論文2003-40TC-12-2

NGN에서의 품질보장형 음성서비스 제공을 위한 대역 설계 방법

(Dimensioning Next Generation Networks for QoS Guaranteed Voice Services)

金 潤 基 *, 李 君 **, 李 廣 輝 **

(Yoon-Kee Kim, Hoon Lee and Kwang-Hui Lee)

요 약

본고는 차세대 IP망(NGN)에서의 대역 설계 방법에 관한 것이다. 특히, VoIP 뿐만 아니라 데이터 서비스를 수용하는 에지 라우터에서의 호레벨 및 패킷 레벨 음성 트래픽 대역 설계 방법을 제시하였다. 호레벨 모델은 실제 연결되는 호의 수를 통계적으로 계산하기 위해 평균, 분산과 같은 통계적 기법을 사용하고, 패킷레벨 모델은 M/G/1 큐잉 모델을 활용하여 음성 및 데이터 트래픽의 부하를 나타내었다. 제시된 트래픽 모델을 통해서 계산된, 음성과 데이터 연결을 위한 대역폭을 기반으로 최대 트래픽 부하를 예측할 수 있고, 또한 수치 시험을 통한 이의 결과를 제시하였다.

Abstract

In this paper we proposea method for estimating the bandwidth in next-generation IP network. Especially, we concentrate on the edge routers accommodating the VoIP connections as well as a group of data connections. Bandwidth dimensioning is carried out at call level and packet level for voice traffic in the next-generation IP network. The model incorporates the statistical estimation approach at a call level for obtaining the number of voice connections simultaneously in the active mode. The call level model incorporates a statistical technique to compute the statistics of the number of active connections such as the mean and variance of the simultaneously connected calls in the network. The packet level model represents a load map for voice and data traffic by using non-preemptive M/G/1 queuing model with strict priority for voice over data buffer. From the proposed traffic model, we can derive a graph for upper bounds on the traffic load in terms of bandwidth for voice and data connections. Via numerical experiments we illustrate the implication of the work.

Keywords : NGN, VoIP, QoS, Bandwidth Dimensioning, AGW, TGW

I. Introduction

Recently the traditional best-effort Internet is evolving toward the next-generation network (NGN). NGN tries to support real-time services as well as current data services in a single framework of IP network. However, NGN includes an evolutionary scenario in which current networks such as PSTN and wireless voice and data networks are included as an access network. There exist network architectures

* 正會員, KT技術硏究所
(KT Technology Laboratory)
** 正會員, 昌原大學校
(Changwon National University, Changwon)
接受日字:2003年11月8日, 수정완료일:2003年12月3日

and service scenarios for NGN in a number of literature[1-3]. There exist a number of works on the performance of voice service over IP network[4]. We can classify those works into two approaches: the approach at the call level and at the packet level. As to the former approach, we can find the works from[5, 6]. In [6] Hoey et al. proposed a method to design NGN transport networks for real-time voice applications. They used the concept of Erlang for the offered load of a single connection and Erlang loss formula in determining the number of links between end-to-end path of a network. We could find some areas that can be further investigated in his work. First, Hoey et al. assumed the traffic load as a multiplication of BHCA (Busy Hour Call Attempt) and MHT (Mean Holding Time) divided by 3600, a classical definition of Erlang[7].

In this paper we propose a new approach to obtain the expected value of the number of active voice connection by direct measurement. To the best of authors' knowledge, we could not find the measurement-based modeling of traffic load in an operational NGN-VoIP traffic in the call level. This is the first point of the motivation of our work. Secondly, Hoey's work adopted a worst-case analysis in which the required bandwidth of a trunk made by an aggregation of a large number of connectionsis computed as a multiplication of the number of link under the Grade of Service (GOS) constraint such as call blocking probability and the peak rate of a voice source. That approach has an implicit assumption that an end-to-end pipe inside IP networks is dedicated to voice connections, which corresponds to a segregated network provisioning approach.

