Anatomy of Delay for Voice Service in NGN

Hoon Lee* and Yong-Chang Baek **
*Changwon National University, Changwon, Korea, 641-773
** KT Technology Headquarters, Bundang, Korea
Corresponding E-mail: hoony@changwon.ac.kr

Abstract

In this paper we propose a method for the evaluation of the quality of service for VoIP services in NGN. Specifically, let us anatomize the elements of delay of a voice connection in the network in an end-to-end manner and investigate expected value at each point. We extract the delay time in each element in the network such as gateway, network node, and terminal equipment, and estimate an upper bound for the tolerable delay in each element.

Keywords: NGN, VoIP, QoS, End-to End delay, Network anatomy

1. Introduction

Recently Internet is undergoing a rapid change toward the next-generation network (NGN). NGN can provide the customers with real time services in addition to the current data services in a single framework of IP network. NGN service includes the voice, data, video, streaming media and Internet TV services via several access networks such as PSTN, IMT-2000 and Ethernet. Among them, voice service is considered to be the most probable service that is introduced into NGN in the immediate future. As such, a quantitative evaluation of the quality of service for the VoIP service such as the delay performance along the end-to-end path is prerequisite before implementation of the service [5]. However, we could find little work in this field. An approach is appeared in the standardization activity of ITU-T in [3]. This work is more embodied realization of [3] by taking into account more realistic network environment.

This paper is composed as follows: In Section 2, network architecture for VoIP services over NGN is described. In Section 3, an anatomy for the network architecture for VoIP service over NGN is carried out. In Section 4, we present an example for a typical decomposition of an end-to-end delay to a set of elementary delay in a network. Finally in Section 5, we summarize the work.

2. VoIP Service in NGN

It is assumed that the traditional voice service from PSTN is migrated into NGN by employing AGW (access gateway) or TGW (trunk gateway) for connecting the Telephone users or the PSTN between the PSTN and the IP networks. TGW/AGW is connected to an NGN backbone network via Network access server (NAS), which is an access router.

Fig.1 illustrates the VoIP service architecture for the NGN network [2]. On the other hand, voice applications from PCs connected to LANs can also access NGN via NAS. In any way, packets from voice sources are transferred to the receivers via IP cloud, which is the backbone network of NGN.

As to the voice service provided by heterogeneous network environments, in which a voice signal traverses the access network of traditional PSTN network or LAN and the IP backbone network, packet delay is one of the most critical parameters of QoS. Packet transfer delay is defined as an upper bound on the mean value of end-to-end delay of an IP packet from an ingress point to an egress point of network for a flow [3]. In a real operating network, individual packets may experience delay that exceeds this bound. However, the average packet transfer delay should normally be less than the upper bound [3].

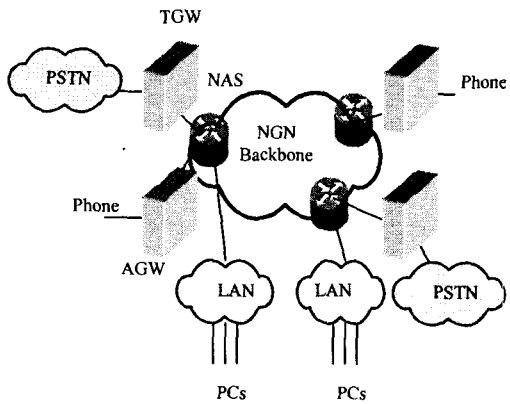

Fig.1. Architecture of VoIP over NGN

In order to guarantee a predefined delay to packets in an end-to-end manner for a voice under the heterogeneous network environments, a detailed anatomy of the element of the network which imposes delay to a voice signal is prerequisite, because delays of voice packets from LAN or PSTN and IP network will be different.

3. Anatomy of End-to-End delay in VoIP

As we could see from Fig.1, there exist two ways for providing VoIP services in NGN. Traditional black phone from PSTN users are connected to IP network via PSTN and trunk gateway (TGW), whereas new IP phone such as SIP phone by PC can be connected to IP network via access gateway (AGW). Fig.2 illustrates a reference path and corresponding QoS region for a VoIP connection in which two different ways are illustrated in the transfer of packets, which corresponds to an abstracted diagram of Fig.1 in Section 2.

