely shielded from other traffic on the network. If voice traffic shares the link with other traffic instead, with PQ used to schedule packets, then as found in Section IV, the increase in queuing delay caused by the residual transmission time of non-voice packets is two-fold: on one hand, the queuing delay $w_{M/D/1}$ is significantly larger than $w_{\sum D_i/D/1}$. Except for lightly utilized high speed links (see Section II), $M/D/1$ leads to significantly larger delay values as compared to $\sum D_i/D/1$. [7] On the other hand, the residual transmission time of lower priority packets is typically much larger than $w_{M/D/1}$; in fact, considering a typical MTU of 1500 bytes, the transmission time of an MTU sized packet is around 16 times larger than that of a typical voice packet[8].

Number of hops and link utilization. In Figure 6, we plot the complementary distribution of queuing for both configurations in the case of G.729A, for one, 2 and 5 hops, over T1 links. The first observation is that in all cases, the tail of the delay distribution is still fast-decaying, which confirms the fact that the $(1 - L_{max})^{th}$ delay percentile is indeed much lower than the maximum delay that could be achieved in the network. However, the tail of the complementary delay distribution widens significantly when voice shares the link with other traffic. For a tolerable loss rate L_{max} set to 10^{-5}, the queuing delay percentile incurred by voice flowing alone on the link does not exceed 6 ms even if the path consists of five T1 links; conversely, when voice traffic is given priority over other traffic on the link, then voice delay is close to 20 ms as soon as two T1 links are traversed. In fact, queuing delay becomes very sensitive to the number of hops (e.g. more than 40 ms incurred for 5 T1 links), which is expected, since for $N \leq 10$, the $(1 - L_{max})^{th}$ percentile of the residual transmission time over N hops is close to N times T_{MTU}. Clearly, in this case, the contribution of the residual transmission time of lower priority packets to queuing delay is large. Similar observations apply to other link speeds (10 Mbit/s, 45 Mbit/s, etc.). In Figure 7 we plot the complementary distribution function of queuing delay for the same two configurations in the case of G.729A for 5 T1 hops, and for voice traffic using up to 10%, 50% and 75% of the link bandwidth[9]. The figure reveal that when voice traffic is given priority over

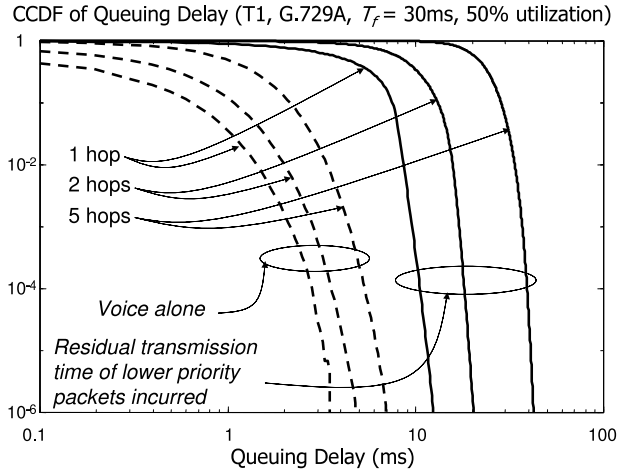

Fig. 6. CCDF of Queuing Delay (T1, G.729A, $T_f = 30$ ms, 50% utilization) for 1, 2 or 5 hops. Two configurations are compared: either voice is transmitted alone on the link, or given high priority over other traffic flowing on the link.