On the other hand, in the real IP network that is operated in DiffServ mode, a link is not dedicated to a single class of traffic but shared statistically bya number of traffic classes, and voice packets are treated as urgent customers by using generic packet scheduling scheme such as SP (Strict Priority) or a family of WFQ (Weighted Fair Queuing) scheme. Under the DiffServ service architecture, there exists a QoS violation problem in which a voice packet can not receive a service in its contracted rate unless a certain policy is included in the scheduler, even though voice packet is treated with higher priority than the other types of packets. Examples of the policies are the call admission control for voice connections or bandwidth allocation (or limitation) policy in the output port of a router[8]. Little work is done for the modeling and bandwidth dimensioning of NGN-VoIP network except[9, 10], even thoughlots of works argue that voice service is to be classified as an EF (Expedited Forwarding) class in DiffServ (Differentiated Service) architecture of IETF (Internet Engineering Task Force), and peak rate of bandwidth equivalent to the packet generation rate of a voice source has to be provisioned by using the strict priority scheduling scheme at the output port of a router in the edge and core of the network[11]. This is the second point of the motivation of our work.

In this work we try to propose a scheme of link provisioning for VoIP services over NGN by using a call level traffic model from measurement-driven approach as well as by using a packet level traffic model that describes an upper limit on the average rate of voice packet in terms of bandwidth.

This paper is composed as follows: In Section 2, network architecture for VoIP services over NGN is described. In Section 3, a call level traffic model for voice traffic is proposed. In Section 4, a method for bandwidth estimation at packet level when a router accommodatesboth voice and data connections in a link is described. In Section 5, the result of numerical experiment is described. Finally in Section 6, we summarize the paper.

Ⅱ. VOIP service over NGN

It is assumed that the traditional voiceservice from PSTN is realized using 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 core network via Network access server (NAS), which is an access router. Fig.1 illustrates the VoIP service architecture for the NGN network[12].

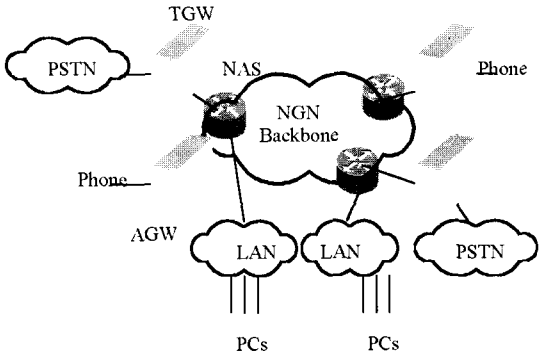


그림 1. NGN에서의 VoIP 구조
Fig. 1. Architecture of VoIP over NGN.

At the sending party, TGW collects circuits from PSTN to NAS, which is an access gateway to 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. TGW and AGWaggregate 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, it is easy to understand the relationship between the number of circuits in theTDM world and the amount of bandwidth in a pipe that accommodatethe aggregate traffic if we represent the traffic load at the input side of NAS in terms of the number of circuit and the traffic load to NAS in terms of bandwidth, respectively. This is themain reason of the discussion about the call level traffic model that is described in Section 3 and the packet level discussion in Section 4.

Ⅲ. Call level model for VoIP services

A traffic model for voice call is well known in the field of PSTN. Whenthe busy hour and its corresponding parameters such as BHCA and MHT of a call is known, we can obtain the offered load by using the Erlang formula[6, 7]. Furthermore, we can determine the number of links required to guarantee the GoS for the given offered load by using the famous Erlang loss formula[7]. This approach has been useful in the design of a PSTN network. However, in the operation and management of NGN network, the number of customer and their activity in the packet level varies with respect to time. In order to utilize the bandwidth in IP network efficiently, one must utilize the fact that the traffic volume varies in a statistical manner, from which we can infer that bandwidth varies in a statistical manner.

Let us focus on the boundary of PSTN and IP network in NGN. Via real measurement and a statistical modeling approach in the determination of the number of connections that are active simultaneously, we obtain the mean and variance of the number of voice connection. Let us assume that N customers of a PSTN network are connected to a TGW/AGW via NGN. Each customer is engaged in a voice communication. Let us monitor the input port of TGW/AGW, and record the duration of the occupancy of the link group every T seconds. The duration of the connection can be measured from the log file of the soft switch over which the connection initiation protocol is run. Fig.2 illustrates the state of the occupancy of the link group with five links during an observation period T. Following the discussion in[13] let us assume that $N=5$ (In a real network, the PSTN link group can accommodate tens of thousand of connections at a time).

The bold line indicates the duration of a call, and let t_i be the time interval at which exactly i links are occupied by active calls(If a call is engaged in voice communication, we call it an active call. We use call and connection synonymously). As one can find from Fig.2, there may exist a case in which the same number of links are occupied by active connections at any different time instant.

