Channel-Based Scheduling Policy for QoS Guarantees in Wireless Links☆

김 정 근*　　　홍 인 기**
Jeong Geun Kim　　　Een-Kee Hong

Abstract

Proportional Fair (PF) share policy has been adopted as a downlink scheduling scheme in CDMA2000 1xEV-DO standard. Although it offers optimal performance in aggregate throughput conditioned on equal time share among users, it cannot provide a bandwidth guarantee and a strict delay bound, which is essential requirements of real-time (RT) applications. In this study, we propose a new scheduling policy that provides quality-of-service (QoS) guarantees to a variety of traffic types demanding diverse service requirements. In our policy data traffic is categorized into three classes, depending on sensitivity of its performance to delay or throughput. And the primary components of our policy, namely, Proportional Fair (PF), Weighted Fair Queuing (WFQ), and delay-based prioritized scheme are intelligently combined to satisfy QoS requirements of each traffic type. In our policy all the traffic categories run on the PF policy as a basis. However the level of emphasis on each of those ingredient policies is changed in an adaptive manner by taking into account the channel conditions and QoS requirements. Such flexibility of our proposed policy leads to offering QoS guarantees effectively and, at the same time, maximizing the throughput. Simulations are used to verify the performance of the proposed scheduling policy. Experimental results show that our proposal can provide guaranteed throughput and maximum delay bound more efficiently compared to other policies.

☞ Keyword : Scheduling, QoS, Proportional Fairness, WFQ, CDMA2000 1xEV-DO

1. Introduction

With the explosive growth of the Internet and rapid proliferation of personal communication services, the demand for data services in wireless networks has been ever-increasing. Responding to this demand, wireless networks have been evolving toward packet-switched architectures that are more flexible and efficient in providing packet data services. The CDMA2000 1xEV-DO standard (abbreviated hereafter as 1xEV-DO) is one of such architectures, that is designed to support data services in the third generation wireless network [1]. Compared to previous wireless networks often characterized by circuit-switched and voice-oriented architecture, the 1xEV-DO system has several unique features. Notably among them is the "opportunistic" scheduling used to schedule downlink (or equivalently forward link) transmission.

The basic idea behind the opportunistic scheduling is that temporal channel variation of multiple users is taken into account in scheduling [2]. Thus at a given time the scheduling decision favors a mobile user currently seeing a better channel. If all the traffic is "elastic" (i.e., having flexible service requirements), such opportunistic scheduling mechanisms then greatly improve throughput performance and yield higher bandwidth utilization. The 1xEV-DO standard adopted an opportunistic scheduling called Proportional Fair (PF) share policy as a downlink scheduler.

---

* 정 회 원 : 경희대학교 전자정보대학 교수
jg_kim@khu.ac.kr(제 1저자)
** 정 회 원 : 경희대학교 전자정보대학 교수
ekhong@khu.ac.kr(공동저자)

☆ 이 연구는 2003년도 경희대학교 지원에 의한 결과임 (KHU-20030456).

Although the PF policy offers optimal performance in aggregate throughput conditioned on equal time share among users [3], it cannot provide a bandwidth guarantee and a strict delay bound, which is essential requirements of real-time applications. As the 1xEV-DO system expands its service realm from non-real-time (NRT) to real-time (RT) applications like video streaming, such a limitation triggers a lot of research efforts toward expanding its capability for supporting various quality-of-service (QoS) requirements of RT traffic.

The authors in [4] proposed a scheduling discipline called the Exponential Rule in which the queue with larger weighted delay of head-of-line (HOL) packet gets higher priority in transmission. This rule behaves like PF when the weighted delay difference is not relatively large. As the difference becomes significant, this policy gracefully adapts from PF to the Exponential rule. However, this policy suppresses the chance the HOL packet with relative lower weighted delay gets selected for transmission even when it currently sees a better channel. This may lead to lower throughput performance. In [5], Liu et al. proposed an utility-based scheduling algorithm in which utility is a decreasing function of packet delay. By scheduling the transmission at each time in a way the total utility rate is maximized, this scheme can provide a strict delay bound. However, it is questionable to how the utility function can be devised for throughput-sensitive applications. Furthermore the existence of utility function and its derivative is not obvious for applications with diverse requirements. To address the issues of fairness and throughput guarantee in the PF policy, a dynamic rate control algorithm was proposed in [6]. In this scheme, a target throughput is set along with the associated minimum and maximum throughput, and the scheduler attempts to maintain user's perceived throughput within those bounds. However this scheme can only work with throughput-sensitive applications, but may not suit other traffic categories.

