IEICE TRANS. COMMUN., VOL.E92–B, NO.6 JUNE 2009

PAPER

Performance Analysis of the ertPS Algorithm and Enhanced ertPS Algorithm for VoIP Services in IEEE 802.16e Systems*

Bong Joo KIM[†], *Nonmember and* **Gang Uk HWANG**[†a)], *Member*

SUMMARY In this paper, we analyze the extended real-time Polling Service (ertPS) algorithm in IEEE 802.16e systems, which is designed to support Voice-over-Internet-Protocol (VoIP) services with data packets of various sizes and silence suppression. The analysis uses a two-dimensional Markov Chain, where the grant size and the voice packet state are considered, and an approximation formula for the total throughput in the ertPS algorithm is derived. Next, to improve the performance of the ertPS algorithm, we propose an enhanced uplink resource allocation algorithm, called the e^{2}rtPS algorithm, for VoIP services in IEEE 802.16e systems. The e^{2}rtPS algorithm considers the queue status information and tries to alleviate the queue congestion as soon as possible by using remaining network resources. Numerical results are provided to show the accuracy of the approximation analysis for the ertPS algorithm and to verify the effectiveness of the e^{2}rtPS algorithm.

key words: *Voice-over-Internet Protocol (VoIP), ertPS scheduling algorithm, IEEE 802.16e system, resource allocation, Quality of Service (QoS)*

1. Introduction

Recently, VoIP (Voice-over-Internet-Protocol) services have attracted a great deal of attention in the design of wireless networks including IEEE 802.16 broadband wireless networks, and there are a few studies which have examined the support of VoIP services over IEEE 802.16 systems. Lee et al. [4] proposed an enhanced uplink scheduling algorithm by using a Grant-Me (GM) bit for VoIP services in IEEE 802.16d/e systems. They [5] also proposed the extended rtPS (ertPS) algorithm adopted in the IEEE 802.16e standard. Yanfeng and Aiqun [7] used the adaptive linear prediction method to estimate the number of active VoIP services and to allocate suitable resources for VoIP services in IEEE 802.16d networks. Zhang et al. [8] proposed a new ertPS scheduling mechanism based on a multi-polling way in which the Base Station (BS) grants resources to user groups, rather than to each user.

The IEEE 802.16e standard [1] provides five uplink scheduling algorithms, Unsolicited Grant Service (UGS), real-time Polling Service (rtPS), extended real-time Polling Service (ertPS), non-real time Polling Service (nrtPS), and Best Effort (BE) service to support various multimedia applications having different QoS requirements. Among them, three algorithms, UGS, rtPS, and ertPS, are suitable for real-time multimedia applications, such as VoIP services. Especially, the ertPS algorithm is a scheduling mechanism which takes the merits of UGS and rtPS algorithms [1], and with this algorithm, the BS can dynamically provide resources to a Subscriber Station (SS). Hence, the ertPS algorithm uses resources more efficiently and supports more users than the UGS and rtPS algorithms [5], [6].

In [5], [6] Lee et al. evaluated the resource utilization and throughput of the ertPS algorithm to show its effectiveness. Zhao and Shen [9] also studied the performance of voice packet transmission and the resource utilization when the ertPS algorithm is used in IEEE 802.16-based backhaul networks. However, to the best of the authors' knowledge, all relevant previous studies used simulations and did not consider any mathematical model to analyze the performance of the ertPS algorithm. In this paper, we develop a two-dimensional discrete-time Markov Chain considering both the grant size (the amount of resources) and the voice packet state. By using it, we analyze the performance of the ertPS algorithm mathematically and obtain an approximation formula for the total throughput in the ertPS algorithm.

Although the ertPS algorithm is an efficient scheme for VoIP services with data packets of various sizes and silence suppression, it allows the SS to use resources only at predefined periodic slot times. So, even when the SS has packets to transmit in its queue and there are remaining (i.e., unused) resources at a slot time, if the slot time is not assigned to the SS, it can not use the resources. Since VoIP services are delay sensitive and require a stringent end-to-end delay bound, it is important to serve the packets of VoIP services as soon as possible so that they are not suffered from queueing delay. In addition, it would be efficient in terms of the resource utilization of the system that remaining resources are assigned to SSs who need to but are not allowed to transmit packets. By additionally assigning the remaining resources to SSs whenever possible, SSs can reduce the packet queueing delay and, accordingly, the maximum number of SSs with VoIP services that satisfy the given maximum delay bound can be increased. With these motivations, we propose an enhanced resource allocation algorithm in IEEE 802.16e systems.

In our proposed resource allocation algorithm, called the enhanced ertPS (e^{2}rtPS) algorithm, an SS informs the BS of its queue status information on whether its queue

Manuscript received December 3, 2007.
Manuscript revised August 18, 2008.

[†]The authors are with the Department of Mathematical Sciences and Telecommunication Engineering Program in the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea.

*This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2008-313-C00087).

a) E-mail: guhwang@kaist.edu

DOI: 10.1587/transcom.E92.B.2000

length is not less than an *a priori* given threshold or not. For this purpose, the e^{2}rtPS algorithm uses the reserved bit in the Generic MAC Header (GMH) of IEEE 802.16e. Then, the BS becomes aware of the queue congestion of the SS, and additional resources can be assigned to the SS if there are remaining resources.

The remainder of this paper is organized as follows. In Sect. 2, we describe the system model for the ertPS algorithm and analyze it using a two-dimensional Markov Chain, where the grant size and the voice packet state are considered. We then obtain an approximation formula for the total throughput in the ertPS algorithm. In Sect. 3, we propose the e^{2}rtPS algorithm, and then we provide numerical results to verify the accuracy of our approximation analysis for the ertPS algorithm and to examine the effectiveness of the e^{2}rtPS algorithm in terms of the total throughput and transmission delay in Sect. 4. Finally, we give our conclusions in Sect. 5.

