Dimensioning Links for NGN VoIP Networks

Yoon-Kee Kim*, Hoon Lee** , Kwang-Hui Lee** Regular Members

ABSTRACT

In this paper we present a theoretical framework for the network design with delay QoS guarantee to a voice at the packet level. Especially, we propose a method for estimating the bandwidth at the ingress edge routers accommodating the voice connections and data sessions in the next-generation IP network. First, we describe network architecture for VoIP (Voice over IP) services in the NGN (Next Generation Network). After that, we propose a procedure for dimensioning the bandwidth at the output port of a router that accommodates voice and data traffic using the non-preemptive queuing system with strict priority service scheme. Via numerical experiments we illustrate the implication of the proposition.

Key Words : NGN, VoIP, Network dimensioning, QoS

I. Introduction

Recently Internet has been undergoing a rapid change toward the next-generation network (NGN). NGN tries to provide the customers with real time services such as a voice service as well as the current high-speed Internet services in a single framework of IP network. NGN service includes the current voice, data or video services via several access networks such as PSTN, IMT-2000 and Ethernet.

There exist network architectures and service scenarios for NGN in a number of literature [3,8,9]. There exist a number of works on the modeling of VoIP in the multiplexed packet level [1,2,5]. However, little work has been done for the modeling and bandwidth dimensioning of NGN-VoIP network except [5]. Hoey et al. proposed a method to design NGN transport networks for real-time voice applications in an end-to-end manner at the call level. They used the concept of Erlang for the offered load and Erlang loss formula in determining the number of links between end-to-end path of a network. Their work is concerned with the calculation of link capacity in connection level, which is in line with the design concept of PSTN (Public switched telephone network) link dimensioning.

Traffic model for voice at call level is well known in the field of PSTN. When the busy hour and the mean call generation rate from a group of customers and mean holding time of a call are given, we can obtain the offered load by using the famous Erlang formula [5]. The number of links required to guarantee the Grade of Service (GoS: e.g., the call blocking probability of 0.1% or 1%) for the given offered load is calculated by using the Erlang loss formula [12]. This approach has been useful at least in the design of a trunk capacity in PSTN network.

However, in IP-based NGN network, the network resource is represented not in the number of circuits but in bandwidth. Therefore, one must know the amount of bandwidth represented in bit per second. Let us review related works in this field. Ahlgren et al. used a packet level model for dimensioning the link capacity for IP Telephony by using MMPP/D/1/K model [1]. Chuah and Katz introduced a bufferless fluid-flow model and central limit theorem for aggregated

* KT Technologies Lab. Jeonmin-Dong, Daejon, Korea, 305-811(yoonkee@kt.co.kr)
** Changwon National University, Changwon, Korea, 641-773 (hoony@changwon.ac.kr)
논문번호 : 030097-0311, 접수일자 : 2003년 3월 11일

voice packets in dimensioning the link capacity of the VoIP network [2]. The two works assumed no QoS requirement in the packet level. However, delay of a packet from VoIP traffic has to be kept under a certain level (as described in Section II), and delay requirement can be guaranteed via a scheduling mechanism at the output port of a router. This motivated the work described in this paper.

There exist a number of scheduling mechanisms for the VoIP: Typical examples are SP (Strict priority) and WRR (Weighted round robin). SP is considered to be more suited to the absolute support of delay to the VoIP traffic. However, simulation requires too much effort if we want to obtain an intuition for the design of a large IP network. In this work we will set up a theoretical framework for the design of link capacity for NGN router with delay priority to the VoIP packet over the data traffic.

This paper is composed as follows: In Section II, network architecture for VoIP services over NGN is described. In Section III, a model for voice services in NGN is described by using the M/G/1 non-preemptive queuing model with strict priority service mechanism for voice traffic. In Section IV, a procedure for the link dimensioning is described. In Section V, the result of numerical experiment is described. Finally in Section VI, we summarize the paper.

II. Voice service over NGN

Currently, it is assumed that voice service in NGN is realized by employing AGW (access gateway) or TGW (trunk gateway) for connecting the telephone users directly to the NGN backbone or connecting the PSTN to the IP networks. TGW/AGW is connected to an NGN core network via a Network access server (NAS), which is an access router. Fig.1 illustrates the VoIP service architecture for the NGN network [6].

