Application-Aware Design to Enhance System Efficiency for VoIP Services in BWA Networks

Sung-Min Oh, *Student Member, IEEE*, and Jae-Hyun Kim, *Member, IEEE*

***Abstract*—This paper has designed a cross-layer framework for voice over Internet protocol (VoIP) services in IEEE 802.16 systems. It uses the application session information of the session description protocol to generate the quality of service parameters in IEEE 802.16 systems. This feature allows the system to efficiently allocate the radio resource because it can exactly estimate the properties of VoIP services such as packet-size and packet-generation-interval. In other words, the cross-layer framework is expected to achieve a novel resource request scheme for a VoIP service that dynamically assigns the radio resource. This paper has analyzed the maximum number of supportable VoIP users for the resource request schemes in terms of the packet-generation-interval in the silent-period, the duration of the silent-period, and the major VoIP speech codec. The numerical results show that the proposed scheme can efficiently support the VoIP services for the various communication environments. Particularly, it can improve the maximum number of supportable VoIP users by $14 \sim 93\%$ compared to an extended real-time polling service.**

***Index Terms*—Cross-layer framework, dynamic resource request, resource efficiency, voice over Internet protocol (VoIP) service.**

I. Introduction

VOICE over Internet protocol (VoIP) service is one of the most important services for next-generation wireless communication systems; it is also very sensitive to end-to-end transmission delays. For this reason, previous papers have focused on the fast transmission of VoIP packets [1], [2]. In recent years, various schemes that improve the efficiency of VoIP services in wireless communication networks have been investigated because VoIP traffic has variable bit rates [3]–[5].

VoIP traffic characteristics can be specified via the VoIP speech codec in the application layer. The main properties of the VoIP speech codecs are as follows [6]–[11]:

- packet size (PS) in a talk-spurt: variable [adaptive multiple rate (AMR), enhanced variable rate codec (EVRC)] or constant (G.7xx that means G.711, G.723.1, and G.729);
- packet generation interval (PGI) in a silent-period: random (G.7xx) or periodic (AMR, EVRC).

The PS and the PGI information of the VoIP speech codecs need to be shared at the medium access control (MAC) layer in order to support VoIP services with efficient usage of wireless resources. This indicates that the quality-of-service (QoS) parameters of the MAC layer are mapped with the traffic characteristics of the VoIP speech codec in the application layer. In addition, this can be associated with the wireless resource allocation scheme. However, this QoS information association process for VoIP services is an open issue.

IEEE 802.16 systems, which are strong candidates for the next wireless communication systems, have defined several uplink scheduling types such as unsolicited grant service (UGS) and extended real-time polling service (ertPS) for VoIP services. The systems have also specified the QoS parameter set required for uplink scheduling types. However, the QoS information sharing scheme in the application layer is out of scope in IEEE 802.16 standards. Moreover, the related work is also insufficient. For this reason, several inefficiencies in the IEEE 802.16 systems, such as the waste of wireless resources and the increase of end-to-end transmission delays, can occur. Therefore, this paper designs a cross-layer framework and proposes a dynamic resource request scheme for efficient usage of wireless resources and QoS provisioning.

II. Related Works

IEEE 802.16 standards have not defined a detailed description for the process of generating the QoS parameter set. For this reason, it is needed to investigate how to generate the QoS parameter set for VoIP services. Unfortunately, to the best of our knowledge, there are few related studies. In [12], the authors have considered the resource reservation protocol (RSVP) in order to make the QoS parameter set for a service flow. However, it is difficult to apply the RSVP to real systems because RSVP has a scalability problem. In addition, it is unable to achieve the exact PGI information for a VoIP speech codec from the QoS information of RSVP. For these reasons, the scheme proposed in [12] is inadequate to the QoS parameter generation scheme for VoIP services.

In IEEE 802.16 systems, several resource request schemes for VoIP services have been defined. Actually, UGS has been defined in the data over cable service interface specification (DOCSIS) in order to reduce the time to request an uplink resource for voice services with a constant bit rate [13]. UGS can cause a serious waste of wireless resources because the VoIP service has a variable bit rate. For this reason, in [3] and [4], the authors have studied about improving resource efficiency

Manuscript received May 18, 2010; revised September 08, 2010; accepted October 29, 2010. Date of publication November 18, 2010; date of current version January 19, 2011. This work was supported in part by the IT R&D program of MKE/KEIT (KI001822, Research on Ubiquitous Mobility Management Methods for Higher Service Availability) and in part by the MKE, Korea, under the ITRC support program supervised by the NIPA (NIPA-2010-(C1090-1021-0011)). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Paal Halvorsen.

The authors are with the School of Electrical and Computer Engineering, Ajou University, Suwon, Korea (e-mail: smallb01@ajou.ac.kr; jkim@ajou.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2010.2093512

for VoIP services; they have taken into account the variable PS. By the results analyzed in [3], real-time polling service (rtPS) and UGS-activity detection (AD) are inefficient to support VoIP services because these schemes use wireless resources to perform the polling scheme for every PGI. Thus, the authors have proposed a Piggyback-based resource request scheme. However, the proposed scheme can waste wireless resources because it periodically assigns a resource to send a packet during the silent-period. In reality, since the PGI during the silent-period is different from that during the talk-spurt for AMR and G.7xx, the resource periodically assigned during the silent-period can be wasted. In order to solve this problem, aperiodic resource request schemes have been proposed in [14] and [15]. In [14], ertPS has been defined to support VoIP services. ertPS requests a resource by sending a channel quality indicator channel (CQICH) codeword during the silent-period in order to send a packet without contention. However, ertPS has a weak point that it cannot explicitly request a required resource. In [15], the authors have proposed a resource request scheme based on ertPS. The proposed scheme explicitly requests a required resource by contention during the silent-period. In this paper, we call the scheme proposed in [15] as the contention-based VoIP resource request scheme (CVRS). Unfortunately, the latter scheme can cause deterioration of QoS performance due to the increase of the contention rate for heavy traffic load.

At this time, we would like to summarize the weaknesses of the previous researches as follows.

- There is no detailed description of the QoS parameter generation for VoIP services.
- The previous schemes are unable to obtain the exact PGI information for the VoIP speech codec in the application layer.
- It is difficult for the previous scheduling types to efficiently support VoIP services due to the various properties of PGI during the silent-period.

Therefore, this paper designs a cross-layer framework such that the IEEE 802.16 systems can use the traffic properties of VoIP services in the application layer. In addition, this paper defines the detailed procedures for generating a QoS parameter set and proposes the dynamic resource request scheme based on the QoS parameter set.

III. Background and Motivation

A. VoIP Traffic Model for Various VoIP Speech Codecs

This section represents traffic models for the major VoIP speech codecs in order to summarize the variety of the traffic properties for VoIP services. Traffic models are used to analyze system performance according to VoIP speech codecs. IEEE 802.16 Task Group m (TGm), which was approved by IEEE to develop an amendment to the IEEE 802.16 standard in 2006, published the draft evaluation methodology document in which the G.711, G.723.1, G.729, EVRC, and AMR are considered [16]. There are several common functions for VoIP speech codecs such as discontinuous transmission (DTX), voice activity detector (VAD), and comfort noise generation (CNG) in order to save bandwidth during the silent-period. By these functions, during the silent-period, the VoIP speech codecs send a silent descriptor (SID) packet that includes background noise information. On the other hand, each VoIP speech codec has specific properties. For G.7xx, voice packets are generated with a constant bit rate during the talk-spurt whereas the SID packets are randomly generated during the silent-period. Unlike G.7xx, for AMR and EVRC, voice packets are generated with variable bit rates according to the network condition during the talk-spurt. In addition, the SID packets are periodically generated during the silent-period. Particularly, PGI differs between the talk-spurt and silent-period for AMR. The main traffic properties of VoIP speech codecs can be modeled as follows.

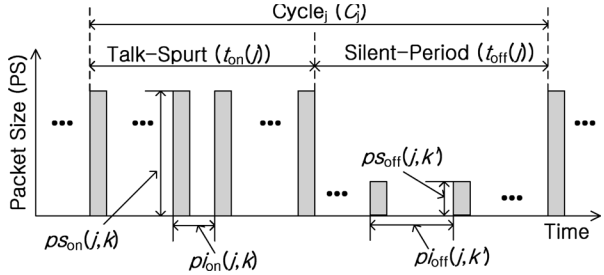


Fig. 1. VoIP traffic model according to the VoIP speech codec.