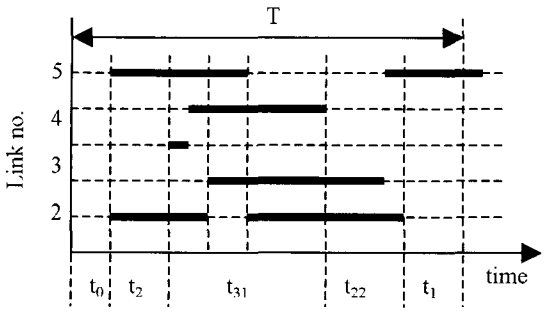


그림 2. 시간대별 링크 점유
Fig. 2. Occupancy of link with respect to time

Let t_{ij} be the j-th time interval at which i links are simultaneously occupied by i active connections. Then, we have

$$t_i = \sum_{j=1} t_{ij} . \tag{1}$$

Note that T satisfies the following formula.

$$T = \sum_{i=0}^{N} t_i . \tag{2}$$

The total time of the link occupancy, Θ, is given by

$$\Theta = \sum_{i=1}^{N} i t_i . \tag{3}$$

Finally, the traffic load, A, can be defined by

$$A = \frac{1}{T} \sum_{i=1}^{N} i t_i . \tag{4}$$

If we rewrite (4), we have

$$A = \sum_{i=1}^{N} i \frac{t_i}{T} . \tag{5}$$

Note that the term t_i/T in the last formula indicates the rate of time during which exactly i connections are active simultaneously during a monitoring interval T. Therefore A is equalto the average number of links that is occupied among N links. Note also that this is very useful as a measure of the mean number of active connection in the network if one can measure the information about the number of active connections from the network. Note that the number of links A that have been obtained from (5) is the average value. However, the number of connections that is active in an instant varies in time. In order to obtain reliable statistics for the value A, let us obtain the value of A in a statistic manner. To that purpose let us repeat the computation of A in a number of L times. So, measurement is carried out in a total of LT seconds. Let us denote each result by A_1, A_2, ... A_L. Then, A_i, i=1, 2, ..., L, where L is the number of sample, constitutes a series of independent random variables. In order to obtain an expected value for the number of links that is occupied in voice communication under steady-state, let us compute the mean and variance of a random variable A. Let us assume that L is sufficiently large so that the random variable Afollows a Gaussian distribution. Let us assume that and σ^2 are the mean and variance of a random variable A, respectively. In order to obtain a safe design value for the bandwidth, let us assume a 95th percentile value of the observed data, which implies that the following equation holds.

$$P(\alpha - k\sigma \le \Gamma \le \alpha + k\sigma) = 0.95, \tag{6}$$

where Γ is the newly determined expected value of the number of links that is occupied in voice communication under the steady-state and k is computed from a standard normal distribution. Our interests lie in the upper part of the value for Γ, so that Γ is given by[14]

$$\Gamma = \alpha + 1.96\sigma. \tag{7}$$

Ⅳ. Estimation of Link Capacity for Voice and Data

In Section 3, we could obtain the expected number of customers for voice connection between PSTN and

TGW, which is represented as Γ. In this section let us presenta method to compute the required bandwidth of a link at the output of TGW that supports a group of VoIP calls, by using a packet level analysis method for guaranteeing the delay. If we sum up those two arguments, we can represent the points of interests for dimensioning capacity of network as illustrated in Fig.3.

Let us represent the amount of bandwidth required for guaranteeing the predetermined delay for a voice packet in a NAS, which is represented as C_v. Let us assume that a voice source generate voice packets every second when a customer is in an active state. Since voice communication is an interactive application in which two persons are speaking and listening at a time, the average ratio of active (speaking) and silent (listening) state is usually assumed to be 0.35 and 0.65. In the world of network design we usually assume the ratio to be 0.5:0.5 in order to given some safeguard[15].

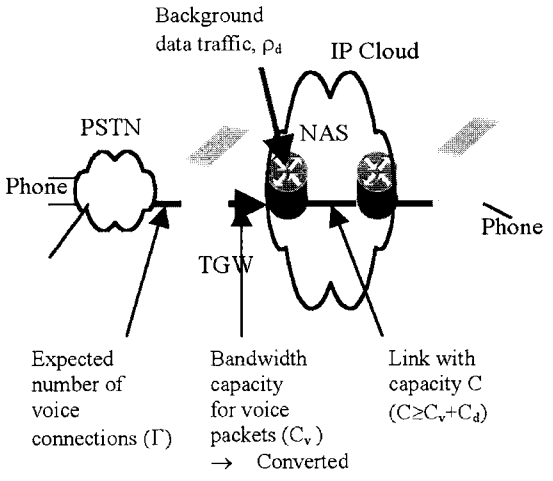


그림 3. 각 요소별 출력단
Fig. 3. Points of interests and corresponding outputs.