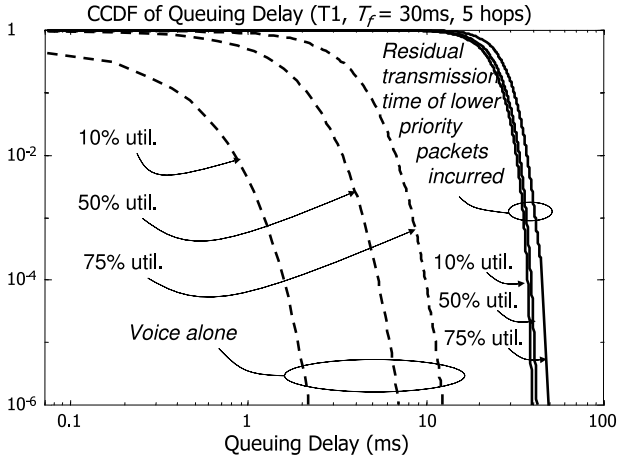

Fig. 7. CCDF of Queuing Delay (T1, G.729A, $T_f = 30$ ms, 5 hops) with voice traffic totaling a maximum utilization of voice 10, 50 and 75%. Two configurations are compared: either voice is transmitted alone on the link, or given high priority over other traffic flowing on the link.

[7] In particular, as the average aggregate utilization approaches 1, $w_{\sum D_i/D/1}$ never exceeds T_f, whereas $w_{M/D/1}$ increases without bound.

[8] That is, a packet generated by G.729A, packetized using T_f= 30 ms and to which a 46 bytes header is appended, for a total of 76 bytes.

[9] When measuring link utilization, we assume a $\sum D_i/D/1$ model, and hence ignore the effect of silence suppression.

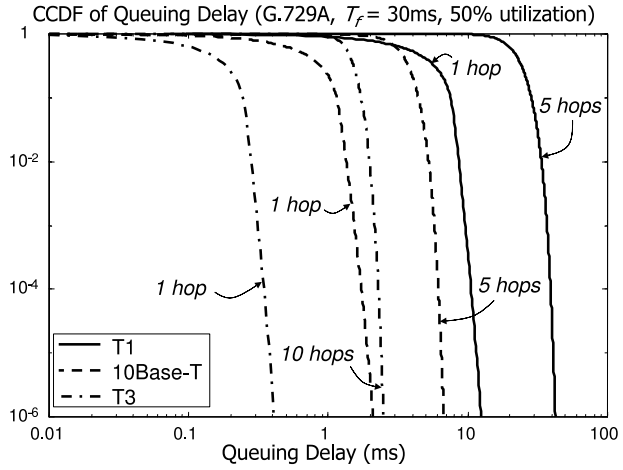

Fig. 8. CCDF of queuing delay (G.729A, $T_f = 30$ ms, 50% utilization) for 1, 5 and 10 hops, and for T1, 10Base-T and T3 links.

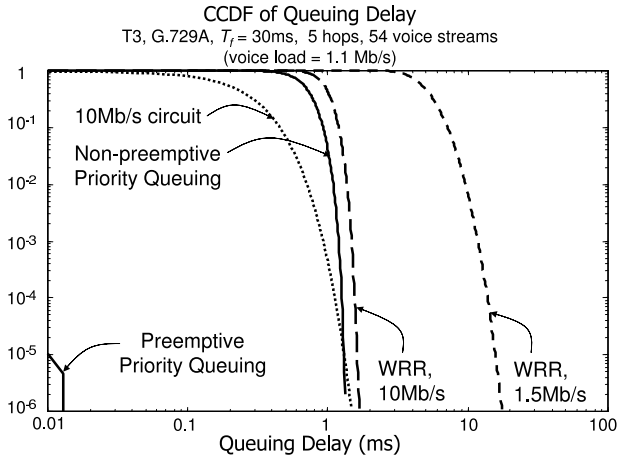

Fig. 9. Complementary delay distribution for voice traffic served using either PQ, WRR or when provided with a separate circuit. (In the case of WRR, the results show a lower bound on the complementary distribution of voice delay for two shares allocated for voice traffic, 1.5 and 10 Mbit/s respectively.)

other traffic on the link, the increase in queuing delay that results from an increase in voice utilization from 10 to 75% becomes less significant, as compared to the total queuing delay incurred. That is, when voice shares the link with other traffic, the characteristics of voice traffic become less important, while the residual transmission time of non-voice traffic becomes the determining factor.