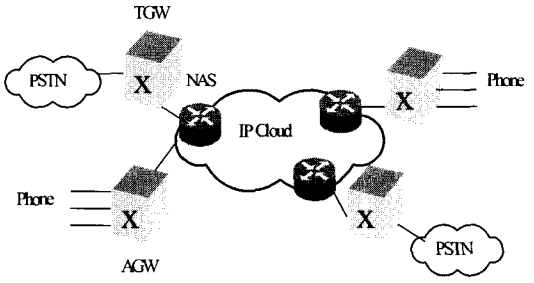


Fig.1. Architecture of Voice service over NGN

At the sending party, TGW connects trunks from PSTN to a NAS, which is an access gateway to an IP cloud, whereas AGW aggregates individual circuits from the users to NAS. At the receiving party, TGW terminates an IP pipe into a number of circuits in PSTN, whereas AGW terminates an IP pipe into a number of circuits for a group of phone terminals. So, TGW and AGW aggregate traffic from the TDM links, whereas NAS aggregates traffic from multiple TGWs and/or AGWs.

As we may find from the above discussion on the behaviors of TGW and AGW, a natural way of representation for an input to the TWG is the number of circuits. On the other hand, it is natural to represent the input of NAS, which is a router, in the unit of bandwidth. This is the main reason for our discussion about the packet level traffic model.

The voice packets from a source have to be transferred to the receiving end within a specified time (eg., 100 or 150 msec), otherwise they can not be reassembled for playout at the receiver. This necessitates the use of a priority service scheme for voice packet at the router in NGN edge and backbone.

III. Delay of Voice Service in NGN Node

As one can find from Fig.1 the voice signal is encoded and transformed into packets at TGW/AGW and transferred into a core network via NAS (Network Access Server). In the core network, we assume a DiffServ architecture,

which is a typical packet service architecture in NGN [10]. Under the DiffServ architecture over NGN, voice packets are classified as an EF (Expedited Forwarding) PHB (Per Hop Behavior) at the ingress edge and served with strict priority (SP) over the data packets. At once voice packets enter the backbone, they pass the core network with high priority over the data traffic. The associated core routers along the end-to-end path fetch the voice packets with strict priority (SP) over the data packets. However, some voice packets will experience buffering delay because of variable size of data packets.

Now let us present a model for computing a delay in delivering the voice traffic between a pair of source and destination. Let us assume that a NAS accommodates voice and data packets from a number of connections. Data packets are considered as aggregated background traffic. Fig 2 illustrates this concept for packet buffering and scheduling at NAS. At each router in an end-to-end path of the network, packets are classified into voice and data, and each packet is fed into corresponding buffer, the voice buffer and data buffer. It is assumed that voice and data packets generated from an arbitrary number of V input interfaces are distributed into V output ports with even distribution. The last assumption renders us to simplify the analysis of a router into a single output port.

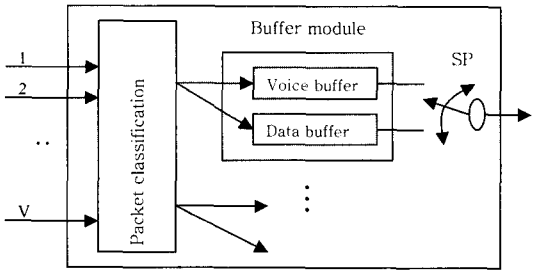


Fig.2. SP scheduling for voice packet at NAS.

As we have stated in the above discussion, the router has an SP service scheme for voice packets. In [11] Yamauchi argued that voice packets can be transmitted almost in their original pattern if voice buffer is served in SP over the data buffer. Using this fact he proposed a service differentiation model for the voice packet at the access router of IP network, which has at least two buffers at the output module, by assuming a single M/D/1 queuing model. In [4] Shiroyama used an M/D/1/K queuing model with SP service scheme in estimating the waiting time of a voice packet in a voice buffer by assuming that the data packets do not affect the delay performance in voice packets if an SP service scheme is adopted in the scheduler.