Then, the average number of voice packets generated by a customer is 0.5ω. Since the expected number of active customer is Γ, we can compute that the required bandwidth C_v for a group of voice customers is given by

$$C_v = 0.5 \times \omega \times \Gamma. \tag{8}$$

Now we can represent the offered load v to a NAS from the bandwidth C_v and the total link capacity C, which is given in eq. (9).

$$\rho_v = \frac{C_v}{C} \tag{9}$$

Note that C_v in eq.(9) represents the required aggregate bandwidth of voice connections. However, in reality, C_v is not dedicated to packets from voice connections, because voice and data packets share a bandwidth with capacity C at any instant, and voice packets are only treated with high priority than data packet with SP policy. This implies that an average bandwidth of C_v/C has to be guaranteed statistically to voice packets.

Now let us compute C_d, the amount of bandwidth a router can allocate to data traffic. As we have described in the above discussion, C_d is limited by the delay requirement of a voice packet. In order to represent a formula for C_d let us assume some parameters for voice and data packet. Packet arrival processes for voice and data from sufficiently large group of customers are assumed to be mutually independent and follow Poisson processes 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 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 follows a SP scheme, which operates in the following manner. Initially, a 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 data 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 us assume that S_v is the sojourn time of a voice packet in the system (buffer and server) and W_v is the waiting time of a voice packet. Then, we have the following relationship between S_v and W_v.

$$S_v = W_v + \frac{1}{\mu_v}. \tag{10}$$

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. A discussion on the analysis of a single class M/G/1 queuing system with vacation is given in[16, 17], and the average value of waiting time for a voice packet in a voice buffer is given as follows:

$$W_v = \frac{\lambda_v(\sigma_v^2 + 1/\mu_v^2) + \lambda_d(\sigma_d^2 + 1/\mu_d^2)}{2(1-\lambda_v/\mu_v)}. \tag{11}$$

Note that we can obtain the sojourn time of a voice packet in the system from (10) and (11). Our final aim lies in the determination of the relationship between ρ_v and ρ_d such that the following inequality is satisfied.

$$S_v < D_{node}, \tag{12}$$

where D_{node} is the target value for the system delay in a single node decomposed from an end-to-end delay, D_{e2e}, of a source-destination pair in a network. In order to compute D_{node} let us classify the delay of voice packet in the network into four typical parts[18]: PCM transcoding delay D_{TC}, propagation delay D_p, delay due to buffering and packet transfer in an end-to-end path D_b and additional delay D_{ex} for some additional packet processing. Among them, note that D_b is the only variable part of the delay. If we let be the target value for the end-to-end delay D_{e2e} of a voice connection, we can obtain the following relationship.

$$D_b \le \tau - (D_{TC} + D_p + D_{ex}). \tag{13}$$

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

$$D_{node} = \frac{\tau - (D_{TC} + D_p + D_{ex})}{H}. \tag{14}$$

If we equate the equations (10), (11) and (12), we can obtain a relationship between λ_v and λ_d under the constraint that eq.(14) issatisfied to a voice packet at each node along the end-to-end path. Finally, we obtain the following formula for the upper bound on λ_d., which is given as follows:

$$\rho_d = \frac{2(D_{node} - \frac{1}{\mu_v})(1 - \frac{\lambda_v}{\mu_v}) - \lambda_v(\sigma_v^2 + \frac{1}{\mu_v^2})}{\mu_d\sigma_d^2 + \frac{1}{\mu_d}}. \tag{15}$$