In order to support a VoIP with satisfactory service, packet delay, delay jitter and loss rate have to be defined within tolerable

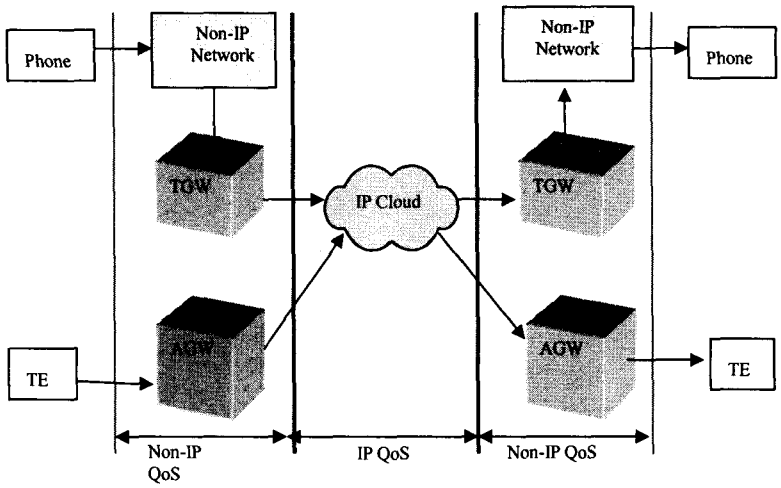

Fig.2. Anatomy of QoS area in NGN.

boundary. Among them, let us consider the packet delay in this work. The delay of a voice is usually defined to be mouth-to-ear (and so an end-to end) delay, which is the total time from the moment a voice sound is produced at one end by the speaker to the moment it is heard at the other end of the receiver [3].

Referring to Fig.2, the end-to-end (E2E) delay of a connection is divided into two parts: the endpoint delay (incurred at end terminal, access gateways in both sending and receiving part or at trunk gateways) at non-IP network and the network delay at IP network cloud.

Let us list the process of voice services and corresponding delay elements [6]. First, the analog voice signal is encoded into digital signal, incurring an encoding delay, which is composed of look-ahead delay, the processing delay, and the sum of frame size:$D_{Enc}=D_{LA} + D_{Proc\text{-}E} + D_{Fr}$, where D_{LA} , $D_{Proc\text{-}E}$, and D_{Fr} is the delay due to look-ahead, processing, and framing, respectively. Table1 summarizes the typical value for the components of D_{Enc} [6].

Table 1. Typical encoding delay.

Encoding scheme	G.711	G.729A	G.723.1
Frame size	125μs	10ms	30ms
Look ahead time	0	5ms	7.5ms
Processing delay	0	10ms	30ms

Second, the encoded bit-stream is packetized, incurring a packetization delay, which is a function of the number of frames k included in a packet: $D_{Pkt}=(k-1)F$, where F is the frame size.

Third, the packetized voice data are fed into the network, incurring a network delay D_{Net}, and inside a network each packet experiences a transmission delay D_{Tx}, buffering delay D_{Qu}, and propagation delay D_{Prop}, along the source-destination path of the network, where $D_{Net} = D_{Tx} + D_{Qu} + D_{Prop}$.

Finally, the packet arrives at the receiver, where it experiences a delay for preparing the playback of the signal, incurring D_{Rec}. For playback, three steps are required. Packet is buffered at a playback buffer, de-packetized, and finally it is decoded into an analog signal. At each step, playback delay D_{PB}, processing delay $D_{Proc\text{-}D}$ and decoding delay D_{Dec} is incurred. Therefore, $D_{Rec} = D_{PB} + D_{Proc\text{-}D} + D_{Dec}$.

One more delay that may be incurred at the receiving side is the PLC (packet loss concealment) time D_{PLC}, which is the time taken to conceal the packet loss. If we summarize the above discussion, we can obtain the following formula for the total end-to-end delay D_{Tot}, of a voice connection:

$$D_{Tot} = D_{Enc} + D_{Pkt} + D_{Net} + D_{Rec} + D_{PLC}. \quad (1)$$

If we list all the components, we have the following equation.

$$D_{Tot} = D_{LA} + D_{Proc\text{-}E} + D_{Fr} + D_{Pkt} + D_{Tx} + D_{Qu} + D_{Prop} + D_{PB} + D_{Proc\text{-}D} + D_{Dec} + D_{PLC}. \quad (2)$$