However, we have to note that simplified M/D/1 or M/D/1/K models can not reflect the NGN architecture where voice and data packets share a link, where the size of data packet is variable in size, and its size is usually much less than or greater than that of voice packet. Therefore, in this work, we assume a more realistic and general model for the evaluation of performance for the prioritized packet service scheme in a node by using M/G/1 queuing model, where the service time is generally distributed. Furthermore, let us model the server as a non-preemptive server with strict priority policy for voice buffer, which faithfully models our packet service architecture.

Using the non-preemptive M/G/1 queuing model with SP scheme in [12], let us describe a procedure for obtaining the waiting time of voice packet by assuming some variables. Packet arrival processes for voice and data are mutually independent and each packet follows a Poisson process with mean arrival rate λ_v for voice packets and λ_d for data packets. The service times for voice and data follow general distributions with mean service rates $1/\mu_v$ for voice and $1/\mu_d$ for data. The variances for the service times of voice and data packets are assumed to be σ_v^2 and σ_d^2, respectively. The mean offered load of the voice and data packets into corresponding buffer is $\rho_v=\lambda_v/\mu_v$ and $\rho_d=\lambda_d/\mu_d$, respectively.

Packet scheduling at the buffer module follows an SP scheme, which operates in the following manner. Initially, server visits a voice buffer. If there exist packets in voice buffer, the server serves them until the buffer is vacant. Otherwise, server visits data

buffer and serves a packet in that buffer, and it returns to voice buffer and repeats the above operation. Let us assume that the moving time between the two buffers is so small that it is ignored. When a voice packet enters a voice buffer while a data packet is receiving service by the server, it waits in the voice buffer until the server finishes service for data packet. Therefore, the service scheme is non-preemptive.

Let S_v be the sojourn time in the system (buffer and server) and W_v be the waiting time of a voice packet, then the following relationship exists between S_v and W_v.

$$S_v = W_v + \frac{1}{\mu_v}. \tag{1}$$

In order to obtain a formula for W_v, let us define variables Q_v, the expected number of voice packets in voice buffer and R, the expected value of the residual service time of a packet in the server. When a server operates in SP for the voice packets, the mean waiting time of a voice packet can be obtained by using the mean waiting time of a customer for a single class M/G/1 queuing system with vacation, where a vacation occurs when a server visits a data buffer in case there is no packet in voice buffer. Therefore, we obtain the following result.

$$W_v = \frac{Q_v}{\mu_v} + R. \tag{2}$$

From Little's formula, eq(2) is rewritten by

$$W_v = \frac{\lambda_v}{\mu_v} W_v + R. \tag{3}$$

If we arrange eq.(3) with respect to W_v, we obtain

$$W_v = \frac{R}{(1-\rho_v)} \tag{4}$$

When packet arrivals follow Poisson distribution, we can use the PASTA (Poisson arrival see time average) property in the computation of R [14]. From [12], we obtain

$$R = \sum_{k=1}^{2} \lambda_k \frac{E[\tau_k^{\,2}]}{2} \tag{5}$$

where τ_k is the service time of a packet, k=1 for voice packet and k=2 for data packet. $E[\tau_k^2]$ is the second moment of τ_k, which is represented by traffic parameters.

$$E[\tau_k^2] = \sigma_k^2 + \frac{1}{\mu_k^2}, k = 1,2. \tag{6}$$

Finally, we can obtain the mean value of waiting time for a voice packet in a voice buffer, which is given as follows:

$$W_v = \frac{\lambda_v(\sigma_v^2 + 1/\mu_v^2) + \lambda_d(\sigma_d^2 + 1/\mu_d^2)}{2(1-\lambda_v/\mu_v)} \tag{7}$$

Note that we can obtain the sojourn time of a voice packet in the system from (1) and (7).