2. System Model and Analysis of the ertPS Algorithm

In this section, we consider the ertPS algorithm for VoIP services. We assume that physical frame duration times, called frame slots, are of fixed length 5 ms, which is generally used in IEEE 802.16e systems. For our analysis in this section, we assume that 4 frame slots (20 ms) are considered as a unit time. So, the time axis is divided into unit times of length 20 ms and time is indexed by $t = 0, 1, 2, \cdots$. Each SS in the network generates VoIP packets and has a queue of infinite size to accommodate the generated VoIP packets. The queue service discipline is First-In-First-Out (FIFO).

2.1 The ertPS Algorithm

The ertPS algorithm is designed to support real-time multimedia applications that periodically generate data packets of various sizes, such as VoIP services with silence suppression [1], [5]. In this algorithm the BS allocates resources to an SS periodically, e.g., every 4 frame slots, and the BS changes the amount of allocated resources to the SS only when it requests a different amount of resources. We use the term "granted frame slots" to denote the frame slots at which the BS assigns resources to the SS of interest.

The assigned resources to the SS can be controlled by using the Grant management SubHeader (GSH) and Bandwidth Request Header (BRH). The formats of the GSH and BRH are given in Table 1 and Fig. 1, respectively. When the size of a voice packet decreases, the SS informs the BS of the required resources by using the extended piggyback request field of the GSH. On the other hand, when the size of a voice packet increases, the SS informs the BS of the required resources by using the BR MSB and BR LSB fields of the BRH [1]. In this case, when the BS can not assign the required resources to the SS at the current granted frame slot due to lack of remaining resources, the SS can be assigned resources at the nearest available frame slot before the next granted frame slot. Refer to Fig. 2. By using GSH and BRH, the BS controls the SS's grant size (the amount of assigned resources) to be as large as the size of the packet to be transmitted. The operation of the ertPS algorithm is shown in Fig. 2.

2.2 Voice Packet Arrival

We assume that each SS uses the EVRC (Enhanced Variable Rate Codec) with variable data rates and silence suppression for VoIP service [2]. The EVRC generates one voice packet every 20 ms with four different rates, i.e., full rate (Rate 1), half rate (Rate 1/2), quarter rate (Rate 1/4), and

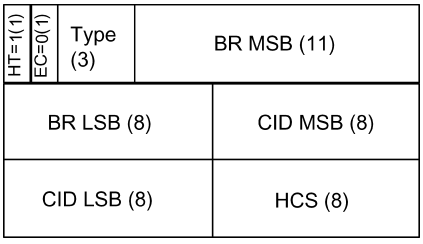


Fig. 1 The Bandwidth Request Header (BRH) format.

Table 1 The Grant management SubHeader (GSH) format.

syntax	size	Notes
Grant Management Subheader {	-	-
if (scheduling service type == UGS) {	-	-
SI	1 bit	-
PM	1 bit	-
FLI	1 bit	-
FL	4 bits	-
Reserved	9 bits	shall be set to zero
} elseif (scheduling service type == Extended rtPS) {	-	-
Extended piggyback request	11 bits	-
FLI	1 bit	-
FL	4 bits	-
} else {	-	-
Piggyback request	16 bits	-
}	-	-
}	-	-

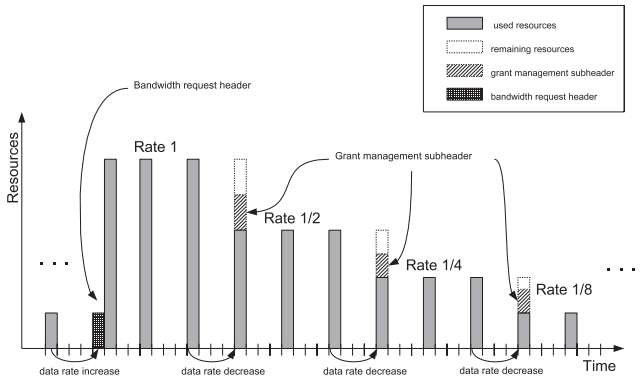


Fig. 2 The operation of the ertPS algorithm.

Table 2 The EVRC parameters.

rate type	voice packet size (bits)	data rate (kbps)	probability
Rate 1	171 (L_1)	8.6	0.29 (p_1)
Rate 1/2	80 (L_2)	4.3	0.04 (p_2)
Rate 1/4	40 (L_3)	2.1	0.07 (p_3)
Rate 1/8	16 (L_4)	1.0	0.60 (p_4)

Table 3 The definition of states in the first order Markov chain.

s(t)	R(t-1)	R(t)
0	1	1
1	1	1/2
2	1	1/4
3	1	1/8
4	1/2	1
5	1/2	1/2
6	1/2	1/4
7	1/2	1/8
8	1/4	1
9	1/4	1/2
10	1/4	1/4
11	1/4	1/8
12	1/8	1
13	1/8	1/2
14	1/8	1/4
15	1/8	1/8

($R(t)$ is the data rate at time t)

eighth rate (Rate 1/8). The full, half, and quarter rates are regarded as the voice data rates during on (talk-spurt) periods, and the eighth rate is regarded as the voice data rate during off (silence) periods. This results in a voice activity factor ($\triangleq \frac{\text{the expected on period}}{\text{the expected on period+the expected off period}}$) of 0.4. The detailed EVRC parameters are given in Table 2. Note that the generated voice packet size is variable according to the rate type of the EVRC, as given in Table 2.