Fig. 1 represents the PS and PGI of VoIP traffic generated in the application layer. Here, since the silent-period is immediately started at the end of the talk-spurt, this paper defines the period of a talk-spurt and a silent-period as a cycle as shown in Fig. 1. C_j means the j_{th} cycle. C_j consists of the talk-spurt with $t_{on}(j)$ and the silent-period with $t_{off}(j)$. The size of the k_{th} packet in $t_{on}(j)$ and the size of the k'_{th} packet in $t_{off}(j)$ are defined as $ps_{on}(j,k)$ and $ps_{off}(j,k')$, respectively. This paper also defines the interval between the k_{th} packet and the $k+1_{th}$ packet in $t_{on}(j)$ and the interval between the $k'-1_{th}$ packet and the k'_{th} packet in $t_{off}(j)$ as $pi_{on}(j,k)$ and $pi_{off}(j,k')$, respectively. For each VoIP speech codec, the traffic models are as follows: The traffic model of G.7xx is represented as

$$\begin{cases} ps_{on}(j,k) = PS_{on} \\ pi_{on}(j,k) = PI_{on} \\ ps_{off}(j,k') = PS_{off} \\ pi_{off}(j,k') = p(k'), 0 < p(k') \le t_{off}(j) - \sum_{q=1}^{k'-1} p(q) \end{cases} \tag{1}$$

where $j = 1,2,\ldots,k = 1,2,\ldots,K_j$, and $k' = 1,2,\ldots,K'_j$. K_j and K'_j mean the total number of packets during $t_{on}(j)$ and $t_{off}(j)$, respectively. In addition, PS_{on}, PS_{off}, and PI_{on} are constants for the packet size during the talk-spurt, the packet size during the silent-period, and the packet generation interval during the talk-spurt, respectively. $p(k')$ is variable in terms of k' and it indicates the k'_{th} interval. Since EVRC uses three different PS as shown in Table I, PS can be represented as $E = \{e_i | e_{i+1} > e_i, e_i \text{ is constant for } i = 1,2,3\}$. e_3 indicates the size of an SID packet. The traffic model of EVRC is as follows:

$$\begin{cases} ps_{on}(j,k) \in \{e_1, e_2\} \\ pi_{on}(j,k) = PI_{on} \\ ps_{off}(j,k') = PS_{off} = e_3 \\ pi_{off}(j,k') = PI_{off} = PI_{on}. \end{cases} \tag{2}$$

TABLE I
PROPERTIES OF VOIP SPEECH CODECS

VoIP Speech Codec	PS (bytes)		PGI (msec)	
G.711	PS_{on}	160	PI_{on}	20
	PS_{off}	2	PI_{off}	random
G.723.1	PS_{on}	19.88	PI_{on}	30
	PS_{off}	2	PI_{off}	random
G.729	PS_{on}	10	PI_{on}	10
	PS_{off}	2	PI_{off}	random
EVRC	PS_{on}	21.375, 10	PI_{on}	20
	PS_{off}	2	PI_{off}	20
AMR	PS_{on}	11.857, 12.875, 14.75,16.75, 18.5, 19.875, 25.5, 30.5	PI_{on}	20
	PS_{off}	5	PI_{off}	160

Since AMR can also adjust the traffic rate, PS can be defined as $A = \{a_n | a_{n+1} > a_n, a_n \text{ is constant for } n = 1, 2, \ldots, 8\}$. The AMR traffic model is as follows:

$$\begin{cases} ps_{on}(j,k) \in A \\ pi_{on}(j,k) = PI_{on} \\ ps_{off}(j,k') = PS_{off} \\ pi_{off}(j,k') = PI_{off}. \end{cases} \quad (3)$$

Table I indicates the values of the constants for VoIP speech codecs.

B. Inefficiencies of the Conventional Resource Request Scheme for VoIP Service

This section describes the operation properties of conventional schemes in detail to show the shortcomings that can occur when the system does not consider all the traffic properties of VoIP speech codecs. In general, conventional resource request schemes of VoIP services use the default value of the grant-size (GS) and grant-interval (GI). Actually, ertPS can efficiently support a variety of PS by polling and piggybacking the bandwidth request message [3]–[5]. However, ertPS cannot deal effectively with a variety of PGI. As shown in Table I, the values of PI_{on} for all VoIP speech codecs are almost similar whereas the values of the packet generation intervals during the silent-period (PI_{off}) are remarkably different. Thus, it is expected that several inefficiencies can occur during the silent-period. Therefore, this paper mainly takes into account the resource request scheme in terms of the variety of PGI during the silent-period. This paper classifies the conventional schemes into the implicit resource request schemes and explicit resource request schemes.

1) Implicit Resource Request Scheme in Silent-Period: ertPS is the scheduling type that uses the implicit resource request scheme to send a VoIP packet during the silent-period. During the silent-period, a subscriber station (SS) sends the CQICH codeword to request the required uplink resource for the VoIP packet when the SS has a transmitting VoIP packet. Since a base station (BS) can know which SS requires the additional bandwidth by receiving the CQICH codeword, sending the CQICH codeword can avoid a contention. However, the CQICH codeword cannot include the information of the required bandwidth. Therefore, the BS allocates a maximum GS to the SS even though the SS needs a small-sized bandwidth when the BS receives a CQICH codeword from the SS. By this feature, an uplink resource can be wasted during the silent-period as shown in Fig. 2.

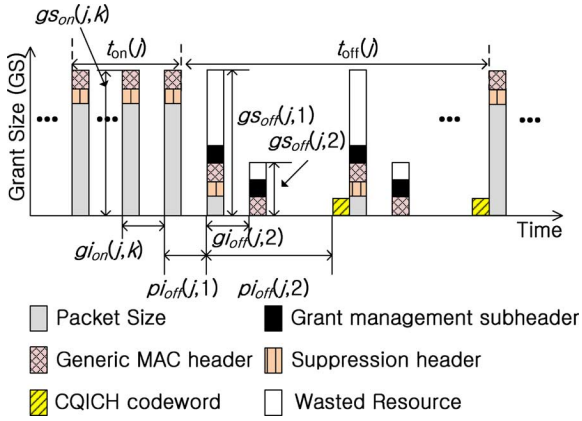


Fig. 2. Inefficiency of ertPS during the silent-period when PGI is different from GI.

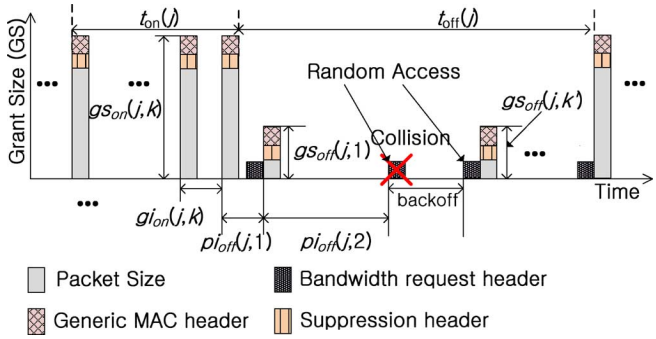


Fig. 3. Inefficiency of CVRS during the silent-period when PGI is different from GI.

Fig. 2 represents the inefficiency of ertPS during the silent-period. This paper considers the IP header suppression that the real-time protocol (RTP), user datagram protocol (UDP), and IP header are suppressed into a suppression header (SH) whose size is 3 bytes [4]. $gs_{on}(j,k)$ and $gs_{off}(j,l)$ mean the size of the k_{th} grant during talk-spurt and that of the l_{th} grant during the silent-period in the j_{th} cycle, respectively. $gi_{on}(j,k)$ and $gi_{off}(j,l)$ indicate the interval between the k_{th} grant and the $k+1_{th}$ grant and the interval between the $l-1_{th}$ grant and the l_{th} grant, respectively. At the beginning of the silent-period, an SS sends a VoIP packet with the grant management subheader (GMSH), that is one of subheader types in IEEE 802.16 standards, to reduce GS through $gs_{off}(j,1)$ and a BS then allocates the reduced grant ($gs_{off}(j,2)$) after GI ($gi_{off}(j,2)$). Actually, since GI is fixed in ertPS, $gi_{on}(j,k) = gi_{off}(j,l) = GI$ for $j = 1, 2, \ldots, k = 1, 2, \ldots, K_j$, and $l = 1, 2, \ldots, L_j$. GI is constant and L_j is the number of the grant during the silent-period. At this time, the SS has no transmitting packet in the VoIP queue; thus, it sends a GMSH whose bandwidth field is zero. By this operation, the BS does not allocate a grant to the SS. If the SS receives an SID packet from the upper layer, then the SS sends a CQICH codeword in order to inform the BS that the SS has a transmitting packet. The BS allocates a grant (gs_{off}(j, 3)) whose size is based on the *maximum sustained traffic rate* of