Available Bandwidth. In general, the effect of bandwidth on queuing delay is well known in queuing theory, as average queuing delay for a given utilization is shown to be inversely proportional to the available bandwidth for most queuing systems of interest. Similarly, delay percentiles (as considered in this study) decrease with an increase in the available bandwidth[10]. As queuing delay decreases with the available bandwidth, other delay components do not vary: in particular, propagation delay, formation time and look-ahead delay are incurred by all voice packets in the stream; therefore, when bandwidth is large enough, worrying about queuing delay becomes futile. In Figure 8, we plot the complementary delay distribution of voice traffic for different values for the number of hops in the path and the available bandwidth. One can see from the graph that the queuing delay percentile incurred on voice packets going through 10 T3 hops remains lower than 3 ms while the delay incurred over one and 5 T1 hops already exceeds 10 and 40 ms, respectively, which is on the order of both the formation time and the propagation delay, and constitutes a large proportion of the tolerable end-to-end delay $D_{max} = 100$ ms.

B. PQ Versus Other Scheduling Schemes

The Section V-A above, we showed that PQ leads to adequate delays when the available bandwidth is large. In this section, we look more closely at the appropriate choice of scheduling scheme, and compare in that respect PQ, WRR, and the provision of a separate circuit for voice traffic. We plot in Figure 9 the complementary delay distribution for voice traffic when serviced using preemptive and non-preemptive PQ, and lower bounds of the complementary distribution of voice traffic when serviced using WRR for two values of W (1.5 Mbit/s and 10 Mbit/s, respectively). Note that when W is set to a value that is close to the maximum voice load (1.5Mbit/s versus 1.1Mbit/s), the delay incurred by voice traffic exceeds 10 ms for 5 T3 hops; however, as W is increased to 10 Mbit/s, then the queuing delay incurred by voice traffic decreases to less then 2 ms, which is very close to the value obtained with non-preemptive PQ. In fact, as long as the ratio of the maximum voice load to W is kept low (below 20%), the delay incurred with WRR converges quickly to that obtained with non-preemptive PQ. Also, note that even though W is much larger than the maximum voice load, resources are not wasted since residual bandwidth that is unused by voice traffic can be shared by traffic in the other queues. In addition, WRR has the well-known advantage of making sure that no traffic from a given queue starves traffic from other queues. However, since voice traffic will remain a small portion of Internet traffic (10% of the total Internet traffic expected within 5 years), we suspect that, in most situations, no special precaution will be needed to prevent voice traffic from starving other traffic. Also, note that WRR can lead to delays that are much larger than those shown in this figure: in particular, when multiple queues share the link, then voice packet could incur the delay pertaining to more than one MTU-sized packets. For this reason, WRR is not ideal for the handling of voice traffic.

We now compare non-preemptive PQ to the provision of a circuit for voice traffic. As shown in Figure 9, the penalty that results from having voice traffic incur the residual transmission time of other traffic is such that carving out a circuit of $W = 10$ Mbit/s for voice traffic on a 45 Mbit/s link (thus shielding it totally from other traffic on the link) leads to lower delays than giving priority to voice traffic over the entire 45 Mbit/s bandwidth. However, contrarily to WRR, the difference between the actual voice load and W is wasted in this case. That is, compared to PQ, lower delays can be obtained by either wasting resources (which is undesirable), or by preempting the transmission of non-voice packets. To achieve a delay for voice traffic that is lower than that offered by a non-preemptive PQ

[10]For $M/D/1$, delay percentiles are inversely proportional to the available bandwidth for a given utilization (that is, for a proportionally larger number of streams supported). For $\sum D_i/D/1$, queuing delay percentile also decreases when the available bandwidth is increased, but the two are not exactly inversely proportional.

scheduler, yet avoiding the waste in bandwidth that results from providing voice with a separate circuit, a PQ scheduler can be made to preempt the transmission of lower priority packets, as has been proposed for low speed links[11] [4]. In this case, the delays obtained become as low as if voice traffic were flowing alone on the link. As shown in Figure 9, the use of preemption reduces the delay incurred by voice traffic across 5 hops over T1 links to less than 0.2 ms. In general, the results show that with preemptive PQ, the queuing delay of voice traffic flowing across 5 hops over T1 links does not exceed 11 ms for a link utilization as high as 75% .