According to [3], a first-order Markov Chain for the packet state is defined as shown in Table 3. Note that the packet state, denoted by $s(t)$, is defined based on the respective rate types with which the previous packet and the current packet (i.e., the Head Of Line (HOL) packet in the queue of the SS of interest) at time t were generated. In Table 3, $R(t)$ denotes the rate type with which the HOL packet at time t was generated. Then, by the definition of $s(t)$, if the packet state $s(t-1)$ is equal to k, then the packet state $s(t)$ can be one of four states as follows (refer to Table 3):

$$s(t) \in \{4 \times (k \mod 4), 4 \times (k \mod 4) + 1, \\ 4 \times (k \mod 4) + 2, 4 \times (k \mod 4) + 3\}.$$

For instance, if $s(t-1) = 3$, then $s(t) \in \{12, 13, 14, 15\}$. Note that the Markov Chain has 16 states. When a packet state $s(t)$ is chosen, the data rate, $R(t)$, is given by $\frac{1}{2^{(s(t) \bmod 4)}} (\in \mathcal{X} \triangleq \{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}\})$. We can obtain one step transition probabilities $a_{sk} \triangleq P[s(t) = k | s(t-1) = s]$ where $s, k \in \mathcal{S} \triangleq \{0, 1, \cdots, 15\}$, as given in [3], which will be used later in our analysis.

2.3 Analysis of the ertPS

To model an SS with the ertPS algorithm, we assume that the granted frame slots of the SS occur at $t = 0, 1, 2, \cdots$. Note that a unit time is 20 ms long by our assumption. Then, the SS is assigned resources at each time $t = 0, 1, 2, \cdots$. Let $g(t)$ denote a random variable representing the grant size of the SS assigned by the BS at time t. For convenience, we let the state space of $g(t)$ be the same as that of $R(t)$.

Note that the grant size $g(t)$ at time t is affected by both the packet state and the grant size at time $t-1$. In addition, a voice packet can be transmitted or kept in the queue at time t according to the values of $g(t)$ and $s(t)$ in the ertPS algorithm. Therefore, in order to analyze the system behavior, we consider a two-dimensional discrete-time Markov Chain $\{g(t), s(t)\}$. With the state pair $\{g(t), s(t)\}$, the following types of transitions can occur in the ertPS algorithm:

$$\begin{cases} [g(t) = i, s(t) = k | g(t-1) = i, s(t-1) = s] \\ \qquad \text{if } i = \frac{1}{2^{(s \bmod 4)}}, k \in \mathcal{S}, \\ [g(t) = j, s(t) = k | g(t-1) = i, s(t-1) = s] \\ \qquad \text{if } i > \frac{1}{2^{(s \bmod 4)}} = j, k \in \mathcal{S}, \\ [g(t) = i, s(t) = s | g(t-1) = i, s(t-1) = s] \\ \qquad \text{if } i < \frac{1}{2^{(s \bmod 4)}}, \\ [g(t) = j, s(t) = k | g(t-1) = i, s(t-1) = s] \\ \qquad \text{if } i < \frac{1}{2^{(s \bmod 4)}} = j, k \in \mathcal{S} \end{cases} \quad (1)$$

for $i \in \mathcal{X}$ and $s \in \mathcal{S}$. The first transition type in (1) is for the case where the grant size is equal to the HOL packet size at time $t-1$, and, accordingly, the HOL packet is transmitted at time $t-1$. The grant size is not changed in this case, i.e., $g(t-1) = g(t)$. The second transition type is for the case where the HOL packet size is smaller than the grant size at time $t-1$. So, in this case, the HOL packet is transmitted at time $t-1$, but the grant size is decreased at time t. The third and fourth transition types are for the cases where the HOL packet size is greater than the grant size at time $t-1$. So, voice packet transmission is not possible due to the insufficiency of allocated resources at time $t-1$. The third transition type occurs when the BS can not also assign the requested amount of resources until time t. Note here that the length of a unit time in our analysis is

20 ms. Since the ertPS algorithm is operated based on frame slots of length 5 ms, the SS can have more chances (exactly three more chances) to obtain resources for packet transmission between two time points, $t-1$ and t. On the other hand, the fourth transition type occurs when the HOL packet is transmitted before time t, i.e., the BS can assign the requested amount of resources before time t. Let $P_{(i,s),(j,k)}$ denote transition probabilities from $\{g(t-1)=i, s(t-1)=s\}$ to $\{g(t)=j, s(t)=k\}$, where $(i,s)\in\mathcal{X}\times\mathcal{S}$ and $(j,k)\in\mathcal{X}\times\mathcal{S}$. We then have the state transition probability matrix in block form as follows:

$$\mathbf{P}=\begin{bmatrix} \mathbf{T}_{1,1} & \mathbf{T}_{1,\frac{1}{2}} & \mathbf{T}_{1,\frac{1}{4}} & \mathbf{T}_{1,\frac{1}{8}} \\ \mathbf{T}_{\frac{1}{2},1} & \mathbf{T}_{\frac{1}{2},\frac{1}{2}} & \mathbf{T}_{\frac{1}{2},\frac{1}{4}} & \mathbf{T}_{\frac{1}{2},\frac{1}{8}} \\ \mathbf{T}_{\frac{1}{4},1} & \mathbf{T}_{\frac{1}{4},\frac{1}{2}} & \mathbf{T}_{\frac{1}{4},\frac{1}{4}} & \mathbf{T}_{\frac{1}{4},\frac{1}{8}} \\ \mathbf{T}_{\frac{1}{8},1} & \mathbf{T}_{\frac{1}{8},\frac{1}{2}} & \mathbf{T}_{\frac{1}{8},\frac{1}{4}} & \mathbf{T}_{\frac{1}{8},\frac{1}{8}} \end{bmatrix},$$

where the submatrix $\mathbf{T}_{i,j}$ is defined as

$$\mathbf{T}_{i,j}=\begin{bmatrix} P_{(i,0),(j,0)} & \cdots & P_{(i,0),(j,15)} \\ \vdots & \ddots & \vdots \\ P_{(i,15),(j,0)} & \cdots & P_{(i,15),(j,15)} \end{bmatrix}.$$