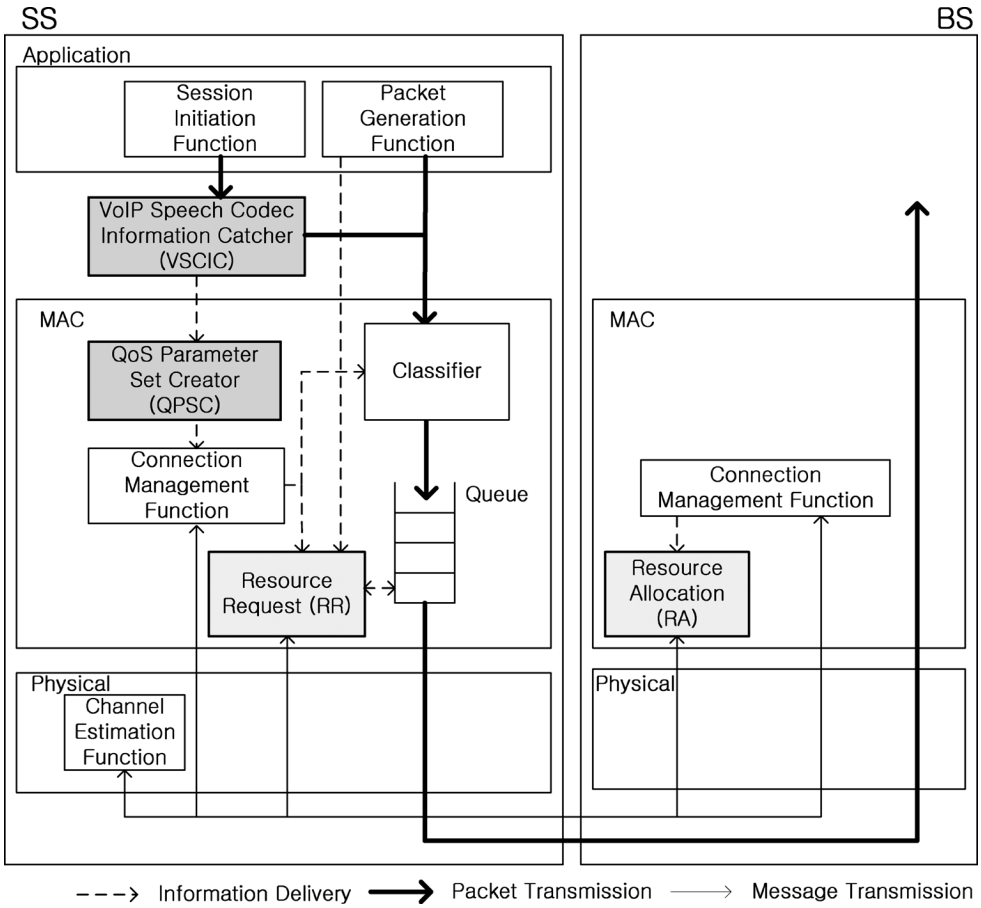


Fig. 4. Block diagram of the cross-layer framework.

the VoIP service. Since GS is larger than the size of the required grant [SID packet + SH + generic MAC header (GMH)], the SS sends a GMSH by piggybacking. After GI, the BS allocates a grant and the SS then sends a GMSH whose bandwidth field is zero. This process is iteratively performed up to the end of the silent-period. Therefore, ertPS can waste uplink resources for every packet transmission during the silent-period.

2) Explicit Resource Request Scheme in Silent-Period: In [15], the authors have proposed a CVRS that is similar to ertPS from the point of view of the resource request method for a variable PS. In CVRS, an SS explicitly requests the required uplink resource for an SID packet by the contention scheme during the silent-period. However, the contention scheme can cause the deterioration of the system efficiency and the QoS performance as the data rate increases. Fig. 3 describes the inefficiency of CVRS during the silent-period. During the silent-period, an SS sends a bandwidth request header (BRH), which is one of header types defined in IEEE 802.16 standards, to request the required uplink resource for the SID packet as shown in Fig. 3. This property has an advantage. The SS can request a grant ($gs_{off}(j,k')$) based on the size of the VoIP queue (SID packet + SH + GMH); thus, CVRS can prevent uplink resource waste. On the other hand, the property can cause CVRS inefficiencies. The SS has to send BRH for every packet transmission during the silent-period. In addition, the number of BRH transmission can be increased due to the collision of BRH. For this reason, BRH transmission can reduce system efficiency. With this disadvantage, CVRS can also deteriorate QoS performance. The reason for this is that the access delay can be increased due to the retransmission of BRH as shown in Fig. 3.

IV. Proposed Cross-Layer Framework for VoIP Services

In order to compensate for the inefficiencies of conventional schemes, this paper designs a cross-layer framework. Fig. 4 depicts the cross-layer framework for VoIP services. In Fig. 4, the VoIP speech codec information catcher (VSCIC) and the QoS parameter set creator (QPSC) are newly defined in this paper. The resource request (RR) and resource allocation (RA) are modified based on the conventional schemes. For the other functions in Fig. 4, this paper refers to the IEEE 802.16 systems and the conventional protocols [4], [14], [17]. VSCIS and QPSC are responsible for interworking QoS information between the MAC layer and the application layer. RR and RA are designed for the efficient use of uplink resources for VoIP services.

The brief descriptions of the conventional functions are as follows.

- **Session Initiation Function**: When a VoIP call is generated in the application layer, a session connection protocol is used to connect the VoIP call between VoIP users through end-to-end networks. This paper considers the session initiation protocol (SIP) as the session connection protocol

TABLE II
QOS PARAMETER SET MAPPING VOIP SCHEDULING TYPES

VoIP speech codec index	VoIP speech codec	*Maximum sustained traffic rate* (tr_{max})	*Minimum reserved traffic rate* (tr_{min})	*Unsolicited grant interval* (ugi)	*Unsolicited polling interval* (upi)
0	G.711	$(PS_{on}+SH)/PI_{on}$	$(PS_{on}+SH)/PI_{on}\cdot\alpha+(PS_{off}+SH)/E[pi_{off}]\cdot\beta$	PI_{on}	$E[pi_{off}]$
1	G.723.1				
2	G.729				
3	EVRC	$(e_1+SH)/PI_{on}$	$(e_1+SH)/PI_{on}\cdot\alpha+(PS_{off}+SH)/E[pi_{off}]\cdot\beta$		PI_{off}
4	AMR	$(e_1+SH)/PI_{on}$	$(a_1+SH)/PI_{on}\cdot\alpha+(PS_{off}+SH)/E[pi_{off}]\cdot\beta$		

[17] that performed this function. By this definition, this function can exchange several messages such as INVITE, 200 OK, BYE, etc.

- **Connection Management Function**: This function is responsible for the connection management defined in IEEE 802.16 systems. For the Connection Management Function, an SS and a BS exchange several messages, such as the dynamic service addition, change, and delete (DSA, DSC, and DSD) messages, in order to specify or manage a service flow. Particularly, this function performs the call admission control and allocates the connection and service flow identifier (CID and SFID), when a connection is generated. In addition, this function delivers CID, SFID, and QoS information to the Classifier and RR/RA for the QoS provisioning functions. QoS information is defined in Section IV-A.
- **Packet Generation Function**: This function includes the main functions of the VoIP speech codec in the application layer. Thus, this function generates a packet and performs DTX, VAD, and CNG according to the VoIP speech codecs. In addition, this function can deliver VAD information to the convergence sublayer (CS) of the MAC layer by primitives [4].
- **Classifier**: This function is responsible for the en/decapsulation of the MAC service data unit (SDU). For this property, this function has a mapping table that consists of IP header information (destination/source IP address, transport protocol, etc.) and the MAC information (CID, SFID, MAC address, etc.). If this function makes an MAC SDU using the mapping table when an IP packet is received from the upper layer, then it stores the MAC SDU into a queue.
- **Channel Estimation Function**: This function is responsible for estimating the channel quality of the packets transmitted through the downlink, and it sends the information of the estimated channel quality to a BS through CQICH. The BS can use this channel quality information for resource allocation.

A. Cross-Layer Interworking for VoIP Speech Codec Information

In order to use the information of the VoIP speech codec in the application layer, it is needed to recognize the context information in the application layer. In addition, it is required to translate from the application information into the system information defined in the MAC layer. For these requirements, this paper defines VSCIC and QPSC as shown in Fig. 4. The detailed description of each function is as follows.

1) VSCIC: VSCIC is located in a upper layer above the IP layer. It has three main properties as follows.

- It can recognize the context information of SIP messages.
- It can make the system information using context information.
- It can transmit system information to the MAC layer.

In the SIP, session description protocol (SDP) is used to deliver the QoS information for a service. In SDP, the "m" field can include information on media type, port number, transport protocol, and media format. For instance, the context of "$\mathrm{m}=\mathrm{audio}\,49170\,\mathrm{RTP/AVP}\,0$" means that the session media in the application layer is audio, the port number is 49170, the transport protocol is RTP, and "0" indicates that the VoIP speech codec is G.711 [18]. Due to this SDP property, this paper uses the session information of the "m" field to obtain the information of the VoIP session in the application layer. VSCIC makes the system information such as

$$\mathrm{VoIP\_SESSION\_INFO} = \{\mathrm{port\ number},\ \mathrm{VoIP\ speech\ codec\ index}\}$$

and transmits it to the QPSC of the MAC layer.

