Wireless Packet Scheduling Algorithm for OFDMA System Based on Time-Utility and Channel State

Seungwan Ryu, Byung-Han Ryu, Hyunhwa Seo, Muyong Shin, and SeiKwon Park

In this paper, we propose an urgency- and efficiency-based wireless packet scheduling (UEPS) algorithm that is able to schedule real-time (RT) and non-real-time (NRT) traffics at the same time while supporting multiple users simultaneously at any given scheduling time instant. The UEPS algorithm is designed to support wireless downlink packet scheduling in an orthogonal frequency division multiple access (OFDMA) system, which is a strong candidate as a wireless access method for the next generation of wireless communications. The UEPS algorithm uses the time-utility function as a scheduling urgency factor and the relative status of the current channel to the average channel status as an efficiency indicator of radio resource usage. The design goal of the UEPS algorithm is to maximize throughput of NRT traffics while satisfying quality-of-service (QoS) requirements of RT traffics. The simulation study shows that the UEPS algorithm is able to give better throughput performance than existing wireless packet scheduling algorithms such as proportional fair (PF) and modified-largest weighted delay first (M-LWDF), while satisfying the QoS requirements of RT traffics such as average delay and packet loss rate under various traffic loads.

Keywords: OFDMA, wireless downlink, packet scheduling, time-utility, channel state.

Manuscript received Jan. 30, 2005; revised Aug. 18, 2005.

The material in this work was presented in part at CIC 2004, Seoul, Korea, Oct. 2004.

Seungwan Ryu (phone: +82 42 860 1505, email: rush@etri.re.kr), Byung-Han Ryu (email: rubh@etri.re.kr), Hyunhwa Seo (email: hhseo@etri.re.kr), and Muyong Shin (email: myshin@etri.re.kr) are with Mobile Telecommunication Research Division, ETRI, Daejeon, Korea.

SeiKwon Park (email: psk3193@cau.ac.kr) is with the Department of Information Systems, Chung-Ang University, Anseong, Korea.

I. Introduction

Wireless networks are evolving in two aspects: 1) adoption of an Internet protocol (IP) as a transport mechanism, and 2) evolution of the wireless access methods toward beyond third generation (B3G) or next generation methods [1]. As a prospect of the former evolution, the next generation wireless networks are expected to be packet-based wireless networks where traffic data, including real-time (RT) and non-real-time (NRT) traffic data, is transmitted over a wireless link through packetization. The latter evolution is expected to bring a wireless system that is able to support a significantly higher system capacity, a higher data rate per user, and a provisioning of quality-of-service (QoS) to users. As a result of those two evolutional aspects, the next generation wireless networks have been envisaged as packet-based wireless networks capable to provide a variety of video, audio, and data services with a number of traffic classes in addition to conventional mobile services such as voice telephony. It is anticipated that packetized transmission over wireless links between a base station and user equipment (UE) will make possible higher radio resource utilization through statistical multiplexing of packets compared to conventional circuit-based communications. Since the packets of all traffic classes have their unique characteristics and quality of service requirements, packet flows in the next generation wireless networks can be categorized into several traffic classes. In addition, increase of the downlink traffic is anticipated to be a major characteristic of the next generation wireless communications [2].

In this paper, we explore how to provide QoS in packet-based next generation wireless networks. Challenges on delivering QoS to users in such packet-based wireless networks have been watched with keen interest, and the packet scheduler operating at the medium access control (MAC) layer is

considered as the key component for QoS provisioning to users. The packet scheduler should be able to guarantee QoS requirements of such traffic classes as well as maximize utilization of the limited radio resources. In this study, we concentrated our work on delivering QoS to users at the downlink between a base station and UE.

In general, traffic scheduling in packet-based networks is a mechanism responsible for determining the transmission order of packets from different competing flows. At a given scheduling instant (that is, a timeslot), the traffic scheduler tries to maximize the system performance in terms of different QoS requirements such as delay, loss rate, throughput, and utilization of limited radio resources using the status of the server and packets waiting at buffers. However, the packet scheduling methods used in a wired realm cannot be applied directly to a wireless environment because of its inherent characteristics and restrictions including location dependent error, time-varying channel capacity stemming from unstable wireless channel status, and bursty error caused by these factors.

There are many packet scheduling algorithms such as proportional fair (PF) [3] and modified-largest weighted delay first (M-LWDF) [4] designed for the CDMA-1x-EVDO (HDR) system. However, these existing packet schedulers are mainly designed to support non-real-time (NRT) data services. In particular, deadline-based general wireless packet scheduling algorithms have been proposed recently [5], [6]. In this approach, packets are discriminated by their deadline and scheduled based on scheduling priority represented as a function of deadline. However, we are not able to distinguish real-time (RT) and NRT traffics in this approach, and thus QoS requirements of packets of various traffic types cannot be met satisfactorily. In contrast, we propose a wireless packet scheduling algorithm that can support RT and NRT data traffics at the same time. The proposed algorithm is designed to schedule packets of different traffic classes at the downlink in the orthogonal frequency division multiple access (OFDMA) system, which is a strong candidate wireless access method for the next generation wireless networks. As a result, the proposed packet scheduling algorithm can also support multiple users simultaneously at any given scheduling time instant using OFDMA, a multi-user-OFDM wireless access method.

In order to schedule packets of RT and NRT traffics at the same time, the proposed algorithm uses two scheduling factors, the *urgency* of scheduling and the *efficiency* of radio resource usage, by taking into account not only the inherent characteristics and restrictions of a wireless environment but also the QoS requirements of each traffic. In the proposed urgency- and efficiency-based packet scheduling (UEPS) algorithm, the time-utility function is used to represent the urgency of scheduling while the channel state is used to indicate efficiency of usage of the radio resource. The idea behind the UEPS algorithm is to maximize the throughput of NRT traffics as long as the QoS requirements of RT traffics such as the packet delay and the loss rate requirements are satisfied. Then, the UEPS scheduler transmits packets of RT and NRT traffic based on their scheduling priorities obtained from the urgency and efficiency factors.

This paper is organized as follows. In the next section, we briefly survey existing wireless packet scheduling algorithms. In section III, we introduce the OFDMA wireless system model used in this study and the structure of the packet scheduler. Time related scheduling approaches including deadlines and time-value functions are discussed in section IV. In section V, we propose the UEPS algorithm that is able not only to schedule RT and NRT traffic packets but also to support multiple users at the same time. In section VI, we evaluate the performance of the UEPS algorithm and compare it with existing algorithms via simulation study. Finally, a summary of this study is given followed by further study issues.