Therefore preemptive PQ is indeed the most appropriate scheduling scheme for handling voice traffic over sub-10 Mbit/s links, while non-preemptive PQ is adequate (that is, leads to a queuing delay that is lower than 10 ms) when the link bandwidth exceeds 10 Mbit/s. [12]

VI. Choice of Packetization Scheme

In this section, we look into the effect of packet size on voice delay and bandwidth utilization. We show that in most cases, choosing a packet size dynamically on a per-connection basis only provides a modest benefit as compared to using a fixed packet size for all connections.

Headers corresponding to the various layers of the protocol stack are appended to voice packets before they can be transmitted on the network[13]. Denoting r the rate of the encoded bit stream, and R the rate of the packetized bit stream, we have $R = r + \frac{8H}{T_f}$. In Figure 10, we plot the ratio $\frac{R}{r}$ as a function of T_f for both G.729A and G.711 over point-to-point links. Clearly, the lower the rate of the encoded stream, the larger the overhead, and hence the larger the ratio $\frac{R}{r}$. When T_f =10 ms, $R = 5.6r$ for G.729A; increasing the formation time to 30 ms already decreases this ratio to $\frac{R}{r} = 2.5$, while a formation time as high as 100 ms is needed to decrease the overhead to 50%. Therefore, there is an incentive to use the largest formation time possible given the maximum tolerable delay D_{max}, the propagation delay $P_{tot} = \sum_{h \in Path} P_h$ and the queuing delay Q. In this section, we intent to investigate whether dynamic packetization (using the largest formation time as described above) is beneficial. We show that in most cases, choosing a packet dynamically on a per-connection basis only provides a modest benefit as compared to using a fixed packet size for all connections.

We start with the end-to-end delay budget equation derived in Section III-B (Equation 2). Ignoring the transmission time

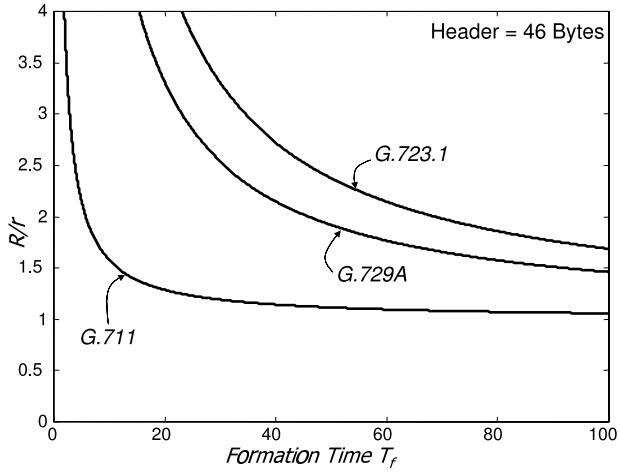

Fig. 10. $\frac{R}{r}$ versus formation time T_f for G.729A and G.711

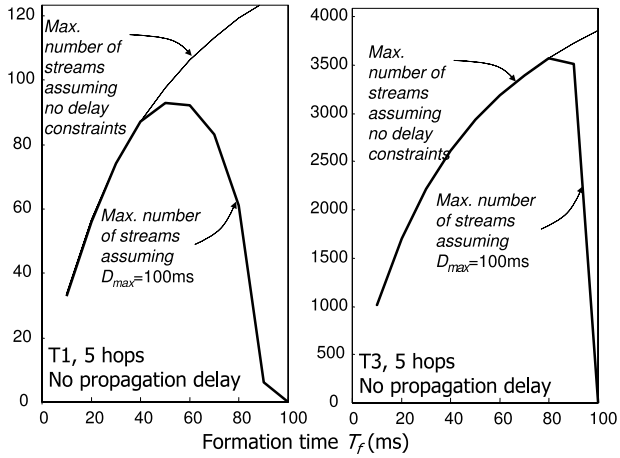

Fig. 11. Maximum number of streams versus T_f on T1 and T3 links, for a 5 hops path and a tolerable end-to-end delay D_{max} = 100 ms. The propagation delay is considered negligible for all voice streams sharing the link.