2) QPSC: The QPSC is responsible for generating the QoS parameters which are defined in IEEE 802.16 standards by using system information received from VSCIC. Table II represents the generation method of the main QoS parameters for VoIP scheduling types. QPSC can estimate the kinds of VoIP speech codecs in the application layer from the VoIP speech codec index delivered by VoIP_SESSION_INFO, as is shown in Table II. For this reason, QPSC can know PS and PGI of the VoIP speech codec by using Table I.

By using the VoIP speech codec information, the main QoS parameters can be generated as follows. Let $\bar{t}_{on}$ and $\bar{t}_{off}$ be the average of $t_{on}(j)$ and $t_{off}(j)$ for all j, respectively, and let $\alpha=\bar{t}_{on}/(\bar{t}_{on}+\bar{t}_{off})$ and $\beta=\bar{t}_{off}/(\bar{t}_{on}+\bar{t}_{off})$. The *maximum sustained traffic rate* ($tr_{\max}$) means the peak data rate of the service flow and the *minimum reserved traffic rate* ($tr_{\min}$) indicates the average data rate of the service flow. Thus, this paper can define the generation method as shown in Table II. Here, $E[pi_{off}]$ means the PGI average. For AMR and EVRC, we can denote that $E[pi_{off}]=PI_{off}$ because $pi_{off}(j,k')=PI_{off}$ for all j and k' as shown in Table I. For G.7xx, $E[pi_{off}]$ can be defined that $E[pi_{off}]=\bar{t}_{off}/E[K'_j]=\bar{t}_{off}/K'$. The *unsolicited grant interval* (ugi) is a QoS parameter for UGS and ertPS, and it usually means PGI in a talk-spurt. In general, the *unsolicited polling interval* (upi) is used for the polling scheme by the rtPS or nrtPS. However, this paper uses upi

to indicate the average of PGI during the silent-period. It is exploited to generate the system parameters required for the dynamic resource request scheme in Section IV-B. QPSC makes the system parameters such as "VoIP_QoS_PARA $= \{\text{port number}, tr_{\max}, tr_{\min}, ugi, upi\}$" and transmits them to the Connection Management Function.

B. Dynamic Resource Request Scheme

As described in Section IV-A, IEEE 802.16 systems can use the QoS information of VoIP speech codecs by using the cross-layer framework. Based on this property, this paper proposes the dynamic resource request scheme in order to compensate the inefficiencies of the conventional resource request schemes. The proposed scheme is performed in RA and RR in Fig. 4, and the main properties of the proposed scheme are as follows.

- It can generate the required parameters by using the system parameters that are delivered from the Connection Management Function.
- It stops the periodic resource allocation to save the wasted resource which can be occurred due to the variety of PGI in the silent-period.
- It uses two types of CQICH codeword such as "0b111101" and "0b111011". It sends the first one when it requests the resource for the SID packet during the silent-period, and it transmits the second one when it requests a resource for the voice packet when the silent-period is changed into the talk-spurt.

1) Required Parameters Generation Method: The proposed scheme needs to specify GS and GI for the talk-spurt and silent-period; thus, this paper uses VoIP_QoS_PARA to make it. $gi_{on}(j,k)$ in Fig. 3 is equal to GI for all j and k where GI is constant because a BS periodically allocates resources during the talk-spurt in the proposed scheme. $gs_{on}(j,k)$ is variable for j and k; however, this paper defines the default value of $gs_{on}(j,k)$ as GS_{on} where GS_{on} means the maximum size of the voice packet. In addition, since the fixed-size grant is allocated to an SS, $gs_{off}(j,k)$ can be defined as GS_{off} (constant) for the SID packet during the silent-period.

In summary, the required parameters for the proposed scheme are GI, GS_{on}, and GS_{off}. These parameters can be derived as (4), (5), and (6) where SH and GMH are the size of the IP suppression header and generic MAC header, respectively:

$$GI = ugi \tag{4}$$

$$GS_{on} = tr_{\max} \cdot ugi + SH + GMH \tag{5}$$

$$GS_{off} = \frac{(tr_{\min} - tr_{\max} \cdot \alpha) \cdot upi}{\beta} + SH + GMH. \tag{6}$$

2) Operational Properties of the Dynamic Resource Request Scheme: The dynamic resource request scheme requires adding the VAD field to GMSH as shown in Table III because the proposed algorithm applies two types of resource request schemes according to voice activity. In the talk-spurt, the proposed scheme uses the periodic resource allocation scheme whereas it uses the implicit resource request scheme during the silent-period.

Fig. 5 represents the operational properties of the proposed scheme, and the descriptions of the properties are as follows.

TABLE III
GRANT MANAGEMENT SUBHEADER FIELD

Name	Length (bit)	Description
Extended bandwidth request	10	Piggyback request.
FLI	1	Frame latency indication (FLI)
FL	4	Frame latency (FL)
VAD	1	Information for the voice activity 0: silent-period 1: talk-spurt

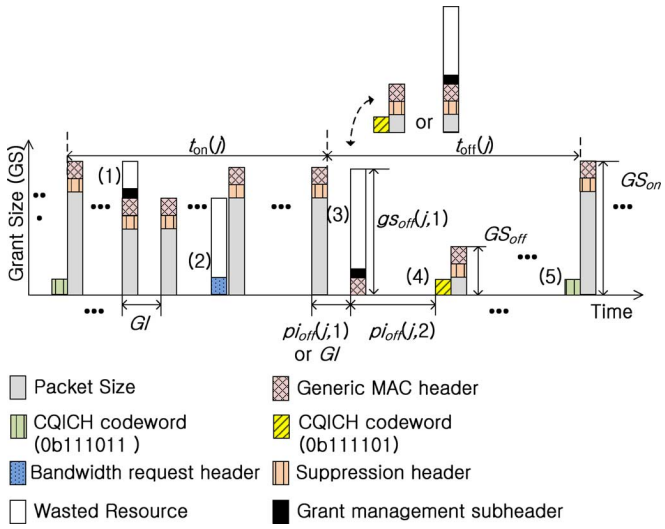


Fig. 5. Proposed dynamic resource request scheme.

1) In talk-spurt, if GS $>$ queue size: An SS sends GMSH with VAD $= 1$ to a BS by piggybacking in order to reduce GS. BS updates GS according to the required resource.
2) In talk-spurt, if GS $<$ queue size: An SS sends the BRH to a BS through the allocated resource in order to request the additional resource. BS allocates the required resource at the next frame, and SS can then transmit a voice packet.
3) In the case of the state change from talk-spurt to silent-period, there are three possibilities, because PGI during the silent-period is variable and BS can periodically allocate a resource in spite of the silent-period because BS cannot know the voice activity of the VoIP service.
   - $pi_{off}(j,1) > GI$: SS sends GMSH with VAD $= 0$ through the allocated resource in order to inform the BS of the change from talk-spurt to silent-period, because SS has no transmitting packets although a resource is allocated. BS stops the periodic resource allocation after receiving GMSH with VAD $= 0$.
   - $pi_{off}(j,1) = GI$: TSS sends an SID packet with GMSH with VAD $= 0$ through the allocated resource. BS stops the periodic resource allocation after receiving the GMSH with VAD $= 0$.
   - $pi_{off}(j,1) < GI$: SS sends the CQICH codeword (0b111101) to BS. BS stops the periodic resource allocation after receiving the CQICH codeword (0b111101) and it allocates GSoff at the next frame.
4) In the silent-period: If SS has a transmitting packet, then it sends the CQICH codeword (0b111101) to BS. BS allo-

cates GSoff at the next frame after receiving the CQICH codeword.

5) In the case of the state change from silent-period to talk-spurt: SS sends CQICH codeword (0b111011) to BS. BS allocates GSon to SS at the next frame after receiving the CQICH codeword. In addition, BS restarts the periodic resource allocation because it knows the change from silent-period to talk-spurt for the VoIP service by receiving the CQICH codeword (0b111011).

Consequently, the proposed scheme can overcome the inefficiencies of the conventional schemes such as the wasted uplink resource and the increase of access delay as mentioned in Section III.

C. System Complexity

The system complexity can be increased by the proposed cross-layer framework. There are two causes. Firstly, the VSCIC and the QPSC newly defined in the paper can increase the system complexity. However, the VSCIC and the QPSC are used only when a session is initiated. Specifically, these functions exploit the QoS information of the SIP and SDP widely used to manage a session, when a session is initiated. For these reasons, the system complexity which is increased by the newly defined functions may not be crucial. Secondly, in the case that multiple VoIP services are supported in an SS, the system complexity can be increased. The reason for this is that the MAC layer has to separately manage each VoIP service. However, this is an inherent property of IEEE 802.16 systems that have been designed to provide a service by using a service flow identifier (SFID) [14]. This means that IEEE 802.16 systems can separately manage each VoIP service in an SS. Hence, we can claim that the system complexity may not be increased by the cross-layer framework.

V. Performance Analysis and Numerical Results

The goal of the proposed cross-layer framework is to increase the system efficiency for VoIP services. For this reason, this paper analyzes VoIP capacity to compare the system efficiency among the resource request schemes. VoIP capacity means the maximum number of supportable VoIP users.

