VoIP Performance on Differentiated Services Enabled Network *

Jogesh. K. Muppala, Terdsak Bancherdvanich and Anurag Tyagi
Department of Computer Science
The Hong Kong University of Science and Technology,
Kowloon, Hong Kong
{muppala, terdsak, anurag}@cs.ust.hk

Abstract

In this paper we study the performance of Voice over Internet Protocol (VoIP) traffic aggregates over Differentiated Services (Diffserv) enabled network using Expedited Forwarding (EF) per hop behavior (PHB). We compare the delay and jitter performance of the VoIP traffic generated by different standard voice codec algorithms, both under Diffserv with EF PHB and with best-effort service. Both homogenous and heterogenous voice traffic aggregates are considered. Our results show that the use of EF yields very good performance improvement for voice traffic compared to best-effort. The improvement is greatest for high coding rate algorithms like G.711 than lower coding rate algorithms like G.723.1. For heterogenous traffic aggregates, the traffic from higher bit rate codecs obtains better performance compared to lower bit rate codecs.

1. Introduction

Voice over Internet Protocol (VoIP) has recently been receiving a lot of attention both in the academia and industry because of its great potential [2, 4]. VoIP traffic requires a minimum quality of service (QoS) from the network to meet its stringent bandwidth, delay and jitter requirements [3]. The current *best-effort* Internet service does not differentiate between packets of different applications. Differentiated services (Diffserv) [1] is designed to provide scalable QoS support within the Internet. For an end-to-end flow, the overall result of this approach is less definite because the actual QoS depends on the momentary load within the network. Hence we need simulation studies to investigate the behavior of services built using Diffserv.

Recently some studies of VoIP performance over Diffserv have appeared in the literature. In [7] the authors examine the performance of voice over a network which supports premium and assured services. They show conditions under which the assured service performance deteriorates while the premium performance is not significantly affected. In [5] the authors compare the performance of EF supported by using either priority round robin (PRR) or weighted round robin (WRR) for traffic generated by CBR sources. In [10] the authors extend the study of [5] to VoIP. In [3] the authors show that a properly configured link scheduling policy can meet the stringent requirements of voice. In [8] we examined VoIP performance over EF which is supported using class based queueing (CBQ).

In this paper we examine: (a) the improvement in VoIP performance using EF compared to best-effort, (b) the performance of homogenous voice aggregates (all the voice sources use the same codec) carried over either EF or best-effort, and (c) the performance and interaction effects of heterogenous voice aggregates (voice packets emerging from sources using one of two different voice codecs are mixed) being carried on the same link and sharing the bandwidth allocated to EF. We notice a significant improvement in VoIP performance when it is carried over EF than best-effort. The improvement is the largest for high bit rate codecs like G.711 than low bit rate codecs like G.723.1. For heterogenous traffic aggregates, the traffic from sources using a high bit rate codec get better performance than those from low bit rate codec sources.

2. Differentiated Services

Diffserv is designed to be a scalable mechanism to be deployed in the core of the Internet. Traffic entering a DS network is classified and conditioned at the edges, then marked and assigned to different behavior aggregates (BA) by setting the value of Diffserv Code Point (DSCP) [1]. Routers look at the DSCP and provide the corresponding PHB to the flow. PHBs are a means of allocating network resources to different traffic aggregates.

The EF PHB requires that the departure rate of the aggregate's packets must equal or exceed the aggregate's maxi-

*This research was supported in part by Hong Kong RGC Earmarked Research Grant No. HKUST6099/99E.

0-7695-0777-8/00 $10.00 © 2000 IEEE

mum arrival rate. It receives a configurable rate independent of any other traffic attempting to transit a node. A number of existing queue scheduling mechanisms like priority queuing (PQ) with the highest priority and a rate policer, weighted round robin (WRR) scheduling, or class based queuing (CBQ) can be used to support EF.