II. Related Works

Many wireless packet scheduling algorithms have been designed to support data traffics in the Third Generation Partnership Project (3GPP) and 3GPP2 wireless systems. For the 3GPP2 system, PF [3] and M-LWDF [4] algorithms are designed mainly to support NRT data services in a CDMA-1x-EVDO system. In the 3GPP wideband-CDMA (WCDMA) system, only NRT data traffic classes such as streaming, interactive, and background traffic classes are subjects of scheduling, and are transmitted through a common or shared channel [7]. On the other hand, conversational traffics such as voice telephony and voice over IP (VoIP) traffics are transmitted on a dedicated channel without scheduling. Moreover, existing packet scheduling algorithms designed to support NRT data traffics in either 3GPP or 3GPP2 are not able to support multiple users at any given scheduling time instant. Furthermore, the traditional voice-based MAC protocol does not perform well in the next generation packet-based multimedia traffic environment because of the bursty nature of such traffic [8].

In this section, we introduce two existing representative wireless packet scheduling algorithms, PF [3] and M-LWDF [4]. The design objective of the PF algorithm is to maximize long-term throughput of a UE whose current channel status (that is, achievable data rate) is better compared to the average throughput. Suppose that $R_i(t)$ and $T_i(t)$ are the current achievable data rate and the estimate of average throughput of user i at timeslot t, where $i \in I = \{1, \cdots, M\}$. Then the PF algorithm works as follows:

- Scheduling: The user with the highest ratio of $R_i(t)/T_i(t)$

among all users will receive a transmission from a base station at each scheduling time. Ties are broken randomly.

• Update the average throughput of each user *i*:

$$T_i(t+1) = (1-1/t_c)T_i(t) + (1/t_c)R_i(t)\Delta_i, \quad (1)$$

where $\Delta_i = 1$ if user i is chosen to transmit; otherwise, $\Delta_i = 0$ and t_c is a low pass filtering parameter.[1)]

Since the PF algorithm is designed to support only data services in the CDMA-1x-EVDO system, it cannot support RT services such as voice and RT-video streaming services.

The M-LWDF algorithm was proposed to support not only NRT data services but also almost real-time services such as video streaming services in the CDMA-1x-EVDO system [4]. The design objective is to maintain the delay of all traffic smaller than a predefined threshold value with probability. The delay and throughput requirements are $\Pr\{W_i > \tau_i\} \le \delta_i$ and $T_i > t_i$, respectively, where W_i is the head-of-line (HOL) delay, W_i is the maximum allowable delay threshold, δ_i is a maximum allowable probability of exceeding τ_i, and t_i is a predefined minimum throughput threshold. In each timeslot t, a user i^* is selected according to the scheduling priority as follows:

$$i^* = \arg\max_{i \in I=\{1,\cdots,M\}} \gamma_i W_i(t) R_i(t), \quad (2)$$

where $\gamma_i = a_i / \overline{R_i}(t)$ is an arbitrary constant, $a_i = -(\log \delta_i)/\tau_i$, and $\overline{R_i}(t)$ is the average channel rate with respect to flow *i*. By setting an appropriate value to each parameter γ_i, the delay requirement can be satisfied. However, it is difficult to find the optimal γ_i value for each traffic class $i \in I$.

III. System Model

1. An Overview of the OFDMA Wireless System

OFDMA, also referred to as multi-user-OFDM, is an extension of orthogonal frequency division multiplexing (OFDM). In current OFDM systems, only a single user can transmit on all of the sub-carriers at any given time, and time division or frequency division multiple access is employed to support multiple users. The major drawback of this static multiple access scheme is the fact that different users see the wireless channel differently, and only one user is selected for packet transmission. In this case, when the channel capacity is larger than the amount of data to be transmitted, the radio resource is not fully utilized.

In this study, we consider an OFDMA system with 20 MHz of bandwidth. It is assumed that there are 1,536 sub-carriers, and all sub-carriers are shared by all users in a cell in terms of sub-channels, a subset of the sub-carriers. We assume that there are twelve sub-channels and each sub-channel is a group of 128 sub-carriers. It is also assumed that all sub-carriers are used for data transmission for simplification, and sub-carriers in each sub-channel are selected by a pre-determined random pattern. The modulation and coding scheme is determined by the prescribed adaptive modulation code table based on the instantaneous signal-interference-ratio (SIR) of each sub-channel. A summary of system parameters is shown in Table 1.

Table 1. A summary of system parameters.

Parameters	Value
System	OFDMA
Downlink channel bandwidth	20 MHz
OFDM symbol duration	100 μs
Total number of sub-carriers	1,536
Number of sub-carriers per sub-channel	128
Number of sub-channels	12
Frame period	12 ms
Slot period	1 ms

2. Structure of the Packet Scheduler in a Base Station

The packet scheduler operating at the MAC layer is the key component for delivering QoS to users. The proposed packet scheduling system in a base station consists of three blocks: a packet classifier, a buffer management block (BMB), and a packet

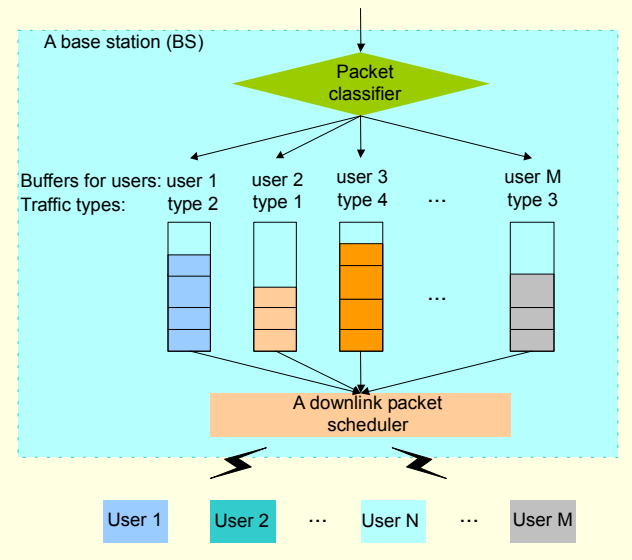

Fig. 1. Structure of the proposed packet scheduler.

1) The value of parameter t_c used in the PF scheduling algorithm is related to the maximum amount of time (i.e., the number of timeslots) for which an individual user can be starved, and t_c = 1000 is recommended [3].

scheduler as shown in Fig. 1. The packet classifier classifies incoming packets according to their types and QoS profiles, and sends them to buffers in the BMB. The BMB maintains QoS statistics such as the arrival time and delay deadline of each packet, the number of packets, and the HOL delay in each buffer. Finally, the packet scheduler transmits packets to users according to the scheduling priority obtained using the channel status reported by user equipment and QoS statistics maintained in the BMB.