If the QoS objective value D_{node} is known we can obtain a graph that represents a load map which represents the relationship between the offered load of voice traffic v and data traffic din a plane at the output port of NAS from eq.(15). If ρ_d is determined as shown in (15), C_d is computed as follows:

$$C_d = \rho_d \times C. \tag{16}$$

Finally, we have an upper limit for the expected utilization of a link, which is given as follows:

$$U = \frac{C_v + C_d}{C}. \tag{17}$$

V. Numerical Results and Discussions

Let us assume that each customer from POTS (Plain Old Telephone Service) network generates voice packet with G.711- encoding. The voice packets undergo VAD (Voice Activity Detection) at AGW/TGW for efficient packet processing in IP network. It is known that voice packets from a CODEC with G.711- law generate *64Kbps* per each connection. The packet has a header with CRTP (Compression of Real Time Protocol) over Ethernet as a layer two network. It is usually assumed that the packet processing delay (D_{TC}) for encoding at source and decoding at receiver ranges from *50* to *80 msec*[10]. Let us assume a worst case in which D_{TC} is equal to *80msec*. Let us assume a delay time for an extra packet processing including the propagation delay D_p and D_{ex} to be *100ms*. Let us assume that the number of node in an end-to-end path is H=20. The target value for an end-to-end delay of a connection is assumed to be 100ms. Let us assume that the size of voice and data packets in order to compute the values of $1/\mu_v$ and $1/\mu_d$. As for $1/\mu_d$ let us assume twocases in order to compare the impact of design areas for different data packet sizes: First, the data packet has a fixed size of *1500bytes* (*1460bytes* of body and *40bytes* of header). Let us call this case as "Scenario A". Second, the data packet has a variable size of mean *500bytes* and standard deviation of *500 bytes*. Let us call this case as "Scenario B". The size of VoIP packet generated from G.711 Vocoder is assumed to be fixed and is equal to *216bytes* (*160bytes* of payload and *56bytes* of headers)[12]. Let us assume that the bandwidth of an output port in NAS is *10Mbps*.

Fig.4 illustrates the load map of voice and data. In Fig.4, there are two sticks in each point: The left one is the result for Scenario A, and the right one is the result for Scenario B. From Fig.4 we can find that the load map of an IP network that accommodates

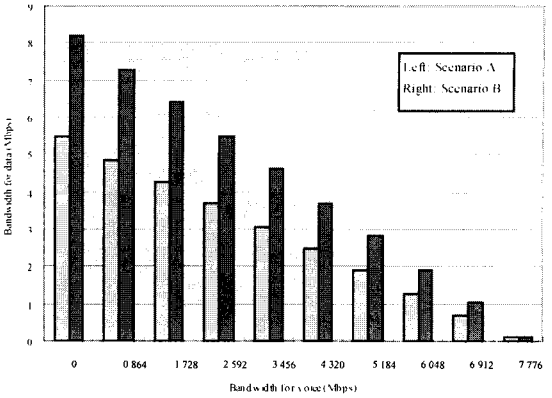


그림 4. 음성 및 데이터 부하도
Fig. 4. Load map for voice and data.

voice and data packets is heavily dependent on the characteristics of data packets such as the size and its distribution. Note that the required bandwidth for the bursty traffic is much higher than that of constant rate traffic. Therefore, a careful investigation on the characteristics of data packets such as the burstness has to be performed in order to set up a design rule for optimal provisioning of QoS for voice services in NGN.

Fig.5 illustrates the utilization of the network with respect to the load of voice traffic for the Scenario A. From Fig.5 we can find that the $C_v+C_d < C$ as we have argued in Fig.3, because utilization is smaller than one.

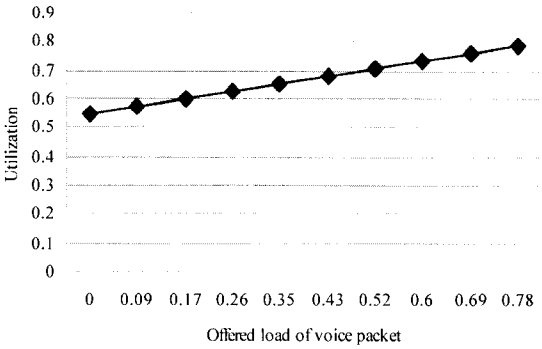


그림 5. 망 점유도
Fig. 5. Utilization of the network.

Ⅵ. Conclusions

In this work we proposed a method to determine the bandwidth of voice and data in a pipe in the access router of NGN VoIP network. The model is described in a hybrid of call and packet level: call level model determines the statistical characteristics of the number of customers in NGN network which is composed of PSTN customers, and packet level model determines the capacity of bandwidth required to satisfy QoS requirement of delay for voice packet. In deriving the call level traffic model we used the statistical method to obtain the mean and variance of the number of active customers in a time axis. The statistical model itself is not new, but it is considered to be useful in the modeling of traffic via measurement of real operating network. In the estimation of bandwidth in the IP networks, we used an M/G/1 queuing system with non-preemptive and strict priority scheme for voice traffic over data traffic, which is faithful to the service philosophy of the DiffServ service architecture of IP network. The information about expected number of active customers is useful in the computation of call level resource between PSTN and TGW/AGW, whereas the information about bandwidth Cv and Cdis useful in the estimation of the IP level resource at NAS.