To compute $P_{(i,s),(j,k)}$, we consider the behavior of the ertPS algorithm in detail. When the grant size is greater than or equal to the packet size at time $t-1$, which corresponds to the first and second transition types in (1), we can derive $P_{(i,s),(j,k)}$ as follows:

$$\begin{aligned} P_{(i,s),(j,k)} &\triangleq P[g(t)=j, s(t)=k | g(t-1)=i, s(t-1)=s] \\ &= P[g(t)=j | g(t-1)=i, s(t-1)=s] \\ &\quad \cdot P[s(t)=k | g(t-1)=i, s(t-1)=s] \\ &= \begin{cases} 1\cdot a_{sk}, & \text{if } i \geqslant j=\dfrac{1}{2^{(s \bmod 4)}}, \\ 0, & \text{otherwise,} \end{cases} \end{aligned}$$

where $a_{sk}=P[s(t)=k|s(t-1)=s]$, $s,k\in\mathcal{S}\triangleq\{0,1,\cdots,15\}$, as defined in Sect. 2.2.

On the other hand, when the grant size is smaller than the packet size at time $t-1$, which corresponds to the third and fourth transition types in (1), we can derive $P_{(i,s),(j,k)}$ as follows:

$$\begin{aligned} P_{(i,s),(j,k)} &\triangleq P[g(t)=j, s(t)=k | g(t-1)=i, s(t-1)=s] \\ &= P[s(t)=k | g(t)=j, g(t-1)=i, s(t-1)=s] \\ &\quad \cdot P[g(t)=j | g(t-1)=i, s(t-1)=s] \\ &= \begin{cases} 1\cdot P[g(t)=i | g(t-1)=i, s(t-1)=s], \\ \text{if } i<\dfrac{1}{2^{(s \bmod 4)}}, j=i \text{ and } s=k, \\ a_{sk}\cdot P[g(t)=j | g(t-1)=i, s(t-1)=s], \\ \text{if } i<j=\dfrac{1}{2^{(s \bmod 4)}}, \\ 0, \quad \text{otherwise.} \end{cases} \end{aligned} \quad (2)$$

Here, $P[g(t)=i|g(t-1)=i, s(t-1)=s], i<\frac{1}{2^{(s \bmod 4)}}$, is the probability that a sufficient grant size is not assigned by the BS until time t. On the other hand, $P[g(t)=j|g(t-1)=i, s(t-1)=s], i<j=\frac{1}{2^{(s \bmod 4)}}$, is the probability that a sufficient grant is assigned by the BS until time t. These probabilities are not easy to compute because they are related to the reserved amount of resources (the sum of all grant sizes) for the other SSs in the network. To approximate these probabilities, we consider the following.

Let TG denote a random variable representing the sum of all grant sizes at an arbitrary frame slot and R_{tot} be the total amount of resources that can be used for VoIP services at a frame slot in the network. Since there are three more frame slots (of length 5 ms) between two time points, $t-1$ and t (of length 20 ms), it follows that

$$\begin{aligned} &P[g(t)=i|g(t-1)=i, s(t-1)=s,] \\ &\quad \approx P[TG > R_{tot}-r_s]^3 \text{ for } i<\frac{1}{2^{(s \bmod 4)}}, \end{aligned} \quad (3)$$

where r_s denotes the required resources to transmit a HOL packet with packet state $s\in\mathcal{S}$. Let $\widetilde{r}_1$, $\widetilde{r}_2$, $\widetilde{r}_3$, and $\widetilde{r}_4$ denote the required resources when data rates are $1, \frac{1}{2}, \frac{1}{4}$, and $\frac{1}{8}$, respectively. As mentioned in Sect. 2.2, for a packet state $s(t)=s$, the data rate $R(t)$ is given by $\frac{1}{2^{(s \bmod 4)}}(\in\{1,\frac{1}{2},\frac{1}{4},\frac{1}{8}\})$. Therefore, r_s is equal to $\widetilde{r}_i, 1\leqslant i\leqslant 4$, when $\frac{1}{2^{(s \bmod 4)}}$ is $\frac{1}{2^{i-1}}$.

To compute $P[TG > R_{tot}-r_s]$, we assume that total VoIP users in the network are equally distributed over 4 frame slots between two time points, $t-1$ and t, and there are N VoIP users in a frame slot. Moreover, we assume that the network is in steady state. If N_1, N_2, N_3, and N_4 are random variables which denote the numbers of users in a frame slot whose packet data rates are $1, \frac{1}{2}, \frac{1}{4}$, and $\frac{1}{8}$, respectively, then N_1, N_2, N_3, and N_4 have a multinomial distribution as follows:

$$\begin{aligned} &P[N_1=n_1, N_2=n_2, N_3=n_3, N_4=n_4] \\ &\quad = \binom{N}{n_1}\binom{N-n_1}{n_2}\binom{N-n_1-n_2}{n_3} p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{(N-n_1-n_2-n_3)}, \end{aligned}$$

where p_1, p_2, p_3, and p_4 are given in Table 2. Using this, we can approximate the probability in (3) as follows:

$$\begin{aligned} &P[TG > R_{tot}-r_s] \\ &\quad \approx \sum_{n_1}\sum_{n_2}\sum_{n_3}\sum_{n_4} 1\left\{\sum_{i=1}^{4}\widetilde{r}_i\cdot n_i > R_{tot}-r_s\right\} \\ &\qquad \times P[N_1=n_1, N_2=n_2, N_3=n_3, N_4=n_4], \end{aligned}$$

where $1\{A\}$ is the indicator function defined by 1 if A is true and by 0 otherwise.