IV. Time Related Scheduling Approaches

1. Value-Based Scheduling Approaches

Scheduling of tasks with time constraints has been studied extensively in the area of computer operating systems to meet certain performance criteria such as maximization of utilization of limited resources such as processors. Scheduling of such tasks can be modeled in various ways: One of the widely used methods is to schedule tasks based on the underlying task model. Two important characteristics of tasks associated with time constraints are the task execution time and the deadline, and these two characteristics can be represented by a value function in time [9]. The task execution time is the amount of time needed to finish the task, and the deadline is the time instant when the task should be finished . The task execution time is a task-local time constraint because it is defined on an individual task-oriented time domain. On the other hand, the deadline is a global time constraint since it is defined on the global time domain by the system (the scheduler in this paper) independent of the individual task time-line.

There are many scheduling approaches on tasks with time constraints. In this paper, we concentrate our work on scheduling approaches developed based on the above two fundamental properties of the underlying tasks. Figures 2 and 3 show two different expressions of value functions associated with two different properties of tasks. Figure 2 describes forms of value functions associated with the deadline expressed functionally in the global time domain. Figure 3 describes forms of value functions associated with the tasks' execution time expressed functionally in the task-local time domain. Time-value functions associated with the global time-line and individual task-oriented local time-line are classified as time-utility functions (TUF) and time-quality functions, respectively [9]. Since the packet scheduler operates with global time, the scheduling time at a MAC layer, time-utility functions are used to schedule RT and NRT traffics in this article.

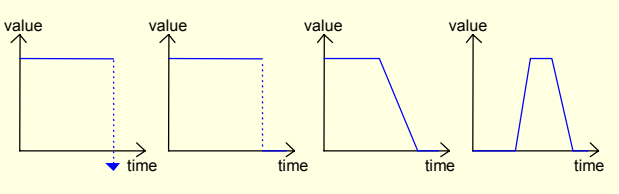

Fig. 2. Forms of value functions associated with the deadline expressed functionally in global time domain.

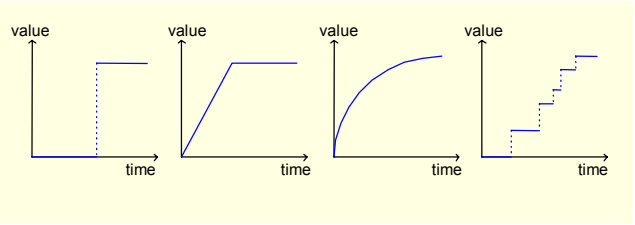

Fig. 3. Forms of value functions associated with the tasks' execution times expressed functionally in task-local time domain.

2. Deadlines and Time-Utility Functions

The major characteristic of time related tasks is the timing requirement expressed in the form of a deadline. A *hard deadline* is a time constraint that is expressed functionally by a binary unit-valued downward step at the deadline [10]. A time related task with a hard deadline is often modeled as a *hard time-utility function*, which has a positive value of 1 if the deadline is met, or 0 if the deadline is missed [10]. On the other hand, the *soft deadline* is a time constraint that yields diminishing values to the system when the deadline has passed. In other words, the soft deadline gives more or less timeliness depending on the completion time of a task with respect to its deadline. A task with a soft deadline is modeled as a *soft time-utility function* with respect to its completion time. Figure 4 shows concepts of the hard and the soft deadlines and the related hard and the soft time-utility functions.

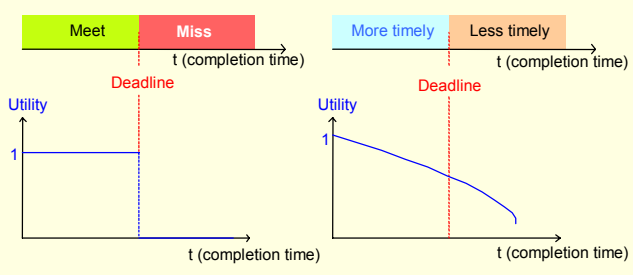

Fig. 4. Concepts of hard (left) and soft (right) deadlines and related time-utility functions.

3. Timeliness and Simultaneousness

If a real-time task should be completed within a deadline, this behavior can be described in terms of *timeliness*. A real-time task having timeliness can be expressed functionally with a hard time-utility function. However, the timeliness timing constraint often means that tasks should be completed by a certain given time instant, that is, a deadline. In this case, the task can be processed

during a certain predefined time interval that contains the deadline. This time interval can be a small time window consisting of delay jitters around the deadline as shown in Fig. 5. This behavior can be described in terms of *simultaneousness* [9]. In this paper, we use the term *simultaneousness* to distinguish the scheduling event occurring during a predefined time interval around a deadline from scheduling events occurring anytime within it. Then, we use a negative delay jitter from the deadline to define such predetermined time interval.

According to Jensen's TUF model [10], the former case implies that utility of the completion of a real-time task is 1 until the deadline. On the other hand, the latter case implies that the utility is 1 only in a small time window. This time window can be created by giving jitters around the deadline as shown in Fig. 5.

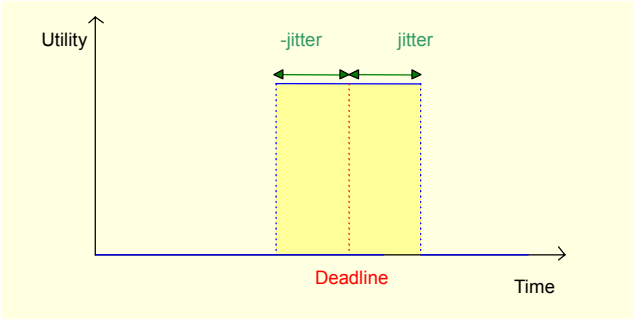

Fig. 5. The concept of simultaneousness with time window around the deadline.

V. The Urgency of Scheduling and Efficiency of Radio Resource Usage

Two scheduling factors, the *urgency* of scheduling and the *efficiency* of radio resource usage, are used to determine the scheduling priority of each user. The TUF is used to represent the urgency of scheduling, while the channel state is used to indicate the efficiency of radio resource usage.

1. The Urgency of Scheduling

Since the utility is decreasing in delay, that is, the longer the delay, the lower the utility, the urgency of scheduling can be expressed as a function in delay [11]. In this work, *the time utility function* is used to indicate the urgency of scheduling. A TUF of a delay-sensitive RT traffic can be expressed as a hard time-utility as shown in Fig. 4. On the other hand, the TUF of an NRT traffic is a continuously decreasing function in delay, in that the utility of an NRT traffic decreases slowly as delay increases. Among NRT traffics, some have a (soft) deadline like WWW traffics as described in the right-hand side of Fig. 4. On the other hand, some NRT traffics such as email and FTP traffics have much longer deadline or no deadline. Then, the unit change of the TUF value indicates the urgency of the scheduling of packets as time passes by.