In order to obtain VoIP capacity, the average number of used slots ($\bar{S}$) for each resource request scheme is derived. For analysis simplicity, this paper assumes that the property of C_j is equal to the property of C_{j+1} for all j. This means that $t_{on}(j) = t_{on}$ and $t_{off}(j) = t_{off}$ where t_{on} and t_{off} are constants. In addition, $K_j = K$ and $K'_j = K'$ where K and K' are constants, we can present $pi_{off}(j, k')$ as $pi_{off}(k')$ for all j. Thus, since $E[pi_{off}(k')]$ is equal to t_{off}/K', $E[pi_{off}(k')]$ can be represented as pi_{off} (constant). By this assumption, we can define the average number of used slots for a second as

$$\bar{S} = \frac{\bar{S}_{on}}{GI_{on}} \cdot \alpha + \frac{\bar{S}_{off}}{pi_{off}} \cdot \beta \tag{7}$$

where $\bar{S}_{on}$ and $\bar{S}_{off}$ mean the average number of used slots for every GI_{on} and pi_{off}, respectively. Thus, we can define $\bar{S}_{on}$ as

$$\bar{S}_{on} = \frac{1}{M_l} \times \frac{1}{K} \times \sum_{k=1}^{K} gs_{on}(k) \tag{8}$$

where M_l is the number of bytes for a slot according to the level of the modulation and coding scheme (MCS). This paper denotes that $gs_{on}(k) = GS_{on}$ for all schemes because the operational method in the talk-spurt of ertPS, CVRS, and the proposed scheme is the same. Thus, $\bar{S}_{on}$ can be given as

$$\begin{aligned}\bar{S}_{on} &= \frac{1}{M_l} \times \frac{1}{K} \times \sum_{k=1}^{K} GS_{on} \\ &= \frac{GS_{on}}{M_l} = \frac{PS_{on} + SH + GMH}{M_l}.\end{aligned} \tag{9}$$

Since $\bar{S}_{off}$ is the average number of used slots for every pi_{off}, $\bar{S}_{off}$ can be defined as

$$\bar{S}_{off} = \frac{1}{M_l} \times \frac{1}{K'} \times \sum_{l=1}^{L} gs_{on}(l). \tag{10}$$

By using (7), (9), and (10), we have derived the average number of used slots ($\bar{S}$) for ertPS, CVRS, and the proposed scheme in the Appendix.

Let S_{tot} be the total number of slots in an uplink subframe. $\bar{S} \cdot F$ means the average number of slots used in an uplink subframe. For this reason, VoIP capacity (m) can be defined as

$$m = \frac{S_{tot}}{\bar{S} \cdot F}. \tag{11}$$

At this time, $\bar{S}$ for CVRS is a function with respect to m because N can be substituted by m in (26). Therefore, (11) can be defined as

$$m = \frac{S_{tot}}{\bar{S}_{CVRS}(m) \cdot F}. \tag{12}$$

Denote that $f(m) = m \cdot \bar{S}_{CVRS}(m) \cdot F - S_{tot}$. The differential of $f(m)$ with respect to m is as follows:

$$f'(m) = F \cdot (\bar{S}_{CVRS}(m) + m \cdot \bar{S}'_{CVRS}(m)). \tag{13}$$

Since $K' = t_{off}/pi_{off} > 1$ and $1 - 1/R < 1$, $\bar{S}_{CVRS}(m) > 0$ for all m. Moreover, $\bar{S}'_{CVRS}(m)$ can be $ln(1 - 1/R) \times (A \cdot (-B)) \cdot (1 - 1/R)^{1-Bm}$ where A and B are positive constants. For this reason, $\bar{S}'_{CVRS}(m) > 0$ for all m; thus, $f'(m) > 0$ for all m. For $m \geq 0$, $f(0) < 0$. Therefore, there exists a solution m that satisfies $f(m) = 0$ for $m \geq 0$, and we can obtain VoIP capacity for CVRS by solving (12).

In this section, we analyze the VoIP capacity to compare the system efficiency among the resource request schemes. In particular, we consider pi_{off} and t_{off} as the major factors that can have an impact on the performance of three schemes such as ertPS, CVRS, and the proposed scheme, because the main difference among the resource request schemes is the packet transmission method during the silent-period. In addition, this paper analyzes the performance of the resource request scheme according to the major VoIP speech codecs such as G.7xx, AMR,

TABLE IV
SYSTEM PARAMETERS

Attributes	Values
Bandwidth	10 MHz
Frame size	5 msec
Fast fourier transform (FFT)	1024 subcarriers
Number of used subcarrier	864 subcarriers
Number of data subcarrier	768 subcarriers
Number of pilot subcarrier	96 subcarriers
Number of subcarriers per subchannel	48 subcarriers
Number of data symbols per frame	44 symbols (UL: 15 symbols, DL: 29 symbols)
Number of symbols per slot	3 symbols
Number of subchannels per slot	1 subchannel
Modulation and coding scheme (MCS) level	QPSK 1/2
GMH	6 bytes
BRH	6 bytes
SH	3 bytes [4]
PS_{on}	20 bytes
PS_{off}	2 bytes
GI_{on}	0.03 sec
Q	1 sec

and EVRC. To calculate the VoIP capacity, the number of slots per a frame is needed. In orthogonal frequency division multiple access (OFDMA) systems, the number of slots per a frame can be calculated by multiplying the number of slots in frequency by the number of slots in symbols. We define the system parameters for the number of subcarriers and symbols as shown in Table IV. We refer to mobile WIMAX for the value of the system parameters [20]. By the system parameters, S_{tot} can be calculated as 90 slots. In addition, we define PS_{on}, PS_{off}, and GI_{on} as a constant as shown in Table IV because we would like to analyze the system performance in terms of pi_{off} and t_{off}.

A. Average Number of Used Slots

The average number of used slots indicates the number of slots that are used to support a VoIP service flow. By this definition, the increase of the average number of used slots means a reduction of system efficiency for a given wireless resource. For CVRS, the collision effect has to be considered, and thus this paper assumes that $R = 16$ slots and $N = 100$ users. This assumption affects only the collision effect of CVRS performance as shown in (26).

From Fig. 6, the average number of used slots decreases for general cases when pi_{off} or t_{off} increases. The reason for this is that the data rate is reduced by the increase of pi_{off} or t_{off}. However, for CVRS and ertPS, there are several cross-points of performance for each scheme when pi_{off} is varied from 0.03 to 0.05 s. The increase of t_{off} means that the portion of silent-period for the VoIP service is increased. The decrease of pi_{off} can cause the increase of SID packet transmission rate. In this case, performance can be deteriorated by the inefficiency of CVRS and ertPS as mentioned in Section III-B. However, we can find out that there is almost no inefficiency of the proposed scheme although the transmission rate of SID packets is high.

The gain of the proposed scheme can be clearly shown in Fig. 7, which indicates the ratio of the number of reduced slots by the proposed scheme to the number of used slots by the other schemes. This means that $D_{ertPS-Pro.} = (\bar{S}_{ertPS} - \bar{S}_{Pro.})/\bar{S}_{ertPS} \times 100$ and

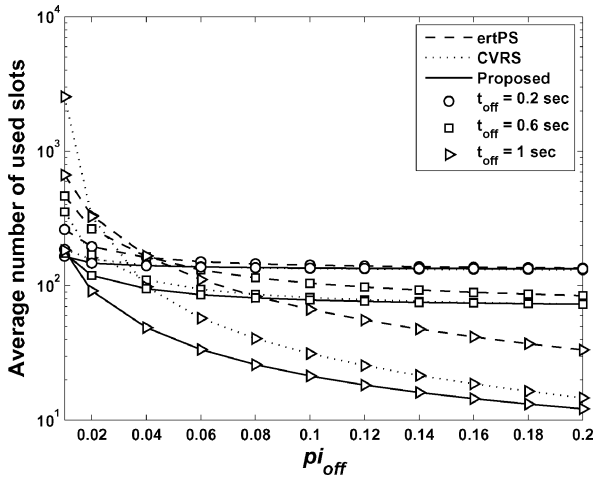


Fig. 6. Average number of used slots versus pi_{off} (seconds) ($R = 16$ slots and $N = 100$ users).

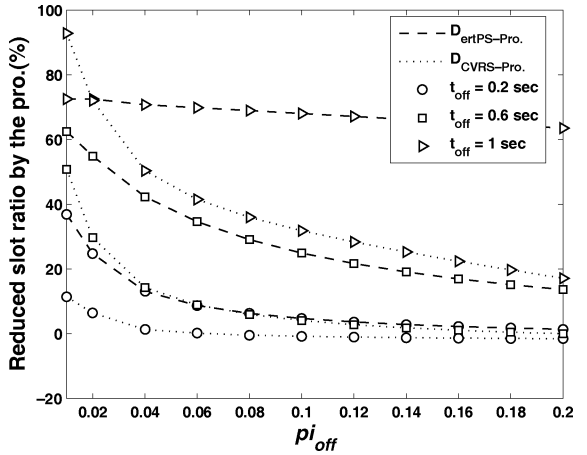


Fig. 7. Reduced slot ratio by the proposed scheme versus pi_{off} (seconds) ($R = 16$ slots and $N = 100$ users).