Similarly, we can approximate the probability $P[g(t)=j|g(t-1)=i, s(t-1)=s], i<j=\frac{1}{2^{(s \bmod 4)}}$, in (2) that a sufficient grant can be assigned by the BS until time t as follows:

$$\begin{aligned} &P[g(t)=j|g(t-1)=i, s(t-1)=s] \\ &\quad \approx 1-P[TG > R_{tot}-r_s]^3 \text{ for } i<j=\frac{1}{2^{(s \bmod 4)}}. \end{aligned}$$

Let $\pi_{(i,s)}=\lim_{t\to\infty}P\{g(t)=i, s(t)=s\}$ be the stationary

probabilities of the two-dimensional Markov Chain. Define a row vector $\boldsymbol{\pi}$ by

$$\boldsymbol{\pi} = [\pi_{(1,0)}, \cdots, \pi_{(1,15)}, \cdots, \pi_{(\frac{1}{8},0)}, \cdots, \pi_{(\frac{1}{8},15)}].$$

Then, the stationary distribution $\boldsymbol{\pi}$ can be computed from

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}, \quad \sum_{i\in\mathcal{X}, s\in\mathcal{S}} \pi_{(i,s)} = 1.$$

Let $L_i (i \in \{1, 2, 3, 4\})$ be the information bits per HOL packet when the data rate of the HOL packet is 1, $\frac{1}{2}$, $\frac{1}{4}$, and $\frac{1}{8}$, respectively, as given in Table 2, and T_{VC} be the frame generation duration 20 ms in the EVRC. As mentioned before, since we assume that total VoIP users in the network are equally distributed over 4 frame slots between two time points and there are N VoIP users in a frame slot, there are a total of $4N$ VoIP users in the network. Then, in the ertPS algorithm the total throughput of the information data can be calculated as follows:

$$S_N^e = \frac{4N}{T_{VC}} \sum_{i=1}^{4} L_i \cdot p_i^u,$$

where p_i^u denotes the probability that a HOL packet with data rate $\frac{1}{2^{i-1}}$ is transmitted at an arbitrary time, and is computed by

$$p_i^u = \sum_{j\geqslant j'} \pi_{(j,s)} + \sum_{j<j'} \pi_{(j,s)} \cdot (1 - P[TG > R_{tot} - r_s]^3),$$

where $j' = \frac{1}{2^{s \mod 4}}$ and $i = (s \mod 4) + 1$.
Here, the first term $\sum_{j\geqslant j'} \pi_{(j,s)}$ is the probability that the grant size is not less than the HOL packet size. The second term is the probability that the grant size is less than the HOL packet size, but the HOL packet is transmitted before the next granted frame slot.

3. Proposed Algorithms for VoIP Services in IEEE 802.16e Systems

In the ertPS algorithm, an SS can not use resources at frame slots which are not granted frame slots (except for the case that the SS sends the BRH to the BS) even though the SS has packets to transmit in its queue and network resources remain at those frame slots. Since VoIP services are delay sensitive and require a stringent end-to-end delay bound, if voice packets suffer from excessive queueing delay even for a short period of time, then the voice quality degrades. Therefore, it is important to serve VoIP packets as soon as possible so that they are not suffered from queueing delay. In addition, it would be efficient in terms of the resource utilization of the system that remaining resources are assigned to SSs who need to but are not allowed to transmit packets. By additionally assigning the remaining resources to SSs whenever possible, SSs can reduce the packet queueing delay and, accordingly, the maximum number of SSs with VoIP services that satisfy the given maximum delay bound can be increased. With these motivations, we propose the enhanced ertPS (e^{2}rtPS) algorithm by modifying the ertPS algorithm in this section.

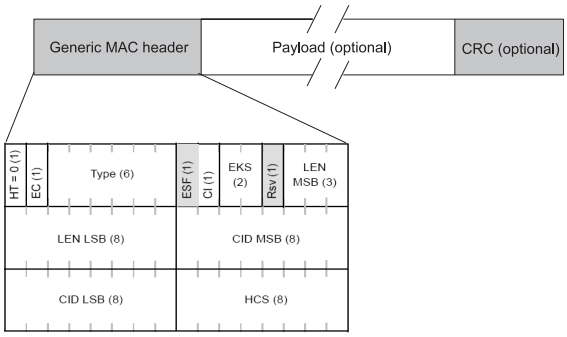


Fig. 3 The Generic MAC Header (GMH) format.

3.1 The e^{2}rtPS Algorithm

Our e^{2}rtPS algorithm considers the queue status of each SS when the BS assigns resources. In the e^{2}rtPS algorithm, an SS informs the BS of its queue status information on whether the queue length of the SS is not less than an *a priori* given threshold K or not. If the queue length of the SS is greater than or equal to K, then the BS tries to assign additional resources to the SS.