3. VoIP

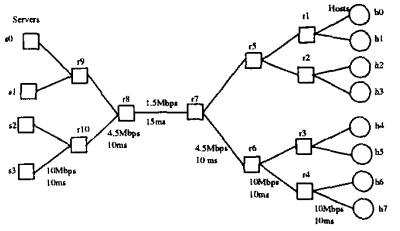
The steps involved in delivering VoIP include: sampling, digitization & encoding, encapsulation, transmission, decoding, buffering and play-out. ITU-T standard voice codec algorithms include [8]: G.711 at 64 Kbps using PCM, G.726 using ADPCM at 40, 32, 24, and 16 Kbps. CELP is used for G.728 at 16 Kbps, G.729/G.729A at 8 Kbps. G.723.1, based on MP-MLQ technology with two transmission rates i.e., 5.3 and 6.3 Kbps, generally provides good speech quality. Table 1 summarizes the details of these algorithms.

VoIP QoS issues unique to packet networks are delay, jitter and loss. Delay gives rise to echo and talker overlap problems. Echo is a significant when round-trip delays are greater than 50 ms, while the talker overlap problem becomes significant when the one-way delay is greater than 400 ms. Since packets experience varying delays in the network, inter-packet time on the receiver side is not constant even if it is so on the sender's side. A play-out buffer is employed to filter out the jitter, which introduces further delay. A late or lost packet will affect the quality of the voice received. To recover from lost packets, voice may be interpolated from the last received packet to fill the place of the lost packet.

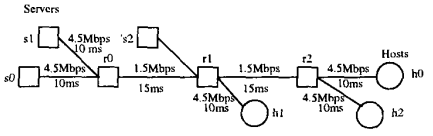
4. Network configuration

The simulation studies presented in this paper were conducted using NS-2 [9] with two different network topologies. The dumbbell topology (Figure 1(a)) is constructed to provide a six-hop path between each source and destination, with a single shared 1.5 Mbps bottleneck link in the middle. The linear topology (Figure 1(b)) consists of core links of 1.5 Mbps bandwidth which acts as the bottleneck links. Similar network topologies have been used for simulation in [3, 5, 7, 10].

The EF PHB is supported using WRR with two queues, one for EF and the other for best-effort (BE). We allocate 30 percent and 70 percent of the link bandwidth respectively to these two queues. The aggregate EF traffic produced by all VoIP sources is restricted to be 450 Kbps, which is 30 percent of the bottleneck link capacity of 1.5 Mbps. The remaining capacity is filled with best-effort traffic. The best-effort traffic is handled at the routers using a random early



(a) Dumbbell



(b) Linear

Figure 1. Network Topologies

detection (RED) queue with a queue size of 200 and the RED parameters [45,90,0.002,0.05] ($[min-th, max-th, q_{wt}, P_{max}]$). The EF queue is implemented as a simple first-in first-out (FIFO) queue.

First, we consider homogeneous flows where voice traffic aggregates consists of packets generated by sources which use the same codec algorithm. Here we study the traffic generated by various codecs listed in Table 1. The voice sources are modelled as constant bit rate (CBR) sources. In all the experiments, if EF is used to transport the voice traffic aggregates, then the EF traffic flows are generated to fill the subscribed rate of the EF class. Since the share of the bandwidth on the bottleneck link for EF class is fixed, the number of flows in the voice aggregate will change depending on the codec used (see Table 1). One reason for considering a fixed share of the bandwidth for voice traffic aggregates is that link bandwidth provisioning in the core of the network is usually done in bandwidth chunks (voice trunks). We further assume that the phases of the various voice sources are *randomized* [3]. To make a fair comparison, the same characteristics are used when the voice traffic aggregates use best-effort. The total traffic generated by the voice traffic aggregates even when best-effort is used, is limited to 30% of the bottleneck link bandwidth, i.e., 450 Kbps. In this case, both the voice traffic and the background TCP traffic share the entire link bandwidth.