$D_{CVRS-Pro.} = (\bar{S}_{CVRS} - \bar{S}_{Pro.})/\bar{S}_{CVRS} \times 100$. The proposed scheme can save the uplink resource up to 70% compared to ertPS because it can compensate for ertPS inefficiency. For CVRS, the gain of the proposed scheme is lower than that of ertPS, except for one case. In the case that pi_{off} is lower than 0.02 and t_{off} is 1 s, the reduced slot ratio of CVRS is very high by $70 \sim 90\%$. This is caused by the increase of the collision rate for BRH transmission. In other cases, CVRS performance is superior to that of ertPS when the transmission rate of SID packets is low. Particularly, the gain of the proposed scheme is smaller than 40% for CVRS when pi_{off} is higher than 0.03 s. The reason is that the first SID packet can be sent through the maximum-sized grant in ertPS and the proposed scheme. Unlike the other schemes, in CVRS, the first SID packet is transmitted by sending a BRH. In general, the size of BRH is much smaller than that of the maximum-sized grant.

CVRS can have good performance when the transmission rate of SID packets is low as shown in Fig. 7; however, CVRS can suffer from collision that can occur during the transmission process of BRH. Figs. 8 and 9 indicate the average number of used slots for CVRS during a second according to pi_{off}, t_{off},

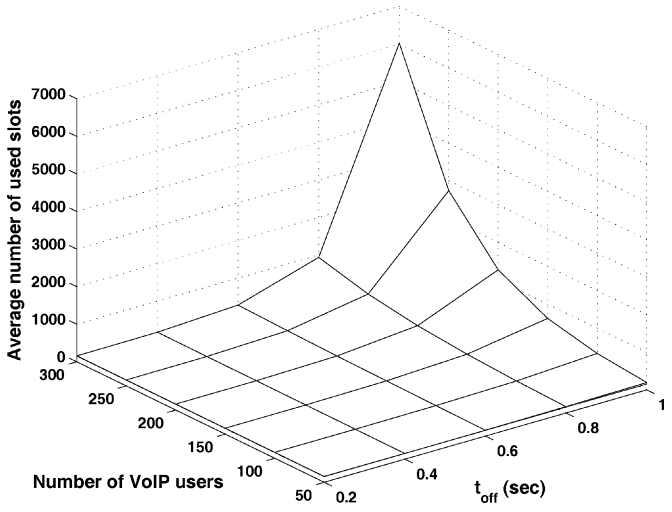


Fig. 8. Average number of used slots versus t_{off} and the number of VoIP users for the CVRS ($R = 16$ slots and $pi_{off} = 0.02$ s).

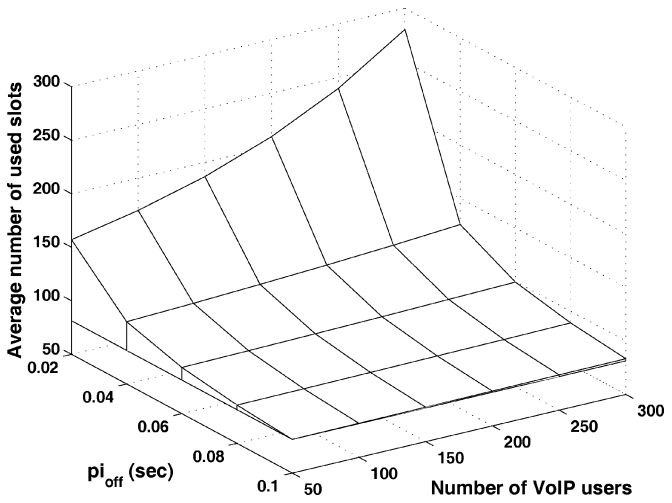


Fig. 9. Average number of used slots versus pi_{off} and the number of VoIP users for the CVRS ($R = 16$ slots and $t_{off} = 0.6$ s).

and the number of VoIP users. As shown in Fig. 8, the average number of used slots exponentially increases when t_{off} and the number of VoIP users increase. Particularly, the average number of used slots is larger than 1000 slots when t_{off} is longer than 0.8 s and the number of VoIP users is 300. If we consider that the total number of slots for a frame is 90 slots, we can easily infer that the performance of CVRS is seriously poor. In addition, the average number of used slots increases more rapidly by the increase of t_{off} when the number of VoIP users is large. This analytic result can also be found in Fig. 9. As shown in Fig. 9, the average number of used slots exponentially increases up to 250 slots by the decrease of pi_{off} when the number of VoIP users is 300. In other words, the number of VoIP users is one of the critical factors that can have an impact on performance. For this reason, the number of VoIP users should be considered when we analyze the VoIP capacity for CVRS.

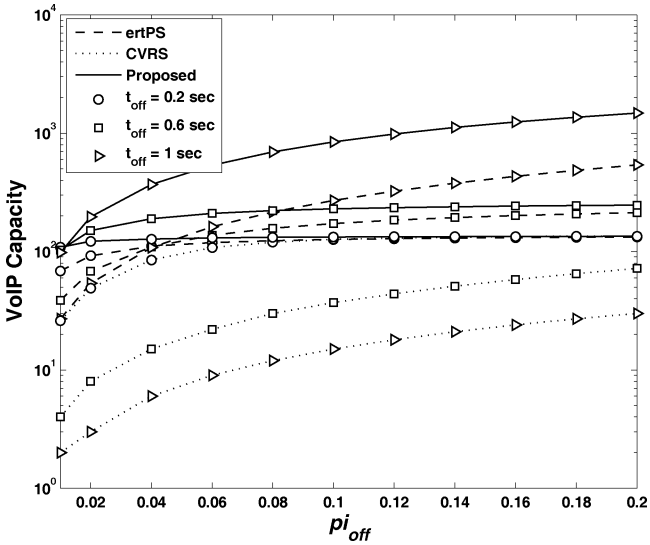


Fig. 10. VoIP capacity versus pi_{off} (seconds) ($R = 16$ slots).

B. VoIP Capacity

Fig. 10 represents the VoIP capacity that is obtained by using (11) for ertPS and the proposed scheme and (12) for CVRS. In Fig. 10, the numerical result for CVRS is obtained under $R = 16$ slots. This means that the numerical result for CVRS includes the collision effect. The VoIP capacity of the proposed scheme, ertPS, and CVRS is 235, 185, and 44, respectively, when pi_{off} is 0.1 s and t_{off} is 0.6 s. The inversion of the VoIP capacity is also generated when pi_{off} is shorter than 0.05 s as shown in Fig. 10. For ertPS and the proposed scheme, the VoIP capacity increases as t_{off} increases. On the other hand, the VoIP capacity of CVRS decreases as t_{off} increases because the average number of used slots exponentially increases due to the increase of the collision rate as shown in Figs. 8 and 9.

From now on, we have to take into account a system that can assign a sufficient R to support a service without the effect of collision probability. This paper considers a collision probability (δ) that does not affect system performance. By using (20), if we assign R with satisfying (14) for all tr, then the collision probability does not affect system performance:

$$R \geq \frac{1}{\left\{1-(1-\delta)^{1/(tr-1)}\right\}}. \tag{14}$$

The right side of (14) is a continuous and increasing function with respect to tr. We can assume that $pi_{off} \geq 0.01$ and $0 \leq \beta \leq 1$, because the minimal value of pi_{on} is 0.01 s as shown in Table I and Q is defined as 1 s. For these reasons, we can derive that $tr \leq 0.5N$ by using (23). By this property, (14) can be

$$R \geq \frac{1}{\left\{1-(1-\delta)^{1/(0.5\cdot N-1)}\right\}}. \tag{15}$$

Therefore, the minimal value of R can be obtained by using (15). Here, we assume that the maximum number of VoIP users in the system is 1500 and δ is equal to 0.01. Then, the minimal value of R is 7425. This means that the collision probability is lower than 0.01 regardless of the length of pi_{off} and t_{off} if the system assigns R as 7425.

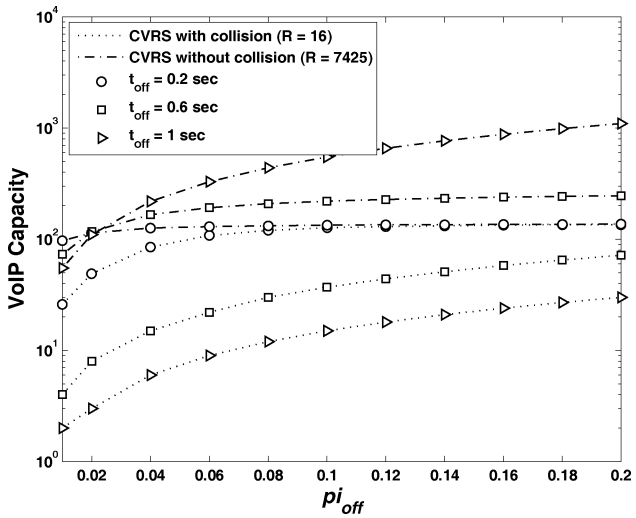


Fig. 11. VoIP capacity versus pi_{off} (seconds) ($R = 7425$ slots).