Ⅳ. Dimensioning Link Capacity

Our final aim lies in the determination of the relationship between the load of voice traffic ρ_v and the load of data traffic ρ_d such that the required end-to-end delay of a voice connection is kept under a certain limit. Let us define D_{e2e} be an end-to-end delay of a source-destination pair in a network and define that D_{node} is the target value for the system delay in a single node decomposed from an end-to-end delay. Then, from the above discussion the following inequality is satisfied.

$$S_v < D_{node} \tag{8}$$

In order to compute D_{node} let us classify the delay of voice packet in the network into four typical parts [2,7]: PCM transcoding delay D_{TC}, propagation delay D_p, delay due to buffering and packet transfer D_b and additional delay D_{ex} for an additional packet

processing. Among them, note that D_b is the only variable part of the delay. If we let δ be the target value for the end-to-end delay D_{e2e} of a voice connection, we can obtain the following relationship

$$D_b < \delta - (D_{TC} + D_p + D_{ex}) \quad (9)$$

Let us assume that the number of node between an end-to-end path of a voice connection is H and delay is evenly distributed at each node along the end-to-end path of a connection. Then, we can obtain

$$D_{node} = \frac{\delta - (D_{TC} + D_p + D_{ex})}{H}. \quad (10)$$

If we equate the equations (1),(7),(9) and (10), we can obtain a graph between ρ_v and ρ_v under the constraint that eq.(8) is satisfied to a voice packet. Note that we can also compute the amount of bandwidth, C_v, required to guarantee a specified delay of a voice packet at the output side of a NAS, which is given by

$$C_v = \rho_v \times C. \quad (11)$$

In eq.(11) C is the total bandwidth of a NAS which accommodates an aggregate of voice and data connections.

V. NumericalResultsandDiscussions

Let us assume that the voice signals from PSTN are transformed into voice packets by G.711- encoding. It is known that voice packets from a CODEC with G.711- μ law generate 64Kbps per each connection. It is usually assumed that the packet processing delay (D_{TC} in eq.(9)) for encoding at source and decoding at receiver ranges from 50 to 80 msec [17]. Let us assume that D_{TC} is equal to 80msec including the propagation delay D_p and D_{ex}. Let us assume that the number of node in an end-to-end path is $H=5$. Let us assume three scenarios for the delay. Table 1 summarizes the resulting parameters computed from eq.(10) [15,16].

Table 1. Target values for the delay

Scenario	δ (msec)	D_{node} (msec)	QoS class
A	100	4	Toll Quality
B	150	14	Mobile phone Quality
C	200	24	IP Phone Quality

Let us assume the size of voice and data packets in order to compute the values of $1/\mu_v$ and $1/\mu_d$. In order to compute $1/\mu_d$ let us first assume a worst case in which the data packet has a size of 1500bytes (1460bytes of body and 40bytes of header). The packet size of VoIP packet generated from G.711 Vocoder is assumed to be 216bytes (160bytes of payload and 56bytes of headers) [7]. Let us assume that the bandwidth of an output port in NAS is 10Mbps.

Finally, using the QoS objectives that is given in Table1, we could obtain a graph that represents a load-map which represents the relationship between the offered load of voice traffic and data traffic at the output port of NAS which satisfies the required QoS target given in eq. (8) and Table 1, which is shown in Fig.3. In Fig.3, there are three sticks at each point: The left one is for scenario A, the middle one is for scenario B, and the right one is for scenario C. The arrival rate at each axis is represented in the unit of the number of packets per second. The packet inter-arrival time for G.711 coder is assumed to be 20 milliseconds, which implies that 50 packets can be generated in a second from an active connection. Then, we can infer that the arrival rate of voice packet equal to 1,000 implies that at least 20 active voice connections can be served from the server simultaneously.

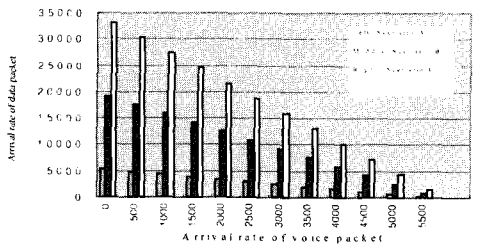


Fig.3. Traffic load-map for voice and data when data packet size is 1500bytes.

As we can find from Fig.3, the load-map (upper bound on the load) between voice and data traffic follows a linear non-increasing function as the arrival rate of voice packet increases. Note also that the offered load of data traffic can be greater than one: For example, for the three scenarios of D_{node} =*24*, *14*, and *4*,the offered load of data traffic can increase to about 40,20,6 times, respectively, that of the configured maximum link capacity (10Mbps in this case) when the offered load of voice packet is zero. This results form the assumption of too large delay target of the proposed model. When one assumes very tight delay targets for both voice and data packets this kind of over-booking may not be found. The phenomenon of over-booking of data traffic can be avoided if one assume an appropriate packet loss rate by limiting the buffer capacity of data buffer, which is remained as a future research area.