Next, we consider heterogeneous flows where voice traffic aggregate consists of packets generated by voice sources, where some sources use G.711 codec while other sources

Table 1. Characteristics of Different Voice Codecs

Codec	G.711	G.723.1	G.726-32	G.729	G.729A
Coding speed (Kbps)	64	5.3/6.3	32	8	8
Frame size (ms)	20	30	20	10	10
Processing Delay (ms)	20	30	20	10	10
Lookahead Delay (ms)	0	7.5	0	5	5
DSP MIPS	0.34	16	14	20	10.5
Payload (bytes)	160	20/24	80	20	20
Number of flows	7	84/71	14	56	56
Subscribed Rate packet time (ms)	20	30.2/30.5	20	20	20

use one of the other codec algorithms Packets from these heterogeneous flows are carried together in one voice aggregate. In all the experiments, EF traffic flows are generated to fill the subscribed rate of the EF class, with traffic generated by sources using G.711 occupying a part of the subscribed bandwidth (varying from 15% to 85%) and the remaining bandwidth is filled by sources using a different codec algorithm. When best-effort is used for all the traffic, the same conditions on the voice traffic are retained. In all the cases, the background traffic is generated by 12 long-term TCP sources that use the Tahoe variation of TCP with a packet size of 576 bytes.

We consider two metrics, namely delay and jitter, to characterize the performance of the VoIP traffic flows. If d_i denotes the time of data.request at the source and a_i denotes the time of data.indication at the destination, then delay D_i may be calculated as $D_i = a_i - d_i$ [6]. We compute the jitter of a packet as $J_i = |D_i - E[D_i]|$ [6], where $E[D_i]$ represents the expected value of the delay. We chose this definition of the jitter to highlight the large variation in delay that can be experienced when using best-effort service. The jitter definition, as used in earlier studies of VoIP [5, 7, 10], considered the absolute value of the difference between the delay of two adjacent packets. This definition yielded very optimistic figures for best-effort traffic in our case. However the jitter definition that measures the difference in delay of adjacent packets ignores the correlation between them. The results are presented in terms of delay and jitter percentiles.

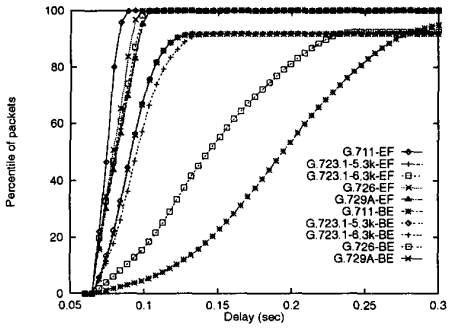
5. Performance Results

First, we consider homogeneous VoIP traffic aggregates. The results of the experiments for the dumbbell and linear topology are presented in Figure 2 and 3 respectively.

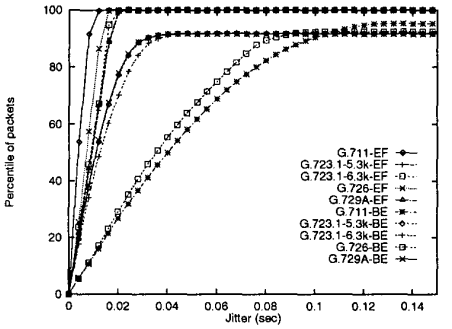
The end-to-end delay for all flows is lower bounded by the end-to-end propagation delay for the two topologies. In the two example topologies considered, this delay is 65 ms for the dumbbell topology, and 50 ms for the linear topology. Using EF for the voice traffic aggregates, we are able to control only the queueing delay experienced by the voice packets in all the routers. Thus in all the results presented, the delay curves start at 65 ms for the dumbbell and 50 ms for the linear topology.

From the delay and jitter characteristics in Figure 2 and Figure 3, we note that for any codec, the delays and jitter experienced by the voice packets improves dramatically when we switch from best-effort to EF for the voice aggregates. The performance improvement is the greatest for the high coding rate codecs like G.711. Furthermore, when we compare the performance of all the codecs using the EF, we notice that the higher coding rate codecs like G.711 yield better performance than the lower coding rate codecs like G.723.1. This is opposite to the trend observed when all the voice sources use best-effort.