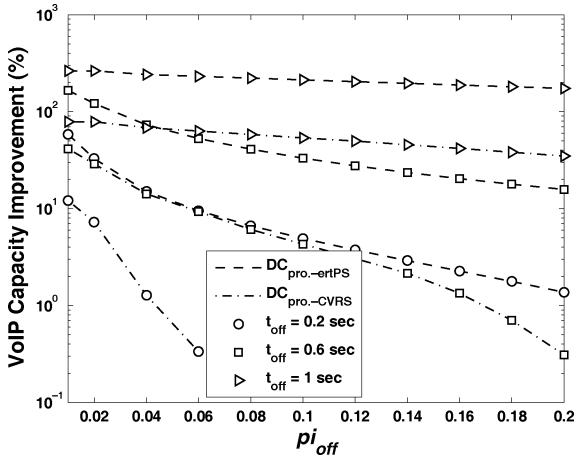


Fig. 12. VoIP capacity improvement by the proposed scheme (%) versus pi_{off} (seconds) ($R = 7425$ slots).

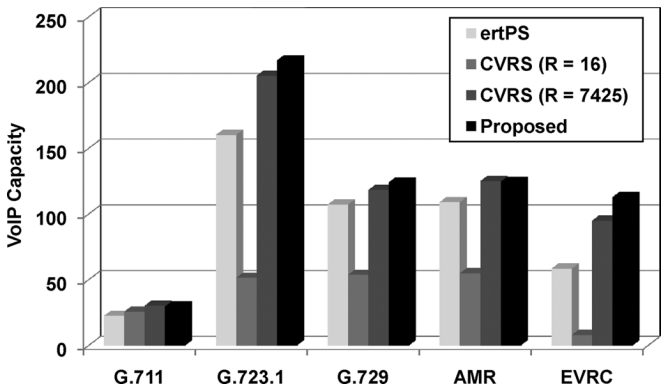


Fig. 13. VoIP capacity for the VoIP speech codecs ($t_{off} = 0.6$ s).

Fig. 11 represents the VoIP capacity for CVRS with or without the collision effect. VoIP capacity is seriously affected by the collision effect. As shown in Fig. 11, VoIP capacity for CVRS with the collision effect is 44 users when pi_{off} is 0.1 s and t_{off} is 0.6 s, whereas that without the collision effect is 228 users. The improved VoIP capacity increases as pi_{off} decreases.

Here, we need to analyze the VoIP capacity of CVRS without the collision effect and that of the other schemes. In order to analyze performance improvement by the proposed scheme, this paper defines the VoIP capacity improvement as $DC_{Pro.-ertPS} = (m_{Pro.} - m_{ertPS})/m_{ertPS} \times 100$ and $DC_{Pro.-CVRS} = (m_{Pro.} - m_{CVRS})/m_{CVRS} \times 100$. Fig. 12 describes the VoIP capacity improvement according to pi_{off}. Note that VoIP capacity for CVRS does not include the collision effect because the system assigns R as 7425. The proposed scheme can improve VoIP capacity by 38% compared to the CVRS without collision effect, when t_{off} is 0.6 s and pi_{off} is 0.01 s. However, the gain of the proposed scheme is smaller than 10% when t_{off} is 0.2 s or t_{off} is 0.6 s and pi_{off} is longer than 0.04 s. This means that the CVRS can also efficiently support a VoIP service in the case that the system can allocate a sufficient resource for random access and the transmission rate of the SID packet is low.

Fig. 13 represents the VoIP capacity for main VoIP speech codecs when t_{off} is equal to 0.6 s. In Fig. 13, the VoIP capacity of G.7xx means the average value in terms of pi_{off}. The numerical result shows that the proposed scheme can efficiently use the uplink resource for all VoIP speech codecs. For CVRS, the gap between the VoIP capacity with and without the collision effect is extraordinarily high. This means that CVRS is not adequate for the wireless communication systems whose conditions, e.g., the number of VoIP users, t_{off}, and pi_{off}, are dynamically changeable. ertPS can maintain VoIP capacity for various environments; however, its performance is lower than that of the proposed scheme by $10 \sim 50\%$. Therefore, it can be assured that the proposed scheme is the most efficient scheme for all VoIP speech codecs.

VI. Conclusion

This paper has proposed a cross-layer framework in order to use the information of VoIP speech codecs at the MAC layer. This paper has defined three main functions for the cross-layer framework. Firstly, it has designed the delivery function of the QoS information for VoIP services from the application layer to the MAC layer. Secondly, it has specified the QoS parameter generation function in the MAC layer using the information of the VoIP speech codec. Thirdly, it has proposed the dynamic resource request scheme in order to reduce the waste of the uplink resource.

By the numerical results, the proposed scheme can most efficiently support VoIP services. In particular, the proposed scheme can improve VoIP capacity by $14 \sim 93\%$ for the various VoIP speech codecs compared to the ertPS which is defined in the IEEE 802.16e standards. In addition, the performance of CVRS depends on the properties of the VoIP speech codecs and the communication environments. For this reason, CVRS is not adequate for wireless communication systems. Consequently, the proposed scheme is the most efficient scheme to support VoIP services.

The proposed scheme allows the system to interwork between the application layer and the MAC layer by using SIP. Thus, it is possible that the next generation communication systems, which consider SIP as the session initiation protocol, can use

the proposed scheme. In addition, the proposed scheme can be applied to video services because video services can be specified by the video codec in the application layer.

APPENDIX
AVERAGE NUMBER OF USED SLOTS FOR ertPS, CVRS, AND THE PROPOSED SCHEME

By the resource request model for ertPS described in Fig. 2, $\bar{S}_{off}$ can be defined as

$$\bar{S}_{off} = \frac{1}{M_l} \times \frac{1}{K'} \times \left\{ \sum_{k'=1}^{K'} (GS_{on} + GS_{off}) \right\}. \quad (16)$$

Thus

$$\begin{aligned} \bar{S}_{off} &= \frac{(GS_{on} + GS_{off})}{M_l} \\ &= \frac{PS_{on} + PS_{off} + 2 \cdot SH + 2 \cdot GMH}{M_l}. \end{aligned} \quad (17)$$

Here, let $t_{on} + t_{off} = Q$ where Q is constant, and then $\alpha = (Q - t_{off})/Q$ and $\beta = t_{off}/Q$. By using (7), (9), and (17), $\bar{S}$ for ertPS can be derived as

$$\begin{aligned} \bar{S}_{ertPS} = &\frac{(PS_{on} + SH + GMH) \cdot (Q - t_{off})}{M_l \cdot Q \cdot GI_{on}} \\ &+ \frac{(PS_{on} + PS_{off} + 2SH + 2GMH) \cdot t_{off}}{Q \cdot pi_{off}}. \end{aligned} \quad (18)$$

As shown in Fig. 3, CVRS transmits an SID packet by random access during the silent-period. For this reason, $\bar{S}_{off}$ for CVRS can defined as follows:

$$\begin{aligned} \bar{S}_{off} &= \frac{1}{M_l} \times \frac{1}{K'} \times \sum_{l=1}^{L} gs_{off}(l) \\ &= \frac{1}{M_l} \times \frac{1}{K'} \times \left\{ \sum_{k'=1}^{K'} (\bar{r} \cdot BRH + GS_{off}) \right\} \\ &= \frac{\bar{r} \cdot BRH + GS_{off}}{M_l} \\ &= \frac{\bar{r} \cdot BRH + PS_{off} + SH + GMH}{M_l} \end{aligned} \quad (19)$$

where BRH is the size of BRH and $\bar{r}$ is the average number of BRH transmissions to transmit an SID packet. In [19]

$$\bar{r} = \frac{1}{(1 - p_c)} \quad (20)$$

where p_c is the probability that a BRH experiences a collision in the slot. The collision probability can be given as $p_c = 1 - (1 - 1/R)^{(tr-1)}$, where R is the number of slots assigned for random access during an uplink subframe, and tr is the number of BRH at the beginning of the frame. tr can be given as

$$tr = \frac{K' \cdot \bar{n} \cdot F}{t_{on} + t_{off}} \quad (21)$$

where $\bar{n}$ is the average number of users during the silent-period, and F is the size of an MAC frame. Let N and n be defined as the total number of VoIP users and the number of users in the silent-period out of N users, respectively. By binomial distribution, $\bar{n}$ can be derived as

$$\begin{aligned} \bar{n} &= \sum_{n=1}^{N} n \cdot \Pr[n \text{ users are in silent-period}] \\ &= \sum_{n=1}^{N} n \cdot \binom{N}{n} \beta^n \alpha^{(N-n)} = N \cdot \beta. \end{aligned} \quad (22)$$