Via some numerical experiments we showed the implication of the work in the design of the bandwidth of the access network in NGN voice over IP network. Themethod proposed in this work can be also used in the determination of whether the link is over-booked or not in a shared-service IP networks if one has multiple number of parameters for the offered load of voice and data traffic as well as the QoS parameters such as an end-to-end delay. Future research areas include the determination of the yardstick for judgment of over-booking as well as the sophistication of the measurement methods.

참 고 문 헌

[1] Jean-Yves Cochennex, "Activities on Next-Generation Networks under global information infrastructure in ITU-T", IEEE Communications, July 2002.

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

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

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

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

[6] G.V. Hoey, et. al, "Dimensioning of NGN transport networks for real-time voice applications", Alcatel Telecommunications Review, 2nd Quarter 2001.

[7] Telephone traffic theory Tables and charts, Part 1, Siemens Aktiengesellschaft, Berlin & Munchen, 1970.

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

[9] S. Ohtani, "Asking for the QoS of the VoIP, Telecommunication", March 2002 (In Japanese).

[10] 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).

[11] E. Nikolouzou, "Network services definition and deployment in a differentiated service architecture", Proc. ICC2002.

[12] 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.

[13] C.-G. Park, T.-W. Jung, Hoon Lee and B.-Y. Lee, "Interpretation on the definition of traffic for multimedia services", Proceedings of 1999 KICS Summer Conference, Korea.

[14] G.R. Cooper and C.D.McGillem, "Probabilistic methods of signal and system analysis" Third Ed., Oxford, 1999.

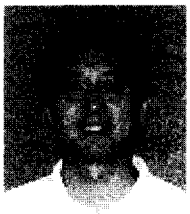
[15] O. Hersent, D. Gurle and J.-P. Petit, "IP Telephony, Packet-based multimedia communication systems" (Addison-Wesley), 2000.

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

[17] Hoon Lee and Yoon Uh, "Dimensioning Links for Voice over IP Networks", paper under submission.

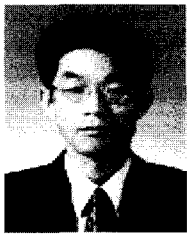
[18] C.-N. · Chuah and R. H. Katz, "Network provisioning & resource management for IP Telephony", http://www.cs.berkeley.edu/~chuah/research/paper/csd-99-1061.pdf.

저 자 소 개



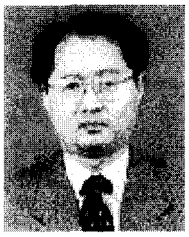
金 潤 基(正會員)

1984년 : 경북대학교 전자공학과(공학사). 1986년 : 경북대학교 전자공학과(공학석사). 1989년~현재 : KT 기술연구소 선임연구원, 실장 <주관심분야 : 차세대네트워크(NGN) 구조 및 설계, 차세대네트워크 제어 및 QoS 보장기술>



李 君(正會員)

1984년 : 경북대학교 전자공학과(공학사). 1986년 : 경북대학교 전자공학과(공학석사). 1996년 : 일본 동북대 통신공학과(공학박사). 1986년~2001년 : KT 연구개발본부 선임연구원. 2001년~현재 : 창원대 조교수 <주관심분야 : 트래픽엔지니어링, 네트워크 설계, 통신망 성능분석 및 QoS 설계>



李 廣 輝(正會員)

1983년 : 고려대학교 전자공학과(공학사). 1985년 : 고려대학교 전자공학과(공학석사). 1989년 : 고려대학교 전자공학과(공학박사). 1991년~1992년 : 영국 Wales 대학 및 Newbridge Networks 방문 연구원. 1994년~1995년 : 영국 UCL 방문 연구원. 1997년~1999년 : 영국 Reading 대학 방문 연구원. 2000년~현재 : Nortel Networks 방문 연구원. 1988년~현재 : 창원대 교수 <주관심분야 : 네트워크/QoS 관리, 정책기반 네트워킹, 멀티캐스트 프로토콜, 모바일 컴퓨팅>