(which, as stated in Section III-B is, in general negligible compared to other the other delay components), we get, for a given stream,

$$T_f + 2Q_{max} \leq D_{max} - (P_{tot} + l) \quad (4)$$

In case different encoding and packetization schemes are used for each voice stream being served by a network node, T_f and l can differ across them. For simplicity, we consider that all voice streams use the same encoder. We first consider G.729A (that is, l = 5 ms), and then comment on both G.711 and G.723.1. Also, in a typical network node serving voice traffic, different streams are characterized by travel paths having different hop counts and different propagation delays. Before investigating such a scenario, we consider the simple scenario in which all streams on a link travel the same number of hops and have a zero propagation delay. In other words, given a maximum tolerable end-to-end delay D_{max}, the end-to-end delay requirement for voice becomes

$$T_f + 2Q_{max} \leq D_{max} - l$$

For this scenario, we plot in Figure 11 the maximum number

[11] In this case, when a voice packet reaches the head of the high priority queue while a lower priority packet is being transmitted, then the transmission of the data packet is interrupted to allow the transmission of the voice packet; only when the high-priority voice queue is empty again does the transmission of the data packet resume.

[12] This result contradicts the findings of [7], in which PQ was shown to lead to extremely large delay values in the network. The reason for the discrepancy lies in the choice of the measure used in [7], the worst-case delay. This once again confirms the importance of taking advantage of the tolerable packet loss L_{max} in the Internet.

[13] A 12 byte RTP header, an 8 byte UDP header, a 20 byte IP header, and either a 6 byte data-link (e.g., PPP or HDLC) or a 29 byte 802.3 MAC header depending on whether the voice packets are crossing a point-to-point link or an Ethernet LAN, respectively, for a total header size H = 46 or 69 bytes, respectively.

TABLE II
DISTRIBUTION OF PROPAGATION DELAY FOR VARIOUS GEOGRAPHIC LOCATIONS.

Propagation Delay (ms)	0	10	20	30	40	50	60	70	80
Distribution for local areas	0.60	0.09	0.09	0.09	0.09	0.01	0.01	0.01	0.01
Distribution for wide areas	0	0.20	0.20	0.20	0.20	0.05	0.05	0.05	0.05
Distribution for trans-oceanic links	0	0	0	0	0	0	0.33	0.33	0.34

of streams as a function of T_f on T1 and T3 links for a maximum tolerable end-to-end delay $D_{max} = 100$ ms. The number of streams increases at first with T_f, as a result of the reduction in bandwidth requirement per stream; however, as the formation time approaches $D_{max} - l$, the total utilization on the link must decrease to keep queueing delay low enough so that end-to-end delay remains below D_{max}; this in turn leads to a decrease in the number of streams. (Clearly, for $T_f \geq D_{max}$, the number of supported streams is zero.) In this case, the number of streams supported increases significantly when the formation time is chosen optimally.

Even though the analysis conducted above suggests a significant benefit in choosing the largest formation time possible, given the end-to-end delay possible, the scenario considered is nevertheless unrealistic, as it assumes that the propagation delay of all streams on the link is negligible. Hereafter, we consider a more realistic scenario in which the propagation delays corresponding to the streams on the link are random. As shown in Table II, we consider various propagation delay distributions, depending on the geographical location of the link of interest. If the link belongs to a local area network, then most of the calls have a relatively low propagation delay. On the other end of the spectrum, all connections belonging to a trans-oceanic link have a large propagation delay. (The distribution of propagation delay for streams flowing in a wide area network lies somewhere between these two extremes.) In each of these cases, we assume that the formation time used for the packetization of each stream is tailored to the propagation and queuing delay it incurs. In order to capture the benefit obtained from dynamically choosing the formation time of a per-connection basis, we plot in Figure 12 the number of streams supported as a function of the maximum allowable formation time. Clearly, as the proportion of streams having a large propagation delay increases, the benefit of tailoring the choice of formation time for each connection decreases significantly: the plots in Figure 12 show that per-flow custom packetization allows the number of streams to increase from around 1000 to more than 2500 on T1 links in local areas; however, the number of streams supported increases to around 2000 in wide areas under the same conditions, while trans-oceanic links do not benefit at all from custom packetization. Also, most of the gain is achieved by increasing the propagation delay from 10 to 30 ms. One can argue that an increase in delay by 20 ms is acceptable, given the significant benefit obtained from the increase. As shown in Figure 13 (in which we