One possible scheduling rule is to schedule head-of-line (HOL) packets waiting in the BMB based on the unit change of the TUF value at a scheduling instant. Let $U_i(t)$ be the TUF of a HOL packet of traffic i at time t. Then, the unit change of the TUF value of traffic i's HOL packet at time t is the absolute value of the first derivative of $U_i(t)$, that is, $|U_i'(t)|$, at time t. A possible packet scheduling rule is to select a traffic based on $|U_i'(t)|$, $\forall i \in I$.

A. Real-Time Traffics

Since the downlink between a base station and user equipments (UEs) is the last link to users, the end-to-end delay can be met as long as packets are delivered to users within the deadline. In other words, the scheduler transmits RT traffic packets any time within the deadline to satisfy the delay requirement. Hence, from this property of the downlink the tight timing constraint of a hard deadline described by the timeliness can be relaxed into a lax one described by the simultaneousness as shown in Fig. 5. For example, by introducing a negative jitter from the deadline, the tight timeliness timing constraint can be relaxed into a lax simultaneousness one. Then a packet of RT traffic i is transmitted only during specific time interval $[d_i - j_i, d_i]$, where d_i is the deadline of the HOL packet of the RT traffic i, and j_i is the length of the given negative delay jitter. In this paper, we call the specific time interval $[d_i - j_i, d_i]$ the *marginal scheduling time interval* (MSTI). Then, during the remaining time interval, $[0, d_i - j_i)$, other packets, especially NRT packets, are transmitted.

Since the TUF of an RT traffic is a hard and discontinuous function in delay, the unit change of the utility, $|U_i'(t)|$, cannot be obtained directly at its delay deadline. In order to address this problem, the TUF of an RT traffic can be relaxed into a continuous z-shaped function that has properties similar to the original hard discontinuous function. An example of z-shaped function relaxation of the TUF of an RT traffic is shown in Fig. 6.

A z-shaped function relaxation of the TUF of an RT traffic can be easily achieved analytically using an s-shaped function having close relation with a z-shaped function. For example, a z-shaped function can be obtained using the s-shaped sigmoid function, $f_{Sigmoid}(t, a, c)=1/(1+e^{-a(t-c)})$, where a and c are parameters that determine the slope and location of the inflection point of the function. Then, the relaxed z-shaped function is $U_{RT}(t)=1-f_{Sigmoid}(t, a, c)=e^{-a(t-c)}/(1+e^{-a(t-c)})$, and the unit change of utility of an RT traffic at the inflection point (t = c) is $|U_{RT}'(t=c)|=a/4$. This value is assigned as the urgency factor of an RT traffic packet during MSTI. As a result, with the z-shaped relaxation of a hard TUF, the concept of simultaneousness can be achieved by giving $|U_i'(t)|$ a positive value only during MSTI and 0 in another time interval.

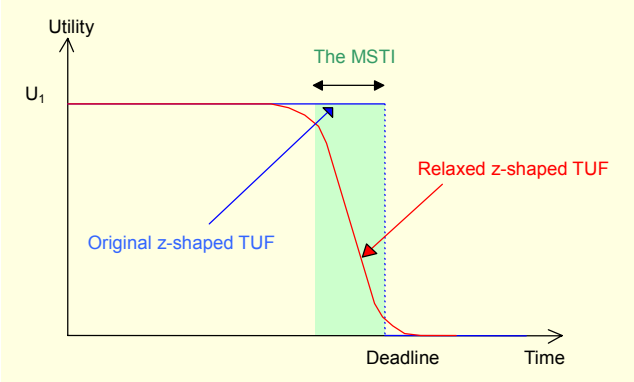

Fig. 6. An example of z-shaped function relaxation of a TUF of an RT traffic.

B. TUFs of NRT Traffics

Since TUFs of NRT traffics are monotonic decreasing functions in time (delay), an analytic model can be easily obtained using related monotonic increasing functions. For example, a truncated exponential function, $f(a_i, t, D_i)=\exp(a_i t)$, can be used, where a_i is an arbitrary parameter and $D_i \geq t \geq 0$ is the delay deadline of NRT traffic i. Then, a possible TUF of NRT traffic i is $f_{NRT_i}(t) = 1 - f(a_i, t, D_i) = 1 - \exp(a_i t)/\exp(D_i)$ [2], and the urgency is $|U'_{NRT_i}(t)| = a_i \exp(a_i t)/\exp(D_i)$.

C. Urgency Factors among RT and NRT Traffics

The urgency factor of each traffic type is used to determine the scheduling priority among HOL packets of different traffic types waiting in buffers. Assignment of the urgency factor among traffic types is dependent on the designer's preference. A rule of thumb is to give a higher scheduling priority to RT traffics over NRT traffics. In this paper, we set the urgency factors of all traffic types in the order of RT voice, RT video, and NRT traffics by setting the urgency factors as follows.

$$\begin{aligned} &|U'_{RT-Voice}(t)| > |U'_{RT-Video}(t)|, \\ &|U'_{NRT-Data1}(t)| > |U'_{NRT-Data2}(t)|. \end{aligned} \tag{3}$$

2. Efficiency of Radio Resource Usage

Efficiency in wireless communications is related to usage of the limited radio resources, that is, a limited number of radio channels or limited bandwidth. Thus, the channel state of available radio channels can be used as an efficiency indicator. For example, the current channel state ($R_i(t)$), the average channel state ($\overline{R}_i(t)$), or the ratio of the current channel state to the average ($R_i(t)/\overline{R}_i(t)$) can be used as an efficiency indicator. In this study, a moving average of the channel state of each user $i \in M$ in past W timeslots, $\overline{R}_i(t) = (1-1/W)\overline{R}_i(t) + (1/W)R_i(t)$, is used for the average channel state, where W is the time window (that is, the number of timeslots) used in calculation of the moving average of the channel state. Note that $R_i(t)$ used in our paper is different from the average throughput of user i, $T_i(t)$, in past t_c timeslots used in the PF algorithm [3]. Therefore, the higher the user's instantaneous channel quality relative to its average value, the higher the chance of a user to transmit data with a rate near to its peak value.

3. The Urgency and Efficiency Based Packet Scheduling (UEPS) Algorithm

We propose the *urgency and efficiency based packet scheduling* (UEPS) algorithm designed to support RT and NRT traffics for a user at the same time. The UEPS algorithm also tries to maximize the throughput of NRT data services while satisfying the maximum allowable QoS requirements of RT traffics such as the maximum allowable delay and loss rate. In detail, the UEPS scheduler transmits NRT traffics during time interval [0, $d_i - j_i$], assuming that the packet has arrived at time 0 and the channel states of all users are the same. In contrast, the scheduler gives an RT traffic a higher scheduling priority over NRT traffics during MSTI, [$d_i - j_i$, d_i].

The UEPS algorithm operates at a base station in three steps; step 0 for packet arrival events, step 1 for scheduling priority of each user, and step 2 for scheduling and transmission of packets.