The results presented in this paper assumes that the fraction of the bottleneck link bandwidth allocated to the voice traffic aggregate remains constant. Thus, different codecs with their different coding rates will result in different number of flows required to fill this bandwidth. In the experiments, the link bandwidth allocated to voice traffic is 450 Kbps (30% of bottleneck link bandwidth). Therefore the number of flows range from 7 for G.711 to 71/84 for G.723.1. Furthermore, the packet sizes are different for different codecs. Thus the granularity (measured in terms of the time required to transmit a packet of a flow) of demand on the link bandwidth varies. Given these conditions, the performance behavior observed can be explained from the queueing perspective. Recall that we are using WRR for scheduling the link. Therefore, the non-EF traffic is guaranteed its fair share of the link bandwidth. When the voice packets are small, and the number of flows are large, queueing of the packets is caused for two different reasons: (a) waiting caused by other voice packets already in the EF queue, and (b) waiting for access to the link once the packet reaches the head of the EF queue. Given the fixed link bandwidth share for EF, the G.711 sources which generate large packets (160 bytes, excluding headers) and has lower number of flows, occupies the link for a longer period of time, once the packet transmission is scheduled. Thus, interference from the TCP packets is less. On the



(a) Delay



(b) Jitter

Figure 2. Performance of Homogeneous Flows (Dumbbell Topology)

other hand, G.723.1 sources with a smaller packet size and larger number of flows, encounter greater interference both from packets within the voice flow aggregate, and also from the competing TCP packets for access to the link. A similar observation that EF packets may have to wait for some time when non-EF queue is being served was also made in [10].

These two factors, the packet size of the flow and the number of flows, are working at cross purposes. The general trend observed in [7, 10] that with increase in the number of flows there is corresponding deterioration of the performance of voice flows is still noticeable in our results. The decrease in the packet size along with the reduction in the bit rate of the voice codecs is unable to offset the effect of the corresponding increase in the number of flows to keep the overall bandwidth for the voice flows constant. We also examined the case where we fixed the codec and changed the number of voice flows. Our results showed the general trend of deterioration in performance with increase in the number of flows, concurring with the observations made in [7, 10].

From Figures 2 and 3 we also notice that the best-effort curves do not reach upto 100% but to around 90%. This is because we consider packet losses in computing the overall delay. For all packets that are lost, the corresponding delay is set to ∞. Therefore, the delay distributions plotted in the figures are all defective distributions (have a non-zero mass at ∞). Recall that the routers implement RED queues for best-effort traffic. Therefore packet drops become significant when the bottleneck link reaches congestion. With EF enabled, the voice packets are serviced in a separate queue, and hence suffer no packet drops. We observe that the simulation results from both the topologies exhibit similar trends, except that the absolute values of jitter and delay are different because of the difference in the number of hops and link delays of the two topologies.

Next we consider heterogeneous VoIP flows in an aggregate, where some sources use G.711 codec while others use one of the other codecs. In this paper, we present results for the cases where G.711 traffic is around 57% (256 Kbps of the 450 Kbps). The results of the experiments are presented in Figure 4 for the dumbbell topology. This is representative of the general behavior observed for a different percentage of G.711 and other codec traffic. We still notice that with best-effort, sources with high bit rate codecs perform worse than those sources which use low bit rate codecs. When EF is used for the voice aggregate, the trend is reversed.

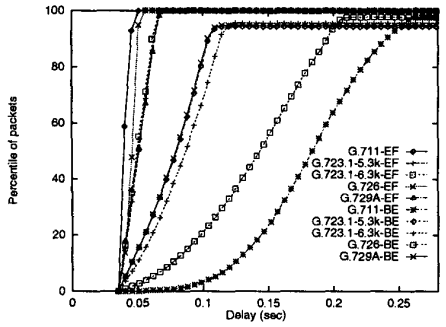
6. Conclusions

We presented a simulation study of VoIP performance over Diffserv using EF or best-effort. We notice a significant improvement in VoIP performance when it is carried over EF, compared to best-effort. The improvement is the largest for high bit rate codecs like G.711 than low bit rate codecs like G.723.1. For heterogenous traffic aggregates, the traffic from high bit rate codec sources get better performance than those from low bit rate codec sources. We are currently investigating other scheduling algorithms like PRR and CBQ to support EF for voice traffic. We are also extending the work to consider on-off models for voice traffic sources.