To inform the BS of the SS's queue status we use the reserved bit in the Generic MAC Header (GMH) of IEEE 802.16e. The GMH format of the IEEE 802.16e standard is shown in Fig. 3. As shown in Fig. 3, each MAC PDU (Protocol Data Unit) begins with a GMH of length 6 bytes, and in the GMH there is one reserved bit for additional future operations. In the e^{2}rtPS algorithm the reserved bit is called the Queue status Indicator (QI) bit. When the SS transmits a voice packet with assigned resources at each granted frame slot, the SS checks whether its queue length is greater than or equal to K or not, i.e., whether queue congestion occurs or not. If the SS detects that its queue length is greater than or equal to K, the QI bit is set to '1.' Otherwise, the QI bit is set to '0.' By doing this, without any additional overhead and amending the overall framework of IEEE 802.16e, the SS can inform the BS of its queue status and can continuously request additional resources until the queue congestion is alleviated. When the BS assigns additional resources to the SS, the amount of additional resources is equal to the currant grant size of the SS. In the e^{2}rtPS algorithm, the SS can inform the BS of its queue status only when the SS transmits a voice packet to the BS because the QI bit in the GMH is used for this purpose. Therefore, we can consider two cases according to the current grant size and the size of the HOL packet in SS's queue. The detailed operations of the e^{2}rtPS algorithm are as follows.

- **Case 1**:
  When the current grant size is greater than or equal to the size of the HOL packet at the granted frame slot and the SS detects its queue congestion, the SS sets the QI

bit to '1' in the GMH and transmits the voice packet. Then, the BS becomes aware of the queue congestion of the SS and determines whether remaining resources at the next frame slot (which is not the granted frame slot of the SS) are available for the SS. This is performed by checking the reserved amount of grants that have already been allocated to other SSs at the frame slot of interest. If remaining resources are not less than the current grant size of the SS, the BS assigns additional resources to the SS. However, if remaining resources are not available, then the BS can not assign additional resources to the SS. Then, the BS continuously checks over the network resource availability at every frame slot until the next granted frame slots of the SS. If there are enough remaining resources at some frame slot, additional resources are allocated to the SS at the frame slot. On the other hand, if there are no remaining resources at any of the frame slots between two consecutive granted frame slots, then the BS assigns the reserved resources to the SS at the next granted frame slot in the same manner as in the ertPS algorithm.

Now assume that the SS is assigned additional resources at some frame slot. Note that the amount of additional resources is equal to the current grant size. If the size of a new HOL packet is not greater than the current grant size, the SS transmits the HOL packet at the frame slot. On the other hand, if the size of the HOL packet is greater than the current grant size, the SS can not transmit the HOL packet at the frame slot. In this case, the SS uses the assigned additional resources at the frame slot to transmit the BRH in order to request an increase in the grant size as in the ertPS algorithm. Then, the BS uses the newly requested grant size to check whether it can assign additional resources, which are equal to the newly requested grant size, to the SS as described above. So, the e^{2}rtPS algorithm makes the SS to send the BRH before the next grant frame slot of the SS, if possible, in this case.

In the e^{2}rtPS algorithm, the BS performs the above procedures at every possible frame slot whenever it receives a packet with the QI bit set to '1.' If more than one SS request additional resources simultaneously at a frame slot, then the BS selects some of them randomly up to an acceptable number of SSs according to the amount of remaining resources at the frame slot.

- **Case 2**:
  When the current grant size is smaller than the size of the HOL packet, then the SS can not inform the BS of its queue congestion because it can not transmit the HOL packet. In this case, the mechanism follows the same manner as the ertPS algorithm. Then, after the SS receives the requested resources, the SS can inform the BS of its queue congestion status by transmitting the HOL packet with a QI bit, which is 1, in the GMH, and the next processes are the same as in **Case 1**.

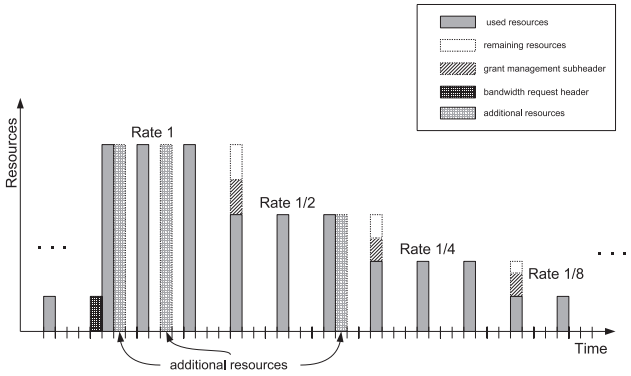


Fig. 4 The operation of the e^{2}rtPS algorithm.

In the e^{2}rtPS algorithm, if the SS no longer wants to receive additional resources because the queue length drops to below K, then the SS sets the QI bit to '0' in the GMH and transmits the HOL packet. Then, the BS stops assigning additional resources to the SS, and the ertPS algorithm is performed. The operation in the e^{2}rtPS algorithm is shown in Fig. 4.

4. Numerical Results

In this section, we provide simulation results to verify the accuracy of our analysis for the ertPS algorithm and the performance improvement of the e^{2}rtPS algorithm. First, we describe the simulation environment. We consider that voice packets are generated every 20 ms at each SS and the generated voice packet sizes are variable according to the rate type of the EVRC, as given in Table 2. The transition probabilities among the four rate types can be obtained from [3], as mentioned in Sect. 2. We assume that the amount of required resources, $\widetilde{r}_1$, $\widetilde{r}_2$, $\widetilde{r}_3$, and $\widetilde{r}_4$ to transmit the voice packets of four rate types are 6, 4, 3, and 2 resource units, respectively. In fact, $\widetilde{r}_i$ depends the value L_i, $1 \leqslant i \leqslant 4$, respectively, and the detailed derivation of the values $\widetilde{r}_i$ from L_i, $1 \leqslant i \leqslant 4$, is given in [5], [6]. Moreover, we assume the total resources R_{tot} are 140 units as in [5].