- In **step 0**, the arrived packet is sent to a user's buffer by the packet classifier based on its user ID. QoS profiles of the arrived packet such as the arrival time, the deadline, the packet type, and the packet size are maintained in the BMB.
- In **step 1**, at each scheduling instant the urgency factor of HOL packets of each buffer, $|U'_i(t)|$, is calculated to represent the urgency factor of user i, that is, $|U'_i(t)|$. In addition, the efficiency factor of user i, $\overline{R}_i(t) = (1-1/W)\overline{R}_i(t) + (1/W)R_i(t)$, is obtained. Finally, the scheduling priority value of user i is $p_i(t) = |U'_i(t)| * R_i(t)/\overline{R}_i(t)$.
- In **step 2**, at each scheduling time instant, multiple users are selected based on their scheduling priority value obtained as follows.

$$i^* = \arg\max_{i \in M} |U'_i(t)| \frac{R_i(t)}{\overline{R}_i(t)}. \tag{4}$$

Then, a sub-channel is allocated to each selected user i^* based on the channel condition represented by the efficiency factor ($R_i(t)/\overline{R}_i(t)$) and the urgency of scheduling ($|U'_i(t)|$). In other words, the traffic class with the larger/smaller ratio of $|U'_i(t)| * R_i(t)/\overline{R}_i(t)$ will be scheduled first/last. The OFDMA

2) It is normalized by the maximum time, D_i, so that it can have a smoother slope.

system considered in this study is designed to support up to twelve users simultaneously at each scheduling time instant by allocating one of twelve sub-channels to each of the selected users. This is one of the main differences of the UEPS algorithm designed for the OFDMA system from the existing wireless packet scheduling algorithms such as PF [3] and M-LWDF [4] designed for the CDMA2000-EVDO system, where only one user is selected in each timeslot for data transmission. The capacity of each allocated sub-channel is determined from the adaptive modulation code option. Finally, the scheduler loads user i^*'s packets on the sub-channel as much as possible when there is room available.

VI. Performance Evaluation

1. Simulation Model and Traffic Parameters

A. Traffic Types

In the simulation study it is assumed that there are four different traffic types, and each user generates one of four traffics. The four types of services are the following:

- **RT voice:** RT voice is assumed to be the voice on IP (VoIP) that periodically generates packets of fixed size. Assuming that silence suppression is used, voice traffic is modeled by a 2-state Markov (ON/OFF) model. The length of the ON and OFF periods follows the exponential distribution with means of one second and 1.35 seconds, respectively.
- **RT video:** RT video is assumed to be the RT video steaming service that periodically generates packets of variable sizes. We use 3GPP streaming video traffic for this type of traffic (3GPP2/TSG-C.R1002). A traffic model and characteristics of an RT video traffic are shown in Fig. 7 and Table 2, respectively.
- **NRT data service type 1:** NRT data service type 1 is assumed to be the NRT data traffics such as web browsing that require wide bandwidth and variable sized bursty data. In this study, we use the web traffic (WWW) model proposed to have a session consisting of several web pages, which contains multiple packets or datagrams as shown in Fig. 8. Characteristics of the WWW traffic model are summarized in Table 3.
- **NRT data service type 2, best effort (BE):** BE is assumed to be NRT data traffics such as emailing traffic. We assume that a message's arrival to a mailbox is modeled by the Poisson process.

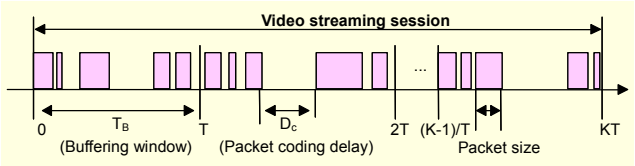

Fig. 7. A real-time video traffic model.

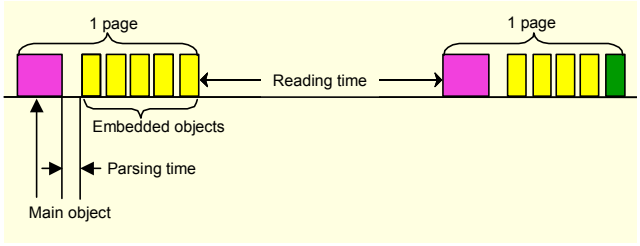

Fig. 8. A web (WWW) traffic model.

Table 2. A summary of the characteristics of a real-time video traffic model.

Characteristics	Distribution	Parameters
Inter-arrival time between frames	Deterministic	100 ms
Number of packets/frame	Deterministic	8
Packet size	Truncated Pareto (Mean:50, Max:125 (bytes))	K=20 bytes, α=1.2
Inter-arrival time between packets	Truncated Pareto (Mean:6, Max:12.5 (ms))	K=2.5ms, α=1.2

Table 3. A summary of the characteristics of a WWW traffic model.

Component	Distribution	Parameters
Main object size	Truncated normal	Mean: 10710 bytes, STD: 25032 bytes Min:100 bytes, Max: 2 Mbytes
Embedded object size	Truncated normal	Mean: 7758 bytes, STD: 12168 bytes Min:50 bytes, Max: 2 Mbytes
Number of embedded objects/page	Truncated Pareto	Mean: 5.64, STD: 53
Reading time	Exponential	Mean=30 s
Parsing time	Exponential	Mean=0.13 s

B. System Parameters

We consider a hexagonal cell structure consisting of a reference cell and six surrounding cells with a 1 km radius. We assume that all cells use an omni-directional antenna. Mobile stations are uniformly distributed in a cell and move with the velocity of uniform distribution in a random direction. The base station transmission power is 12 W evenly distributed to all twelve sub-channels. A summary of simulation parameters for the system model is shown in Table 4.

Table 4. A summary of simulation parameters for the system model.

Parameters	Value
System	OFDMA
Downlink channel bandwidth	20 MHz
OFDM symbol duration	100 μs
Total number of sub-carriers	1,536
Number of sub-carriers per sub-channel	128
Number of sub-channels	12
Frame period	12 ms
Slot period	1 ms

2. Performance Metrics

We evaluate and compare the performance of the proposed UEPS, PF, and M-LWDF algorithms in terms of three different performance metrics such as the packet loss rate, the average packet delay, and the average throughput via simulation study. For the delay-sensitive RT traffics, the average packet delay is mainly used to evaluate performance. Although the RT traffic is tolerant to packet loss, it has a maximum allowable packet loss rate. For example, the packet loss rate of RT voice should be less than 3% [13]. Therefore, the performance of RT traffics is also evaluated in terms of the packet loss rate. For RT traffics, performance of the UEPS algorithm is compared with that of the MLWDF algorithm in terms of the average delay and packet loss rate. The QoS performance requirements of RT voice and video traffics [13] are

- RT Voice: delay < 40 ms, loss rate < 3%, and
- RT Video: delay <150 ms, loss rate < 1%.