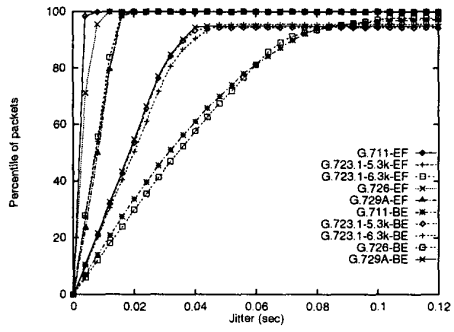
References

[1] D. Black et al., An Architecture for Differentiated Services, *IETF RFC 2475*, Dec. 1998.

[2] *Computer Networks, Special Issue on Internet Telephony*, Vol. 31, No. 3, Feb. 1999.

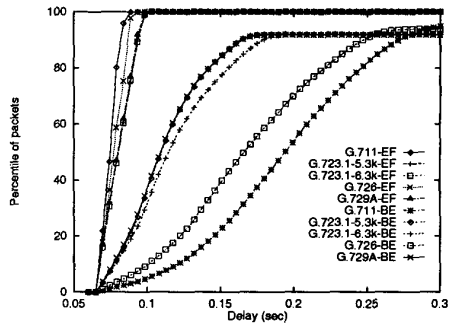


(a) Delay

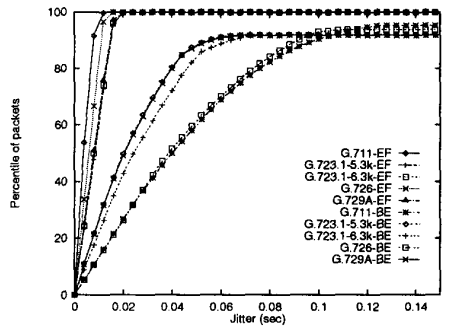


(b) Jitter

Figure 3. Performance of Homogeneous Flows (Linear Topology)



(a) Delay



(b) Jitter

Figure 4. Performance of Heterogeneous Flows (Dumbbell Topology)

[3] P. Goyal et al., Integration of Call Signaling and Resource Management for IP Telephony, *IEEE Network,*, Vol. 13, No. 3, May/June 1999, pp. 24–32.

[4] *IEEE Network, Special Issue on Internet Telephony,* Vol. 13, No. 3, May/Jun. 1999.

[5] V. Jacobson, K. Nichols and K. Poduri, An Expedited Forwarding PHB, *IETF RFC 2598,* Jun. 1999.

[6] H. Knoche and H. de Meer, QoS Parameters: A Comparative Study for Mapping Purposes, *Tech. Rept., Computer Science Department, University of Hamburg,* August 1998.

[7] H. Naser, A. Leon-Garcia and O. Aboul-Magd, Voice over Differentiated Services, *Internet Draft,* draft-naser-voice-diffserv-eval-00.txt, Dec. 1998.

[8] A. Tyagi, J. K. Muppala and H. de Meer, VoIP Support on Differentiated Services using Expedited Forwarding, *Proc. IPCCC 2000,* Phoenix, AZ, USA, Feb. 2000, pp. 574–580.

[9] UCB, LBNL, VINT Network Simulator - NS http://www-mash.cs.berkeley.edu/ns/ns.html

[10] A. Ziviani, J. F. de Rezende and O. C. M. B. Duarte, Towards a Differentiated Services Support for Voice Traffic, *Proc. Globecom'99,* IEEE Comp. Soc. Press, Dec. 1999.