As to the delay element at non-IP network, which is $D_{LA} + D_{Proc\text{-}E} + D_{Fr} + D_{Pkt} + D_{PB} + D_{Proc\text{-}D} + D_{Dec} + D_{PLC}$, almost all the delay elements are fixed if the type of coding or error correction scheme is determined. On the other hand, the delay composed of $D_{Tx} + D_{Qu} + D_{Prop} + D_{PLC}$ is not fixed, where packet delay is heavily dependent on the states of the network such as the bandwidth of the link, the capacity of the router, offered traffic and the attribute of the packets, etc. If the packets are transmitted through the link with the same link capacity along the same route and the same physical medium without loss $D_{Tx} + D_{Prop} + D_{PLC}$ will be also fixed. Otherwise, the network delay is a random variable, which depends on the network environment. Therefore, we have to anatomize the IP network in more detail, via which we can compute the delay budget at every element of the network. In order to simplify the discussion, let us divide the total delay into two parts: the IP network delay and non-IP-network delay, which is given as follows.

$$D_{tot} = D_{IP} + D_{NIP}, \quad (3)$$

where $D_{NIP} = D_{Enc} + D_{Pkt} + D_{Rec} + D_{PLC}$, and $D_{IP} = D_{Net}$.

Because we have stated that D_{NIP} is constant, we can estimate the delay budget outside the IP network if traffic source profile such as the signal generation rate and terminal type, the coding scheme, and the attribute of receiver are identified. From now on let us consider the IP network, and seek for a method for computing the delay budget in an IP cloud.

IP network cloud can be composed of multiple network domains. A network domain is an autonomous system in which the network policy is the same. An IP network domain can be composed of two kinds of nodes: two edge nodes at ingress and egress boundary and a group of core nodes. Fig.3 illustrates a detailed picture of an IP cloud which is composed of two IP network domains (In Fig.3, IP network domain is represented as AS (Autonomous System)). Let us assume that an AS is composed of one edge router, two core routers and one gateway router.

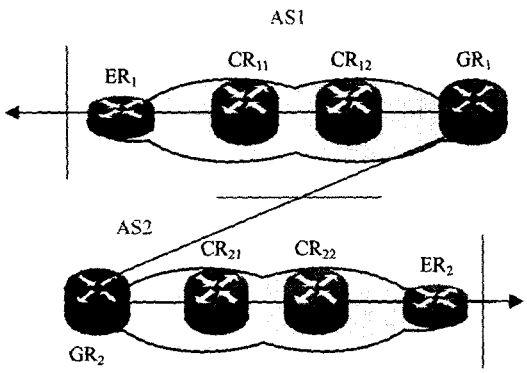

Fig.3 IP network cloud with two ASs.

If we represent the end-to-end packet delay in an IP cloud which is composed of two ASs, it is represented as follows:

$$\begin{aligned} D_{IP} &= \text{Propagation delay in links} + \text{Nodal delay} \\ &\quad (\text{Transmission delay} + \text{buffering delay}) \\ &= D_{Prop} + D_{Tx} + D_{Qu} \\ &\approx L\times 5 + D_{ER1} + D_{CR11} + D_{CR12} + D_{GR1} + D_{GR2} + D_{CR21} + \\ &\quad D_{CR22} + D_{ER2}. \end{aligned} \tag{4}$$

In eq.(4) the unit of delay is in millisecond, L (unit: kilometer) is the end-to-end length of IP network and 5 is computed from the speed of fiber optic per kilometer of 0.2×10^{6}Km/sec.

There may exist various ways for the allocation of the delay budget to an IP network element from E2E delay requirement. The simplest approach is to allocate delay budget in a uniform manner to every associated routers along the path. However, this assumption is effective if and only if the offered load is uniformly distributed, otherwise delay budget is allocated in more efficient way such that greater delay budget is allocated to more heavily loaded routers [4].

4. Decomposition of delay to network elements

In general edge router has a small capacity and the delay time required for the processing of packet header (packet classification, policing, etc.) is high. On the other hand, core router has a very large capacity and it is free from the complex processing of packet headers. Therefore, we have to allocate high delay budget to edge routers compared to the core routers unless core routers are heavily loaded, which is usually unlikely to occur in a well-provisioned commercial IP network.