For the loss-sensitive NRT traffics, the average throughput is used to evaluate the performance of the UEPS, PF and M-LWDF algorithms. Parameters of the PF and M-LWDF algorithms are

- PF: t_c=1000, and
- M-LWDF: W_{Max}=40 ms, δ_{Voice} =0.03 for RT voice and W_{Max} =150 ms, δ_{Video} =0.01 for RT video.

3. Performance Evaluation

In order to evaluate the performance of the UEPS, PF, and MLWDF algorithms, various traffic loads are generated from a light to heavy traffic load. Since the proposed UEPS scheduler selects twelve users at each timeslot, the number of users arrived in each timeslot is used as the offered traffic load. In the simulation study, the offered traffic load (λ) distributes (2, 20) users/timeslot, that is, λ= 2/12 to 20/12 = 0.167 to 1.67.

Since the length of MSTI of RT traffics is one of the important design factors, performance of the UEPS algorithm has been evaluated extensively via simulation study under different sets of MSTI values. Different MSTI[3] values are obtained by giving different negative jitter values from the deadline. For example, if 10 ms of negative delay jitter is given for an RT traffic, the length of MSTI will be 10 ms prior to its deadline, that is, [deadline−10, deadline]. In this case, the RT traffic packets can be scheduled and transmitted to a user only during this time interval, and NRT traffics waiting in buffers are transmitted to maximize throughput during the remaining time interval. In this paper, because of page limitation we only present some important performance evaluation results with a limited set of MSTI values for RT voice and video traffics. We are presenting performance evaluation results when values of MSTI are set to 10, 20 and 30 ms for RT voice traffic and 30 ms for RT video traffic.

A. Performance of RT traffics

Figure 9 shows the performances of the UEPS and M-LWDF algorithms for RT voice traffic in terms of the packet loss rate under various traffic loads. The packet loss rate of the UEPS algorithm stays below the maximum allowable packet loss rate (3%) for all $MSTI_{Voice}$ values under all traffic loads. In detail, in terms of the packet loss rate of the RT voice traffic, performance of the UEPS algorithm is better than the M-LWDF algorithm when $MSTI_{Voice}$= 30 or 20, and is almost the same as that of M-LWDF when $MSTI_{Voice}$ = 10. In general, in case of RT voice traffic, the larger the MSTI value, the lower the packet loss rate with the UEPS algorithm.

In the case of RT video traffic, on the other hand, the packet loss rate of the UEPS algorithm is lower than that of the M-LWDF algorithm when $MSTI_{Voice}$ = 10 or 20, and is almost the same as

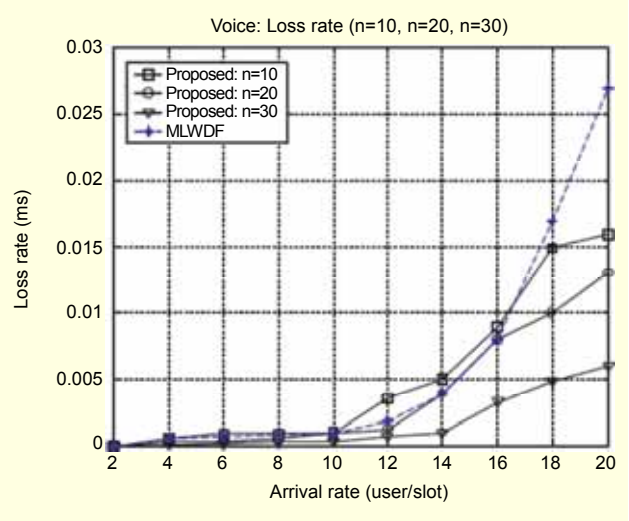

Fig. 9. Average packet loss rates of RT voice traffic under different traffic loads.

3) The MSTI value is represented as "n" in the following figures.

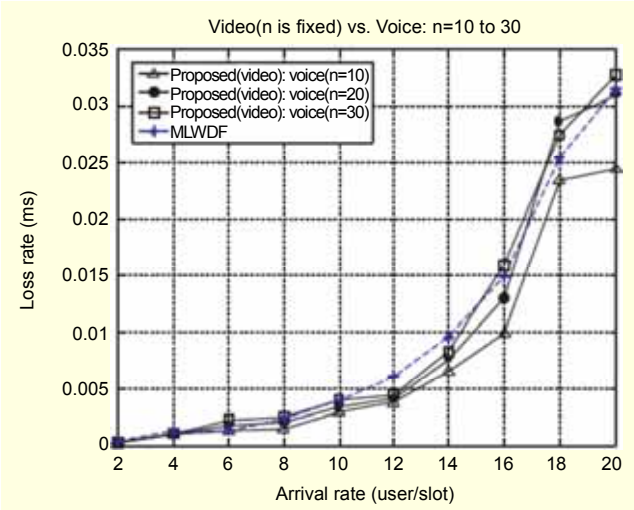

Fig. 10. Average loss rates of RT video traffic under different traffic loads.

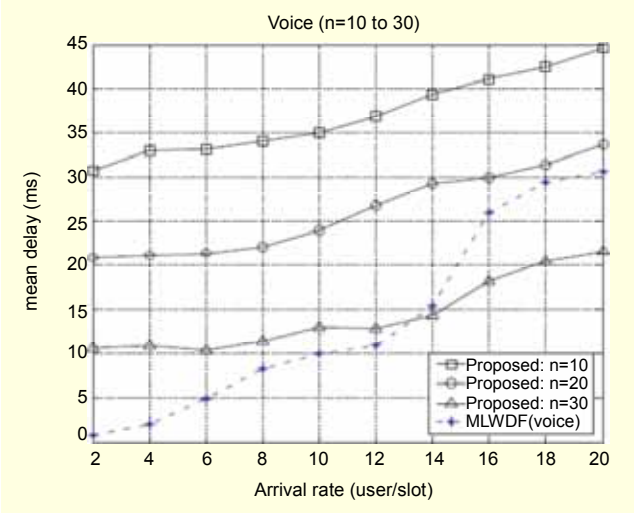

Fig. 11. Average delays of RT voice traffic under different traffic loads.

that of the M-LWDF algorithm when $MSTI_{Voice}$ = 30 as shown in Fig. 9. In particular, the packet loss requirement of the RT video traffic (< 1%) is satisfied with the UEPS algorithm until the offered traffic load reaches 16 users/timeslot when $MSTI_{Voice}$ = 10. However, the requirement is satisfied until the traffic load reaches 15 or 14 users/timeslot when $MSTI_{Voice}$ = 20 or 30, respectively.