By putting (22) to (21)

$$tr = \frac{K' \cdot N \cdot \beta \cdot F}{t_{on} + t_{off}} = \frac{F \cdot N \cdot \beta^2}{pi_{off}}. \quad (23)$$

By using (19), (20), and (23)

$$\bar{S}_{off} = \frac{\frac{BRH}{\left(1 - \frac{1}{R}\right)^{F \cdot N \cdot \beta^2 / pi_{off} - 1}} + PS_{off} + SH + GMH}{M_l}. \quad (24)$$

By using (7) and (24), $\bar{S}$ for CVRS can be obtained as

$$\begin{aligned} \bar{S}_{CVRS} = &\frac{\bar{S}_{on} \cdot \alpha}{M_l \cdot GI_{on}} \\ &+ \frac{\left(\frac{BRH}{\left(1 - \frac{1}{R}\right)^{F \cdot N \cdot \beta^2 / pi_{off} - 1}} + PS_{off} + SH + GMH \right) \cdot \beta}{M_l \cdot pi_{off}}. \end{aligned} \quad (25)$$

Let (25) be summarized with respect to pi_{off}, we can obtain (26) and (27) at the bottom of the page. For the proposed scheme, $\bar{S}_{off}$ can be defined as

$$\bar{S}_{off} = \frac{1}{M_l} \times \frac{1}{K'} \times \sum_{l=1}^{L} gs_{off}(l)$$

$$\bar{S}_{CVRS} = \frac{1}{M_l} \cdot \left\{ \frac{(PS_{on} + SH + GMH) \cdot (Q - t_{off})}{Q \cdot GI_{on}} + \frac{(PS_{off} + SH + GMH) \cdot t_{off}}{Q \cdot pi_{off}} + \frac{BRH \cdot t_{off}}{\left(1 - \frac{1}{R}\right)^{F \cdot N \cdot \beta^2 / pi_{off} - 1} \cdot Q \cdot pi_{off}} \right\} \quad (26)$$

$$\bar{S}_{proposed} = \begin{cases} \frac{1}{M_l} \cdot \left\{ \frac{(PS_{on} + SH + GMH) \cdot (Q - t_{off})}{Q \cdot GI_{on}} + \frac{PS_{on} - PS_{off}}{Q} + \frac{(PS_{off} + SH + GMH) \cdot t_{off}}{Q \cdot pi_{off}} \right\}, & pi_{off} \geq GI_{on} \\ \frac{1}{M_l} \cdot \left\{ \frac{(PS_{on} + SH + GMH) \cdot (Q - t_{off})}{Q \cdot GI_{on}} + \frac{(PS_{off} + SH + GMH) \cdot t_{off}}{Q \cdot pi_{off}} \right\}, & pi_{off} < GI_{on} \end{cases} \quad (27)$$

$$= \begin{cases} \dfrac{GS_{on} + \sum_{k'=2}^{K'} GS_{off}}{M_l \cdot K'}, & pi_{off} \geq GI_{on} \\ \dfrac{\sum_{k'=1}^{K'} GS_{off}}{M_l \cdot K'}, & pi_{off} < GI_{on}. \end{cases} \quad (28)$$

By using (7) and (28), $\bar{S}$ for the proposed scheme can be derived as (27).

REFERENCES

[1] K. Wonghavarawat and A. Ganz, "Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems," *Int. J. Commun. Syst.*, vol. 16, pp. 81–96, 2003.

[2] G. Chu, D. Wang, and S. Mei, "A QoS architecture for the MAC protocol of IEEE 802.16 BWA system," in *Proc. IEEE ICCCAS*, Jul. 2002, vol. 1, pp. 435–439.

[3] H. W. Lee, T. S. Kwon, and D. H. Cho, "An efficient uplink scheduling algorithm for VoIP services in IEEE 802.16 BWA systems," in *Proc. IEEE Vehicular Technology Conf.*, 2004, vol. 5, pp. 3070–3074.

[4] H. W. Lee, T. S. Kwon, and D. H. Cho, "An enhanced uplink scheduling algorithm based on voice activity for VoIP services in IEEE 802.16d/e system," *IEEE Commun. Lett.*, vol. 8, no. 8, pp. 691–693, Aug. 2005.

[5] T. Kwon *et al.*, "Design and implementation of a simulator based on a cross-layer protocol between MAC and PHY layers in a WiBro compatible IEEE 802.16e OFDMA system," *IEEE Commun. Mag.*, vol. 43, no. 12, pp. 136–146, Dec. 2005.

[6] Annex A: Silence Compression Scheme, Int. Telecommun. Union, 1996, ITU-T Rec. G.723.1.

[7] Appendix II: A Comfort Noise Payload Definition for ITU-T G.711 Use in Packet-Based Multimedia Communication Systems, Int. Telecommun. Union, 2000, ITU-T Rec. G.711.

[8] Coding of Speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP), Int. Telecommun. Union, 2007, ITU-T Rec. G. 729.

[9] Technical Specification Group Services and System Aspects; Speech Codec Speech Processing Functions; AMR Wideband Speech Codec; Frame Structure, 2001, 3GPP TS 26.201 V5.0.0.

[10] Technical Specification Group Services and System Aspects; Mandatory Speech Codec Speech Processing Functions; Adaptive Multi-Rate (AMR) Speech Codec; Comfort Noise Aspects (Release 5), 2002, 3GPP TS 26.092 V5.0.0.

[11] Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems, 2004, 3GPP2 C.S0014-A.

[12] S.-E. Hong and O. H. Kwon, "Considerations for VoIP services in IEEE 802.16 broadband wireless access systems," in *Proc. Vehicular Technology Conf.*, 2006, vol. 3, pp. 1226–1230.

[13] Data Over Cable Service Interface Specification: Radio Frequency Interface Specification, 2000, SP-RFI v1.1-I04-000407.

[14] *IEEE Standard for Local and Metropolitan Area Networks—Part 16: Air Interface for Broadband Wireless Access Systems*, IEEE 802.16e™-2009, May 2009.

[15] S.-M. Oh, S. Cho, J.-H. Kwun, and J.-H. Kim, "VoIP scheduling algorithm for AMR speech codec in IEEE 802.16e/m system," *IEEE Commun. Lett.*, vol. 12, no. 5, pp. 374–376, May 2008.

[16] R. Srinivasan *et al.*, Draft IEEE 802.16m Evaluation Methodology Document, 2007.

[17] WiMAX Forum Network Architecture; Architecture, Detailed Protocols and Procedures IP Multimedia Subsystems (IMS) Interworking, 2009, Draft-T33-101-R015v01-C.

[18] M. Handley and V. Jacobson, SDP: Session Description Protocol, 1998, RFC 2327.

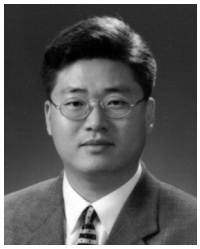
[19] S.-M. Oh and J.-H. Kim, "The analysis of the optimal contention period for broadband wireless access network," in *Proc. PWN05*, Mar. 2005, pp. 215–220.

[20] "Mobile WiMAX – Part I: A technical overview and performance evaluation," in *Proc. WiMAX Forum*, Aug. 2006.

Sung-Min Oh (S'07) received the B.S. and M. S. degrees in electrical engineering from Ajou University, Suwon, Korea, in 2004 and 2006, respectively. He is pursuing the Ph.D. degree in electrical engineering at Ajou University.

His research interests QoS performance analysis and 4G network.

Mr. Oh is a member of the Korean Institute of Communication Sciences (KICS).



Jae-Hyun Kim (M'90) received the B.S., M.S., and Ph.D. degrees, all in computer science and engineering, from Hanyang University, Ansan, Korea, in 1991, 1993, and 1996 respectively.

In 1996, he was with the Communication Research Laboratory, Tokyo, Japan, as a visiting scholar. From April 1997 to October 1998, he was a post-doctoral fellow at the Department of Electrical Engineering, University of California, Los Angeles. From November 1998 to February 2003, he worked as a member of technical staff in the Performance Modeling and QoS Management Department, Bell Laboratories, Lucent Technologies, Holmdel, NJ. He has been with the School of Electrical and Computer Engineering, Ajou University, Suwon, Korea, as an Associate Professor since 2003. His research interests include QoS issues and cross-layer optimization for wireless communication.

Dr. Kim was the recipient of the LGIC Thesis Prize and Samsung HumanTech Thesis Prize in 1993 and 1997, respectively. He is a member of the Korean Institute of Communication Sciences (KICS), the Korea Institute of Telematics and Electronis (KITE), and the Korea Information Science Society (KISS). He also has been serving as Vice Chair in the Information Service Committee at IEEE Comsoc Asia Pacific Board since 2008. More information about him can be found at http://ajou.ac.kr/ jkim.