Table 2 illustrates an example of the distribution of the delay budget for the voice traffic with G.711 coding in an end-to-end path for the domestic Korean VoIP network. Typical values are assumed by following the recommendation of ITU-T [3].

Table 2 Decomposition of delay budget

Network element		Delay budget (ms)	
Propagation delay (1,000Km)		5	
Sending end point	Packet formation (2 frames in a packet)	40	Sum= 81
	Packet insertion time (200bytes)	1	
Receiving end point	Jitter buffering (center of 60ms delay jitter buffer)	30	
	PLC (packet loss concealment) (1 PLC frame)	10	
Delay at IP network 1	Edge Node 1 (sum of queuing and processing)	10	Sum= 17
	Core Nodes at network 1 (sum of queuing and processing)	$2\times N_C$ (N_C=2)	
	Inter-Gateway 1 (sum of queuing and processing)	3	
Delay At IP network 2	Edge Node 2 (sum of queuing and processing)	10	Sum= 17
	Core Nodes at network 2 (sum of queuing and processing)	$2\times N_C$ (N_C=2)	
	Inter-Gateway 2 (sum of queuing and processing)	3	
Delay at Non IP network1 (at source side)		15	
Delay at Non IP network2 (at destination side)		15	
Total E2E delay		150	

In Table 2 the delay components $D_{Proc\text{-}E}$, $D_{Proc\text{-}D}$ and D_{Dec} are assumed to be negligible. From Table 2 we can find that, among the total E2E delay of 150ms, the network delay is 69ms, whereas the endpoint delay is 81ms. From various research works for the mapping between the objective and subjective QoS parameters, there exists a way for the mapping between the E2E delay, R-value of E-model, MOS (Mean Opinion Score) and QoS perceived by an end user. Table 3 summaries the relationship between them [8]. R-value is defined by ITU-T [7] and the relationship between R-value and MOS is defined in [1]. From the above discussion and Table 3, we can find that the

delay budget of 69ms of the network delay corresponds to an R-value of 90 or above and MOS of 4.34~4.5, and the perceived QoS is "Satisfied".

Table 3 Delay performance and voice QoS

Network delay, d	R Value	MOS	Quality Perceived by a user	ITU-Y.1541 IP QoS class
100ms	90	4.34 ~ 4.5	Satisfied	0
150ms	82	4.03 ~ 4.34	Partially unsatisfied	1
233ms	72	3.6 ~ 4.03	Unsatisfied	1

5. Conclusions

In this work we proposed a method to anatomize the element of delay for the VoIP services in NGN in an end-to-end manner. By investigating the points of delay in the end-to-end path of a voice connection, we could estimate the delay budget for an IP transport network.

From simple but comprehensive estimation of the delay budget, we could extract the delay requirement in the backbone network of the network service provider in a quantitative manner.

Even though this result does not tell all the facts in the end-to-end delay of voice application in NGN, especially in Korea, we can infer some level of guideline for the delay budget of NGN VoIP services that is composed of PSTN as an access network and IP network as a backbone network.

Future research area includes the modeling of more realistic network environment and the sophistication of the modeling method to estimate the end-to-end delay.

References

[1] R.G. Cole and J.H. Rosenbruth, "Voice over IP performance monitoring", SIGCOMM Computer Communication Review, Vol.31, No.2,April 2001.

[2] KO Y.-K. and LEE I.-S., "An evolution scenario for Pre-NGN toward a genuine NGN", KT Technical Review, Vol.16, No.2, June 2002.

[3] ITU-T Recommendation Y.1541, Geneva, May 2002.

[4] H. Ren and K. Park, "Performance evaluation of optimal aggregate-flow scheduling: a simulation study", Computer Communications 26 (2003) 222-236.

[5] Iain Lockyer, "Deploying QoS in service provider networks", http://www.securitytechnet.com/resource/rsccenter/presentation/cisco/networkers02-australia/.

[6] M. Karam and F. Tobagi, "Analysis of the delay and jitter of voice traffic over the Internet", INFOCOM 2001.

[7] ITU-T Recommendation G.107, "The E-model, a computational model for use in transmission planning", December 1998.

[8] S. Ohtani, "Asking for the QoS of the VoIP", Telecommunication, March 2002.