Figure 11 shows the performances of the UEPS and M-LWDF algorithms for RT voice traffic in terms of the average delay under various traffic loads. The average delay of the UEPS algorithm is higher than that of M-LWDF algorithm when $MSTI_{Voice}$ =10 or 20 under all traffic loads. When $MSTI_{Voice}$ = 30, the average delay of the UEPS algorithm is higher than that of the M-LWDF algorithm, especially under light and medium traffic load. However, the difference vanishes as the traffic load increases, and finally the

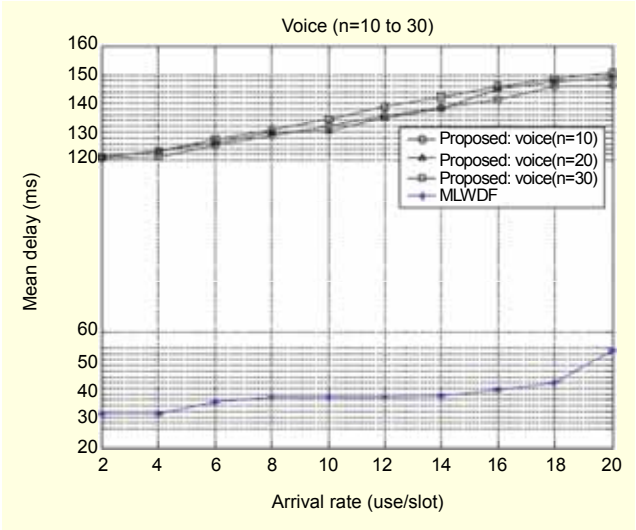

Fig. 12. Average delays of RT video traffics under different traffic loads.

average delay of the UEPS algorithm becomes lower than that of MLWDF under heavy traffic load. The delay requirement of the RT voice traffic (< 40ms) is satisfied under all traffic loads when $MSTI_{Voice}$ = 20 or 30. However, the requirement is satisfied until the traffic load reaches 14 users/timeslot when $MSTI_{Voice}$ = 10.

In the case of RT video traffic, the average delay of the UEPS algorithm is higher that that of M-LWDF under all traffic loads as shown in Fig. 12. However, the delay requirement of RT video traffic (< 150ms) is satisfied with the UEPS algorithm regardless of traffic load and $MSTI_{Voice}$ values.

B. Throughput of RT Traffics

To evaluate the throughput performance of the UEPS, PF, and M-LWDF algorithms, we generate two different traffic environments. First, we evaluate the throughput of the UEPS, PF, and M-LWDF algorithms under a traffic environment where only two NRT traffics, WWW and email traffics, are generated. The throughput performances of UEPS, PF and M-LWDF algorithms in this traffic environment under various traffic loads are shown in Fig. 13. For WWW traffic, the throughput of UEPS is higher than that of PF and M-LWDF under all traffic loads except light traffic load. In particular, the higher the traffic load, the higher the throughput of UEPS is than those of other algorithms. For email traffic, the throughput of UEPS and that of PF are almost the same, and higher than that of M-LWDF under all traffic loads.

Since the PF and M-LWDF algorithms are mainly designed to support NRT data traffics, in this subsection we evaluate the throughput performance of the UEPS algorithm and compare it with those of the PF and M-LWDF algorithms. To evaluate the throughput performance of the UEPS, PF, and M-LWDF algorithms, we generate two different traffic environments. First, we evaluate the

throughput of the UEPS, PF, and M-LWDF algorithms under a traffic environment where only two NRT traffics, WWW and email traffics, are generated. The throughput performance of the UEPS, PF, and M-LWDF algorithms in this traffic environment under various traffic loads are shown in Fig. 13. In the case of WWW traffic, the throughput of the UEPS is higher than those of PF and M-LWDF under all traffic loads except light traffic load. In particular, the higher the traffic load, the higher the throughput of the UEPS is compared to those of PF and MLWDF algorithms. In the case of email traffic, the throughputs of UEPS and PF are almost the same, and are higher than that of M-LWDF under all traffic loads.

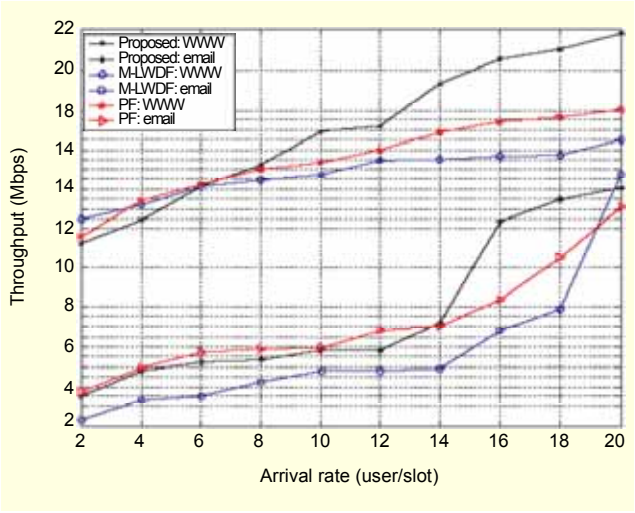

Fig. 13. Throughput of NRT traffics (WWW and email) under different traffic loads when only NRT traffics are generated.

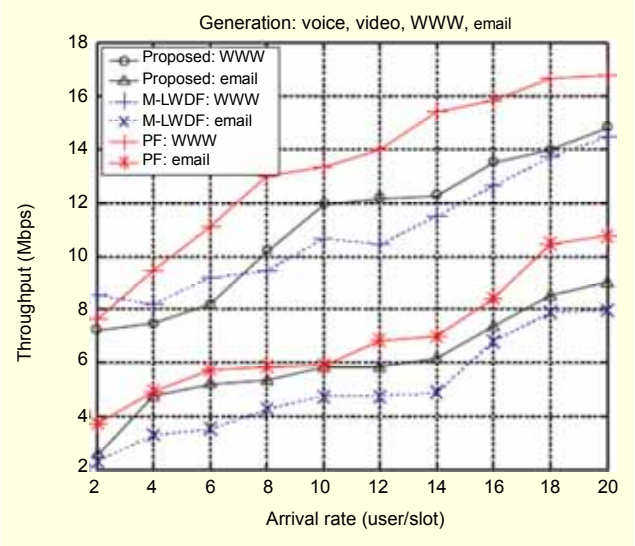

Fig. 14. Throughput of NRT traffics (WWW and email) under different traffic loads when RT and NRT traffics are generated together.

Next, we evaluate the throughput performance of the UEPS, PF, and M-LWDF algorithms for NRT traffics under a traffic environment where four traffic classes, RT voice, RT video, WWW, and email traffics, are generated. In this traffic environment, since the PF algorithm is only able to support NRT traffics, RT traffics are not generated for the PF algorithm. As a result, the PF algorithm shows higher throughput performance for NRT traffics in this traffic environment as shown in Fig. 14. However, for both WWW and email traffics, the UEPS algorithm shows higher throughput performance than the M-LWDF algorithm under most traffic load while supporting RT and NRT traffics at the same time.