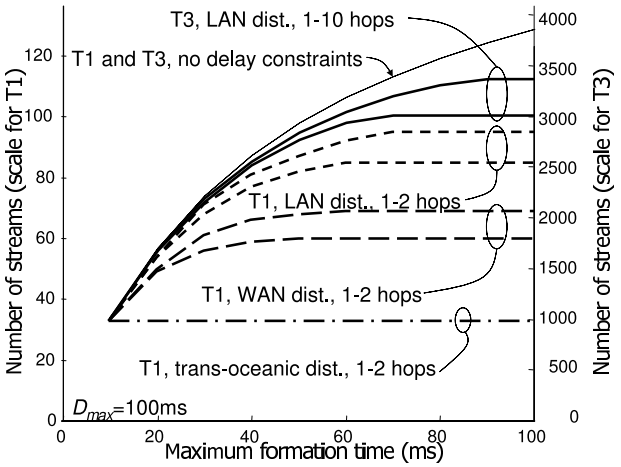

Fig. 12. Maximum number of streams versus the maximum formation time used by the encoder for T1 and T3 links (10 ms minimum formation time allowed).

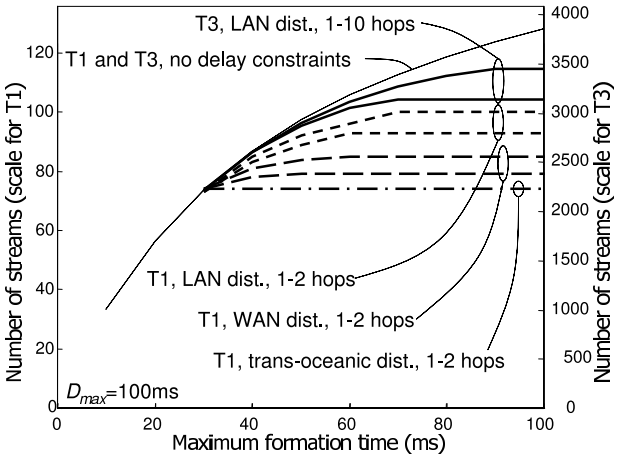

Fig. 13. Maximum number of streams versus the maximum formation time used by the encoder for T1 and T3 (30 ms minimum formation time allowed).

plot the same set of graphs as in Figure 12, but with restricting the minimum formation time to 30 ms), the benefit from custom packetization beyond $T_f = 30$ ms appears low: in WANs, bandwidth is usually largely available and propagation delay is large; therefore, this case is the least interesting, since the incentive for bandwidth reduction is low, and the actual potential for bandwidth reduction is low too. On the other hand, the incentive for bandwidth reduction is the largest in the case of trans-oceanic links, where the bandwidth is scare, and the cost of installation high; unfortunately, as seen in Figure 13, the large propagation delays on the links dictate the choice of low formation times. Finally, the potential for bandwidth reduction is largest in local networks; even when restricting the minimum formation time to 30 ms, the increase in number of streams is still significant (from around 2200 to more than 3000 on T3 links, around 75 to more than 90 on T1 links) when formation time is chosen on a per-connection basis. Still, one could argue that with the advance in fiber optics technology, available bandwidth in local networks will also be large. However, in the situation in which T1 and T3 links are used to aggregate voice traffic in local LANs, dynamic packetization on a per-connection basis could allow the support of a factor of 20% and 36% more streams on the T1 and T3 links,

respectively; conversely, for a given number of streams, the gain can allow the use of fewer T3 links, yielding a significant cost decrease.