In order to investigate the effect of the different data packet sizes into the design of link capacity, let us assume two cases: First, data packet has the same size as the voice packet, which implies that the network is flooded with small packets, which is usual in the real Internet. Second, data packet is assumed to be variable with mean packet size of 500bytes, which implies that the network is loaded with small voice packets and a moderate size of data packets.

Fig.4 illustrates the result for the data packet with a uniform size of 216bytes, which is assumed to be the same as that of voice packet. This illustrates the network environment in which all the packets are small and homogeneous irrespective of the traffic type. Fig. 5 illustrates the result for the data packet with a mean size of 500bytes and a variance given by $\sigma_d^2=0.5/\mu_d^2$, where variance is arbitrarily assumed. This illustrates the inhomogeneous network environment in which the sizes of voice and data packets are different and the size of data packet is assumed to be a random variable, which is more realistic than those two examples in Fig.3 and 4.

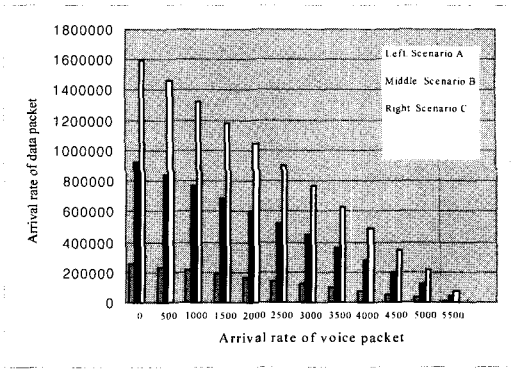


Fig.4. Traffic load map for voice and data when data packet size is 216 bytes.

If we investigate those three figures, we can find that the load-map of an IP network that accommodates voice and data packets is heavily dependent on the characteristics of data packets .We can also note that the total load of the system can exceeds the configured link capacity if one assumes a loose delay target for high priority packet (voice packet in this work) and no delay limit on the low priority packet (data packet) in a SP packet scheduling scheme, because the buffer capacity of data packet is assumed to be infinite. This problem is also discussed in [13]. Therefore, a more detailed investigation on the characteristics of data packets as well as an anatomy on the delay budget has to be carried out in order to establish an optimal rule for the provisioning of bandwidth resources in the NGN networks.

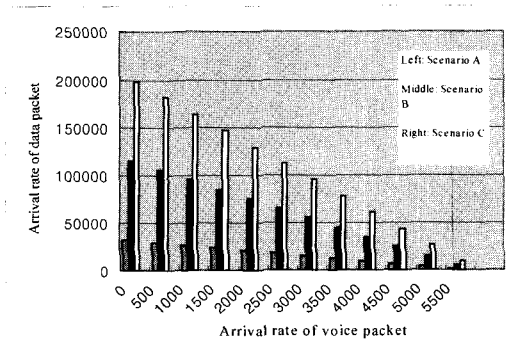


Fig.5. Traffic load map for voice and data when the data packet size is variable with mean value of 500bytes.

VI. Conclusions

In this work we proposed a method to quantify the relationship between the required bandwidth of voice and data traffic that shares a pipe in the access router of NGN VoIP network. The model is described at packet level in order to incorporate the generic QoS characteristics, especially, the delay requirement of a voice packet in NGN that is composed of PSTN and IP networks, where the aggregated bandwidth is estimated.

Via some numerical experiments we could show the implication of the work in the design of the bandwidth of the access network in NGN voice over IP network. The result will be useful in the estimation of an upper limit on the performance and network dimensioning of an access network of VoIP services in the NGN.

The future research area includes the trimming of load-map into realizable limit by taking into account realistic assumption on the end-to-end delay budget of voice source as well as the assumption of finite buffer for data traffic. Also, measurement of real traffic data at the operational network as well as the sophistication of the measurement methods can be the area of future study.