Figure 5 shows the total throughput in the ertPS algorithm. The results obtained from our approximation formula are denoted by 'ertPS$_{\text{ana}}$,' and the results estimated by simulations for the ertPS algorithm are denoted by 'ertPS$_{\text{sim}}$.' As shown in Fig. 5, the results from analysis are matched well with the results from simulations. So, we confirm that our analysis of the ertPS algorithm is appropriate.

Next, we show the effectiveness of the e^{2}rtPS algorithm through simulation. In simulation, we first use the threshold value $K = 1$, and the results are plotted in Fig. 6. As shown in Fig. 6, we can verify that the total throughput of the e^{2}rtPS algorithm is improved compared with that of the ertPS algorithm. In particular, when the total number of SSs is more than 100, the total throughput of the e^{2}rtPS algorithm is significantly greater than that of the ertPS algorithm. This result implies that the e^{2}rtPS algorithm works effectively when the number of SSs is large. In fact, we observe that the to-

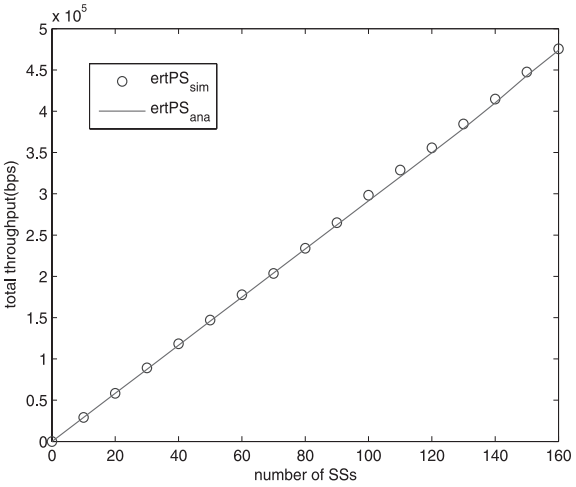


Fig. 5 The total throughput of the ertPS algorithm (bps).

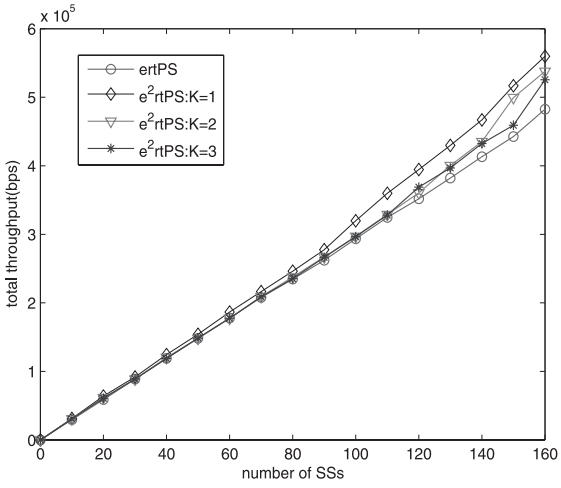


Fig. 7 The total throughput: The effect of the threshold K.

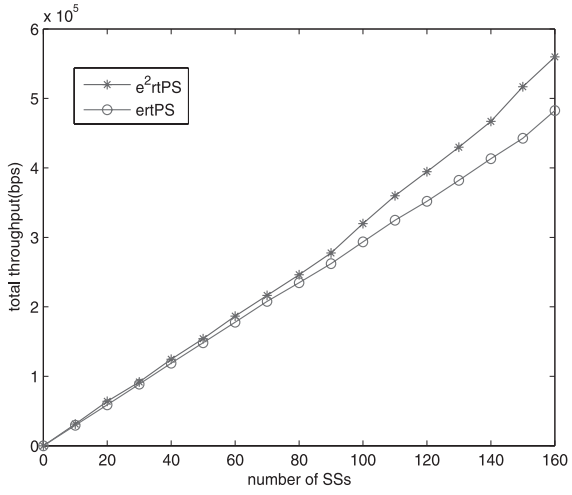


Fig. 6 The total throughput (bps): ertPS vs. e^{2}rtPS.

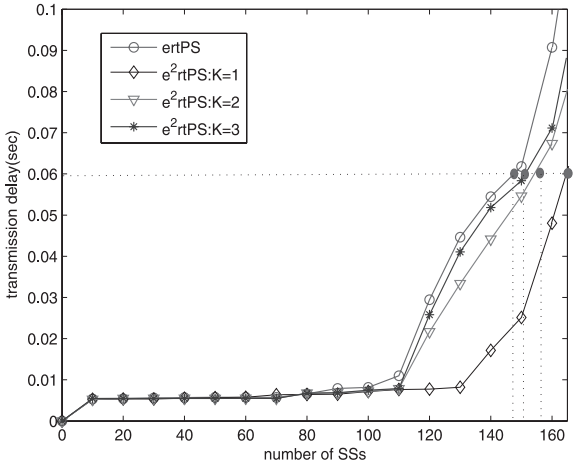


Fig. 8 The transmission delay for an SS.

tal throughput increases up to 15% in the e^{2}rtPS algorithm compared to that in the ertPS algorithm when the number of SSs is 150.