The main contribution of this paper is that we propose a novel "opportunistic" service rule that can flexibly support a wide range of traffic categories. Most of previous works in [5,6] focuses on applications sensitive to either delay or throughput. In contrast, our scheme can deal with QoS needs required by both traffic categories using a simple metric. In addition, throughput guarantee can be supported, depending on urgency, in a either strict or relaxed manner via configurable parameters. Such a capability is useful when throughput guarantee is not necessarily required over microscopic time-scale, but rather needed on long-term basis. Moreover in order to maximize the link capacity, the RT traffic runs on PF rule as long as delay or throughput of its packet does not deviate much from the target. However, as its QoS target is more likely to be missed, the corresponding traffic gains more weight in scheduling decision and increases the chance for transmission. This feature of graceful adaptation from PF to the prioritized mode can provide a maximized throughput and, at the same time, offers an accurate tool for exercising QoS. Our numerical results indicate that the proposed rule can deliver QoS guarantees under realistic service scenarios over wireless links which are primarily characterized by heterogeneous channel conditions and mixed QoS requirements. Moreover, the throughput gains are observed against other schemes and the flexible throughput guarantees are proven to work effectively as designed and give a capacity improvement.

The rest of the paper is organized as follows. In Section 2 we give overview of CDMA2000 1xEV-DO standard and describe the proposed scheduling algorithm. Numerical results and simulations are reported in Section 3, followed by concluding remarks in Section 4.

2. Scheduling Algorithm

2.1 Background

To utilize temporal channel variation, the PF share policy in 1xEV-DO requires the mobile terminals (MTs) to report continuously their channel state information to the scheduler in the access point (AP). Accordingly, the MTs report back the channel condition to the network every 1.667 msec. The pilot bursts from the AP enable the MT to accurately estimate the channel conditions. The channel state information is sent back to the AP in the form of data rate request (see Table 1) through data rate request channel (DRC) in reverse link [1]. Once data on DRC from each MT is gathered, the PF share scheduler selects the MT i^* which satisfies

$$i^* = arg_i \max \frac{DRC_i(t)}{\hat{R}(t)} \tag{1}$$

where $DRC_i(t)$ is the data rate request of MT i at time $t, \hat{R}_i(t)$ is average transmission rate of MT i by time t.

In 1xEV-DO, downlink sharing among multiple MTs is done in time-division multiplexing (TDM) basis, where fixed-size time-slot(s) are dedicated to each MT based on scheduling decision. In particular, different modulation schemes including QPSK, 8PSK, and 16QAM are employed in an adaptive manner, depending on the channel condition of the target MT. The set of available data rates for the corresponding channel conditions is listed in Table 1. To provide fairness in channel usage among the MTs, the average rate $\hat{R}_i(t)$ of ith MT is updated every time-slot as follows:

$$\hat{R}_i(t+1) = (1 - \frac{1}{t_c})\hat{R}_i(t) + \frac{1}{t_c}R_i(t)$$

where is t_c the time constant set to 1000 slots [2], and $R_i(t)$ is the current rate of i th MT. At the end of current transmission the scheduler determines the next MT to transmit based on the criterion in (1) and this procedure repeats.

<Table 1> Data rate options in CDMA2000 1xEV-DO.

Nominal Data Rate (Kbps)	Nominal Slots per PHY packet	Total Bits per PHY Packet	E_c/N_t (dB) thresholds for DRC selection
38.4	16	1024	-13.5
76.8	8	1024	-10.5
153.6	4	1024	-7.4
307.2	2	1024	-4.3
614.4	1	1024	-1.0
921.6	2	3072	1.5
1228.8	1	2048	3.7
1843.2	1	3072	7.1
2457.6	1	4096	9.1

2.2 Algorithm

In a majority of previous works, prioritized services and associated QoS capability have been implemented by incorporating a weighting function $f_i(\cdot)$ into the PF metric in (1). The modified metrics in general have the following form:

$$i^* = arg_i max \frac{DRC_i(t)}{\hat{R}_i(t)} f(\cdot).$$

Different weight functions $f_i(\cdot)$'s have been chosen depending on which performance factor (e.g., bandwidth, delay) is prioritized. The Exponential rule in [4] has a weight function given by

$$f(\cdot) = exp\left(\frac{a_i W_i(t) - \overline{aW}(t)}{\hat{R}_i(t)}\right) \tag{2}$$

where a_i is the weight of ith flow, $W_i(t)$ is the waiting time of the packet, and $\overline{aW}$ is mean weighted delay. In [6], $f_i(\cdot)$ is defined as a group of functions whose values are proportional to deviation from the target rate. See [6] for details.