REFERENCE

[1] B. Ahlgren, A. Andersson, O. Hagsand and I. Marsh, *Dimensioning links for IP Telephony*, Internet Telephony Workshop 2001.

[2] C.-N. Chuah and R. H. Katz, *Network provisioning & resource management for IP Telephony*, http://www.cs.berkeley.edu/~chuah/research/paper/csd-99-1061.pdf.

[3] Jean-Yves Cochennec, *Activities on Next-Generation Networks under global information infrastructure in ITU-T*, IEEE Communications, July 2002.

[4] K. Shiroyama et al., *Performance evaluation of a hardware router with QoS control capabilities*, TECHNICAL REPORT of IEICE, NS2001-258 (March 2002).

[5] G.V. Hoey, S.Van den Bosch, P. de La Vallee Poussin, H. De Neve and G.H. Petit, *Dimensioning of NGN transport networks for real-time voice applications*, Alcatel Telecommunications Review, 2nd Quarter 2001.

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

[7] Hoon Lee, *Methods for supporting guaranteed-Quality voice services over NGN*, Final report, KT Telecommunications Network Laboratories, August 2002.

[8] Stewart D. Personick, *Evolving toward the Next-Generation Internet: Challenges in the path forward*, IEEE Communications, July 2002.

[9] Sang-il Lee, *The technology development strategy for KT-NGN*, Korea Telecom Technical Review, Vol.16, No.2, June 2002.

[10] D. Goderis, H. De Neve, Y. T'Joens, J. De Vriendt, and T. Soetens, *Towards an integrated solution for multimedia over IP*, Alcatel Telecommunications Review 2nd Quarter 2001.

[11] K. Yamauchi et al., *Performance evaluation of a hardware router with QoS control capabilities*, Technical Report of IEICE, NS2001-258 (2002-03).

[12] P. Nain and D. Towsley, *Performance evaluation of computer systems: Lecture notes*, May 1994.

[13] C. Filsfils and J. Evans, *Engineering a multiservice IP backbone to support tight SLAs*, Computer Networks 40 (2002) 131-148.

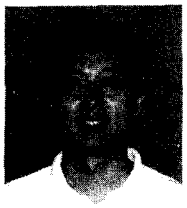
[14] R.W. Wolff, *Stochastic modeling and the theory of queues*, Prentice-Hall International, 1989.

[15] Seiji Ohtani, *Asking for the QoS of the VoIP, Telecommunication*, March 2002 (In Japanese).

[16] Debasis Mitra et al., *New directions in service management*, Bell Labs Technical Journal, Jan-Mar 2000.

[17] T. Sakaguchi, S. Yamamoto and A. Kawabata, *Diffserv scheduling mechanism for a real-time service with a guaranteed data traffic*, TECHNICAL REPORT of IEICE, NS2002-84 (July 2002).

Yoon-Kee Kim



February 1984: B.E. degree in Electronics at Kyungbook National University
February 1987: M.E. degree in Communications at Kyungbook National University
February 1987 - : KT Technology Group, Now he is a Director of NGN research team
Research interests: Network architecture, design, implementation of next generation networks.
Yoon-Kee Kim is a member of KICS.

Hoon Lee

February 1984: B.E. degree in Electronics at Kyungbook National University
February 1986: M.E. degree in Communications at Kyungbook National University
March 1996: Ph.D. degree in Electrical & Communication Engineering from Tohoku University, Sendai, Japan
February 1986 - February 2001: KT R&D Group
March 2001 ~ : Assistant Professor of Changwon National University
Research interests: Teletraffic engineering, Network design, performance analysis and provision of QoS for high speed telecommunication networks.
Dr. Lee is a member of IEEE, KICS and IEEK.

Kwang-Hui Lee

February 1983: B.E. degree in Electronics at Korea University
February 1985: M.E. degree in Electronics at Korea University
February 1989: Ph.D. degree in Electronics at Korea University
1991-1992: Visiting researcher, University of Wales and Newbridge Networks, England
1994-1995: Visiting researcher, UCL, England
1997-1999: Visiting researcher, University of Reading, England
2000 - : Visiting researcher, Nortel Networks
1988 - : Professor of Changwon National University
Research interests: Network/QoS management, Policy-based Networking, Multicasting protocol, Mobile computing