VII. Conclusions and Further Study Issues

In this paper, we deigned a novel wireless packet scheduling algorithm, the *UEPS algorithm*, which is able to schedule RT and NRT traffics simultaneously by taking the urgency of scheduling and the efficiency of radio resource usage into account. The UEPS algorithm uses the time-utility function as an urgency factor and the relative status of the current channel to the average one as an efficiency factor. The main design goal of the UEPS algorithm is to maximize throughput of NRT traffics while satisfying the QoS requirements of RT traffics. The simulation study shows that the proposed UEPS algorithm shows better throughput performance than PF and M-LWDF while satisfying the QoS requirements of RT traffics under various traffic loads.

However, there are further study issues. First, it is necessary to study the UEPS algorithm for more general wireless packet scheduling situations. One possible case is when a user requests a download of RT traffics and NRT traffics from a base station at the same time. Now we are exploiting the application of the UEPS algorithm in such a general scheduling situation. In order to meet the QoS requirements of RT traffics adaptively to the various traffic situations such as different traffic load levels and different traffic mix between RT and NRT traffics, it is also needed to adjust the length of MSTI adaptively to various traffic situations. We are also developing an enhanced UEPS algorithm that is able to adjust the length of MSTI adaptively to dynamically varying traffic situations.

References

[1] J. Chen and T. Zhang, *IP-Based Next Generation Wireless Networks*, John Wiley and Sons, 2004.

[2] S. Ryu, D. Oh, G. Sihn, K. Han, and S. Park, "Research Activities on the Next Generation Mobile Communications and Services in Korea," *IEEE Communications Magazine*, vol. 43, no. 9, Sept. 2005, pp. 122-131.

[3] R. Padovani, A. Jalali, and R. Pankaj, "Data Throughput of CDMA

HDR a High Efficiency-High Data Rate Personal Communication Wireless System," *Proc. VTC2000-Spring*, July 2000, pp. 1854–1858.
[4] K. Ramanan, A. Stolyar, P. Whiting, M. Andrews, K. Kumaran, and R. Vijayakumar, "Providing Quality of Service over a Shared Wireless Link," *IEEE Communications Magazine*, vol. 39, no. 2, Feb. 2001, pp. 150–154.
[5] K. Teh, P. Kong, and S. Jiang, "Proactive Earliest Due-Date Scheduling in Wireless Packet Scheduling," *Proc. ICCT2003*, Beijing, China, 9-11, April 2003, pp. 816–820.
[6] S. Kang and A. Zakhor, "Packet Scheduling Algorithm for Wireless Video Streaming," *Proc. Packet Video 2002*, Pittsburgh, PA, 24-26 April 2002.
[7] H. Holma and A. Toskala, *WCDMA for UMTS*, 2nd ed, John Wiley and Sons, Ltd., 2002.
[8] V. Huang and W. Zhuang, "QoS-Oriented Packet Scheduling for Wireless Multimedia Communications," *IEEE Trans. Mobile Computing*, vol. 3, no. 1, Jan.-Mar. 2004, pp. 73–85.
[9] T. Schwarzfisher, "Quality and Utility - towards a Generalization of Deadline and Anytime Scheduling," *Proc. 13th Int'l. Conf. Automated Planning and Scheduling*, June 2003.
[10] E. Jensen, Real-time systems, http://www.real-time.org/realtime.htm.
[11] J. Wang et al. "Time-Utility Function-Driven Switched Ethernet: Packet Scheduling Algorithm, Implementation, and Feasibility Analysis," *IEEE Trans. Parallel and Distributed Systems*, vol. 15, no. 2, 2004, pp. 119–133.
[12] 3GPP, *Physical Layer Aspects of UTRA High Speed Downlink Packet Access* (release 2000), 3G TR25.848 V4.0.0, March 2001.
[13] T. Janevski, *Traffic Analysis and Design of Wireless IP Networks*, Artech House, Norwood, MA, 2003.

Seungwan Ryu received the BS and MS degrees from Korea University, Seoul, Korea, in 1988 and 1991, and the PhD degree from University at Buffalo, Buffalo NY, USA, in 2003, all in operations research. From 1993 to 2004, he was with Mobile Telecommunications Research Division (MTRD) at Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea, where he has participated in various research projects including CDMA mobile system, IMT-2000 system, and the next generation mobile systems. In 2004, he joined the Department of Information Systems at Chung-Ang University, Anseong, Korea, where he is an Assistant Professor. He is also working at MTRD, ETRI as an invited researcher. His research interests include the design and analysis of wireless MAC protocols, wireless packet scheduling, modelling and control of Internet traffic, and design of beyond-third-generation wireless communication systems and services.

Byung-Han Ryu received the BS and MS degrees in industrial engineering from Hanyang University, Seoul, Korea, and Seoul National University, Seoul, Korea, in 1985 and 1988, respectively, and the PhD degree in information and computer sciences from Osaka University, Osaka, Japan, in 1997. He joined the Switching Technology Division, ETRI in 1988. He is currently with Broadband Mobile MAC Research Team, Mobile Telecommunication Research Division, as a team leader. From 2000, he has been a principal member of engineering staff in ETRI. His research interests are in design, analysis, and performance evaluation of communication networks and mobile communication systems. He is a member of KIIE, KICS, IEEK, and IEICE.

Hyunhwa Seo received the BS degree from Korea University, Korea, in 2002 and the MS degree from Korea University, Korea, in 2004, all in computer and information science. In 2004, she joined ETRI, where she has participated in the next generation mobile systems. Her research interests are in random access protocol, network traffic modelling and network performance evaluation, and design of beyond-third-generation wireless communication systems and services.

Muyong Shin received the BS and MS degrees from Kyungpook National University, Daegu, Korea, in 1991 and 1994. He has been with MTRD at ETRI, where he has participated in various research projects including CDMA mobile system, IMT-2000 system, and the next generation mobile systems. He is currently researching 4G wireless communication systems.

SeiKwon Park received the BS and MS degrees from Seoul National University, Seoul, Korea, in 1978 and 1981, and the PhD degree from Texas A&M University, USA, in 1985, all in industrial engineering. From 1985 to 1987, he was at Network Planning Division of ETRI. From 1987 to 1990, he was at Software House of Ministry of Agriculture and Forestry (MAF) and Korea Rural Economics Institute (KREI), Seoul, Korea, as a senior research fellow. In 1990 he joined the Department of Information Systems at Chung-Ang University, Ansung, Korea, where he is a Professor. His research interests include system analysis/design methodologies for web-based information services and modelling of performance evaluation systems.