Finally, note that when voice traffic is not alone on the link but shares it with other traffic using either PQ or WRR, then the queuing delay is significantly larger, reducing in turn the number of streams supported and the potential gain provided by per-flow custom packetization. In summary, using $T_f = 30$ ms leads to a negligible waste in resources, and constitutes a good compromise in terms of delay and bandwidth utilization. Repeating the same study for G.711 and G.723.1 leads to the same observations. The recommended formation time for these encoders is 10 ms (that is, the recommended packet size is 80 bytes) and 30 ms (that is, the recommended packet size is 20 bytes), respectively.

VII. Conclusion

In this paper, we have focused on networks where a separate queue for voice traffic is provided. In this context, we have first described the particularity of voice traffic as compared to other traffic on the network. We then assessed voice delay, and looked at the effect of network parameters on the delay incurred by voice traffic; in particular, we focused on the effect of the residual transmission time of non-voice packets of voice delay, and showed that it constitutes the largest portion of voice delay in case PQ is used to schedule voice traffic. We also showed the importance of bandwidth to reduce the delay percentile incurred by voice in a network, and concluded that network delay becomes a negligible portion of end-to-end delay in case available bandwidth exceeds 10 Mbit/s. We then compared different scheduling schemes (that is, PQ, WRR and the provision of different circuit for voice traffic), and showed that PQ leads to the best compromise between bandwidth utilization and delay minimization, as long as the preemption of non-voice packets is implemented on sub-10 Mbit/s links. Finally, we studied the effect of packet size on voice delay and bandwidth utilization, and showed that a packet formation time of 30 ms for G.729A and G.723.1 (that is, a packet size of 30 and 20 bytes, respectively) and 10 ms for G.711 (that is, a packet size of 80 bytes) constitute a good compromise between low delay and efficient network utilization.

References

[1] D. Awduche, Y. Rekhter, J. Drake, and R. Coltun, "Multi-Protocol Lambda Switching: Combining MPLS Traffic Engineering Control With Optical Crossconnects, " IETF working draft, work in progress, Nov. 1999.

[2] Y. Bernet, J. Binder, S. Blake, M. Carlson, B. Carpenter, S. Keshav, E. Davies, B. Ohman, D. Verma, Z. Wang, and W. Weiss, "A Framework for Differentiated Services," IETF working draft <draft-ietf-diffserv-framework-02.txt>, work in progress, Feb. 1999.

[3] J.-C. Bolot, "End-to-End Packet Delay and Loss Behavior in the Internet," *Proceedings of SIGCOMM'93*, Sept. 93.

[4] C. Bormann, "PPP in a Real-time Oriented HDLC-like Framing," RFC 2687, Sept. 1999.

[5] P. T. Brady, "A Statistical Analysis of On-Off Patterns in 16 Conversations," *Bell Syst. Tech. Journal*, Vol. 47, pp. 73-91, Jan. 1968.

[6] K. Bullingtom and J. Fraser, "Engineering Aspects of TASI," *Bell Syst. Tech. Journal*, Vol. 38, March 1959.

[7] A. Charny and J.-Y Le Boudec, "Delay Bounds in a Network with Aggregate Scheduling," EPFL-DSC Technical Report DSC2000/022, April 2000.

[8] J. Daigle and J. Langford, "Models for analysis of packet voice communications systems," *IEEE Journal on Selected Areas in Communications*, Vol. SAC-4, No. 6, pp. 847-55, Sept. 1986.

[9] A. Eckberg, "The single server queue with periodic arrival process and deterministic service times," *IEEE Transactions on Communications*, Vol. COM-27, No. 3, pp. 556-62, March 1979.

[10] Recommendation G.711, "Pulse Code Modulation (PCM) of Voice Frequencies," ITU, Nov. 1988.

[11] Recommendation G.726, "40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)," ITU, Dec. 1990.

[12] Recommendation G.723.1, "Speech Coders: Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s," ITU, March 1996.