Next, we examine the effect of threshold value K in the e^{2}rtPS algorithm. In several studies [5], [6], [9], the maximum delay bound is generally assumed to be 60 ms. Note that even if there are only 3 packets in the queue, some of them are likely to be delayed over 60 ms (and consequently may be dropped) because packet transmission intervals for an SS are 20 ms in length. Therefore, in simulations, we change the threshold value K from 1 to 3 and plot the simulation results for the total throughput, as shown in Fig. 7. We observe that the total throughput in the e^{2}rtPS algorithm increases as the threshold value K decreases. As mentioned in Fig. 6, the e^{2}rtPS algorithm with threshold value $K = 1$ works effectively when the total number of SSs is more than 100. On the other hand, we observe that the e^{2}rtPS algorithm with threshold value $K = 2$ or $K = 3$ works effectively when the total number of of SSs is more than 120.

Finally, the simulation results for transmission delay of a voice packet in an SS are plotted in Fig. 8. In Fig. 8, transmission delay of a voice packet is the sum of the queueing delay and physical frame duration time 5 ms for transmitting the voice packet. As shown in Fig. 8, when the total number of SSs is more than 110, transmission delay in the e^{2}rtPS algorithm with $K = 1$ or $K = 2$ is remarkably less than that in the ertPS algorithm. When $K = 3$, transmission delay in the e^{2}rtPS algorithm is a little bit less than that in the ertPS algorithm until the number of SSs reaches around 150. However, when the number of SSs is greater than 150, the difference of transmission delays in both algorithms becomes significant. Therefore, we conclude that the e^{2}rtPS algorithm can greatly alleviate the queue congestion. Figure 8 also shows that the maximum number of SSs that satisfy a given maximum delay bound is increased by using the e^{2}rtPS algorithm. For instance, assume that the transmission delay requirement is given to be 60 ms. Then, from Fig. 8, we see that the maximum numbers of SSs that satisfy the given delay bound in the ertPS algorithm and the e^{2}rtPS algorithms with $K = 1$, $K = 2$, and $K = 3$ are 147, 167, 155, and 152, respectively. So, the e^{2}rtPS algorithm with

$K = 1$ has an increase of 14% in the maximum number of SSs compared with the ertPS algorithm, which also shows the usefulness of the e^{2}rtPS algorithm.

5. Conclusion

In this paper, we analyze the ertPS algorithm by using a two-dimensional Markov Chain, where the grant size and the voice packet state are considered. From the analysis, we obtain an approximation formula for the total throughput in the ertPS algorithm. To improve the performance of the ertPS algorithm, we propose an enhanced ertPS algorithm, called the e^{2}rtPS algorithm, for VoIP services in IEEE 802.16e systems. Through numerical studies, we show that our approximation results for the ertPS algorithm are matched well with the simulation results. In addition, we verify that the e^{2}rtPS algorithm can improve the total throughput and decrease transmission delay.

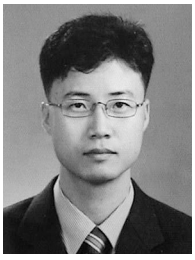
Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions, which significantly improved the presentation of this paper.

References

[1] IEEE 802.16e/D12-2005, "IEEE Standard for local and metropolitan area networks — Part 16: Air interface for fixed and mobile broadband wireless access systems — amendment for physical and medium access control layers for combined fixed and mobile operation in licensed bands and corrigendum 1," Oct. 2005.
[2] TIA/EIA/IS-127, "Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital systems," 1996.
[3] TIA/EIA/IS-871, "Markov service option (MSO) for cdma 2000 spread spectrum systems," April 2001.
[4] H. Lee, T. 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.9, no.8, pp.691–693, 2005.
[5] H. Lee, T. Kwon, and D.-H. Cho, "Extended-rtPS algorithm for VoIP services in IEEE 802.16 systems," Proc. IEEE International Conference on Communications, vol.5, pp.2060–2065, 2006.
[6] H. Lee, T. Kwon, D.-H. Cho, G. Lim, and Y. Chang, "Performance analysis of scheduling algorithms for VoIP services in IEEE 802.16e systems," Proc. 63rd IEEE Vehicular Technology Conference, vol.3, pp.1231–1235, 2006.
[7] G. Yangfeng and H. Aiqun, "Bandwidth allocation algorithm of VoIP based on the adaptive linear prediction in the IEEE 802.16 system," Proc. 6th International Conference on ITS Telecommunications, pp.16–19, 2006.
[8] H. Zhang, Y. Li, S. Feng, and W. Wu, "A new extended rtPS scheduling mechanism based on multi-polling for VoIP service in IEEE 802.16e system," Proc. IEEE International Conference on Communication Technology, pp.1–4, Nov. 2006.
[9] D. Zhao and X.(S.) Shen, "Performance of packet voice transmission using IEEE 802.16 protocol," IEEE Wireless Commun., vol.14, no.1, pp.44–51, Feb. 2007.

Bong Joo Kim received her B.Sc. degree in Mathematics from Ajou University, Suwon, Republic of Korea in 2002. In 2004, she received her M.Sc. degree in Applied Mathematics from KAIST, Daejeon, Republic of Korea. She is working toward the Ph.D. degree in the Department of Mathematical Sciences (formerly, Division of Applied Mathematics) at KAIST. Her research interests include teletraffic theory, performance evaluation of communication systems, quality of service provisioing for wired/wireless networks.



Gang Uk Hwang received his B.Sc., M.Sc., and Ph.D. degrees in Mathematics (Applied Probability) from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, in 1991, 1993 and 1997, respectively. From February 1997 to March 2000, he was with Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea. From March 2000 to February 2002, he was a visiting scholar at the School of Interdisciplinary Computing and Engineering in University of Missouri-Kansas City. Since March 2002, he has been with the Department of Mathematical Sciences and Telecommunication Engineering Program at KAIST, where he is currently an Associate Professor. His research interests include teletraffic theory, performance evaluation of communication systems, quality of service provisioing for wired/wireless networks and cross-layer design for wireless networks.