[13] Annex A to Recommendation G.729, "Coding of Speech at 8kbit/s using Conjugate Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)," Annex A: "Reduced Complexity 8 kbit/s CS-ACELP Speech Codec", ITU, Nov. 1996.

[14] S. Golestani, "A Self-Clocked Fair Queueing Scheme for Broadband Applications," *Proceedings of IEEE INFOCOMM'94*, pp. 636-646, June 1994.

[15] P. M. Gopal and B. Kadaba, "A Simulation Study of Network Delay for Packetized Voice," in *Proceedings of GLOBECOM'86*, Dec. 1986.

[16] P. Goyal, A. Greenberg, C. Kalmanek, W. Marshall, P. Mishra, D. Nortz, and K. Ramakrishnan, "Integration of Call Signaling and Resource Management for IP Telephony," *IEEE Network*, pp. 24-32, May/June 1999.

[17] J. Gruber and L. Strawczynski, "Subjective Effects of Variable Delay in Speech Clipping in Dynamically Managed Voice Systems," *IEEE Transactions on Communications*, Vol. COM-33, No. 8, Aug. 1985.

[18] IEEE 802.1D, "Standard for Local and Metropolitan Area Networks: Media Access Control (MAC) Bridges," 1990.

[19] V. Jacobson, K. Nichols, and K. Poduri, "Expedited Forwarding PHB," RFC 2598, June 1999.

[20] V. Jacobson, K. Nichols, and K. Poduri, "The 'Virtual Wire' Per-Domain Behavior," Internet working draft <draft-ietf-diffserv-pdb-vw-00.txt> , work in progress, July 2000.

[21] Y. C. Jenq, "Approximations for Packetized Voice Traffic in Statistical Multiplexer," in *Proceedings of INFOCOM'84*, April 1984.

[22] M. Karam, F. Tobagi, "On Traffic Types and Service Classes in the Internet," in *Proceedings of GLOBCOM'2000*, Dec. 2000.

[23] M. Karol and M. Hluchyi, "Using a Packet Switch for Circuit-Switched Traffic: A Queueing System with Periodic Input Traffic," *IEEE Transactions on Communications*, Vol. 37, No. 6, pp. 623-625, June 1989.

[24] M. Katevenis, S. Sidiropoulos, and C. Courcoubetis, "Weighted Round-Robin Cell Multiplexing in a General-Purpose ATM Switch Chip", *IEEE Journal on Selected Areas in Communications*, Vol. 9, No. 8, pp. 1265-1279, Oct.1991.

[25] L. Kleinrock, "Queuing Systems-Vol. II: Computer Applications," New York; Wiley, 1975.

[26] M. Mandjes, K. van der Wal, R. Kooij, and H. Bastiaansen, in *Proceedings of the 7th IFIP ATM and IP Workshop*, June 1999.

[27] G. Mercankosk, "The 'Virtual Wire' Per Domain Behavior - Analysis and Extensions," IETF working draft <draft-mercankosk-diffserv-pdb-vw-00.txt>, work in progress, July 2000.

[28] S. Moon, J. Kurose, P. Skelly, and D. Towsley, "Correlation of Packet Delay and Loss in the Internet," Technical Report 98-11, Department of Computer Science, University of Massachussetts, Amherst, Jan. 1998.

[29] G. Ramamurthy, B. Sengupta, "Delay Analysis of a Packet Voice Multiplexer by the $\sum D_i/D/1$ Queue," *IEEE Transactions on Communications*, Vol. 39, No. 7, pp. 1107-1114, July 1991.

[30] M. Shreddhar and G. Varghese, "Efficient Fair Queueing Using Deficit Round Robin," *Proceedings of SIGCOMM'95*, Sept. 1995.

[31] K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data," *IEEE Journal on Selected Areas in Communications*, Vol. SAC-4, No. 6, pp. 833-46, Sept. 1986.

[32] L. Zhang, "Virtual Clock: A New Traffic Control Algorithm for Packet Switching Networks," *Proceedings of ACM SIGCOMM'90*, Sept. 1990.