As pointed earlier, these rules lack the flexibility of handling a diverse requirements. To overcome this limitation, our rule is designed to support three primary types of traffic which most network applications can be categorized into:

Class I : Delay-sensitive
Class II : Throughput-sensitive
Class III : Best-effort.

Such a capability of providing differentiation in service is essential in operating an anticipating multi-service wireless networks where a wide array of traffic types coexist demanding diverse requirements. In our rule, the metric is simple to calculate and the associated parameters are flexibly configured depending on the service category.

We propose the rule named the "Adaptive policy" whose metric is given by:

$$i^* = arg_i \max \frac{DRC_i(t)}{\hat{R}_i(t)} (C_{wd}(t)\overline{(D_i(t)} + 1)^3 + 1)^{w_i(t)} \tag{3}$$

In the above equation $C_{wd}(t)$ is a coefficient that determines the impact of the weight function inside the parenthesis and is set to the maximum ratio of DRC over the average rate at each scheduling epoch, that is,

$$C_{wd}(t) = msx_i \frac{DRC_i(t)}{\hat{R}_i(t)}$$

$W_i(t)$ is the weighting factor controlling the level of emphasis on the weighting function in (3). $\overline{D_i(t)}$ is the normalized waiting time given by

$$\overline{D_i(t)} = \frac{D_i(t) - D_{i,max}}{D_{i,max}} \tag{4}$$

where $D_i(t)$ is the waiting time of the HOL packet in ith queue and $D_{i,max}$ is the maximum tolerable delay of the HOL packet. $D_{i,max}$ is specified as a QoS parameter for each class, which is given by:

$$D_{i,max} = \begin{cases} D_{i,max}, & \text{for } Class\ I \\ max(t_{a,i}, F_{i,-1}) + \frac{L_i}{R_i}, & \text{for } Class\ II \\ \infty, & \text{for } Class\ III \end{cases} \tag{5}$$

For the traffic sensitive to delay (Class I), the maximum delay is specified by $D_{i,max}$ as input QoS parameter. For best-effort traffic (Class III), the maximum delay is set to infinity, thus the weighting function in (3) becomes the unity and the bandwidth is shared by the PF rule among the MTs of the same class. For the Class II traffic which is sensitive to throughput, $D_{i,max}$ is set to the finish time of the HOL packet of the i th flow. Finish time is the notion central to service discipline called weighted fair queueing (WFQ) and is defined for the ith flow by [7]:

$$F_i(t) = max(t_{a,i}, F_{i,-1}) + \frac{L_i}{R_i} \tag{6}$$

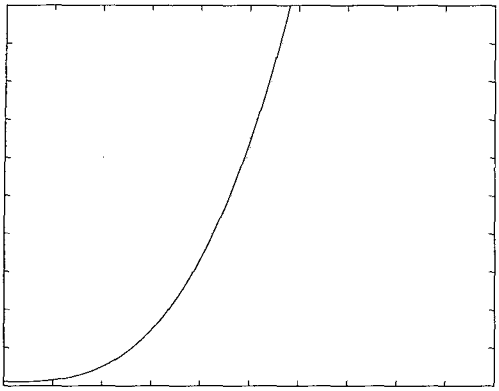
where $t_{a,i}$ is the arrival time of the HOL packet, $F_{i,-1}$ is the finish time of the previous packet, L_i is the packet length, and R_i is the promised rate (or bandwidth). WFQ scheme attempts to emulate packet flow in ideal fluid model by calculating the departure time of a packet (i.e., finish time) in a corresponding fluid model and using this virtual time stamp to schedule packets. The advantage in offering throughput guarantees via WFQ-like policy against the periodic counter in [4] is more accurate and fair in distribution of bandwidth among the competing flows. Note that a significant difference in running WFQ over wireless channels compared against wireline channels is the available bandwidth of wireless link fluctuates according to the channel conditions. The 1xEV-DO also has a finite set of achievable transmission rates for each channel condition as listed in Table 1.

Since there often exists mismatch between physical packet size and data size from the higher layer, the expression of finish time in (6) needs to be modified to take into account the case in which multiple data packets are transmitted over a single physical packet, or vice versa. In this case, the representative arrival time t_a of a group of packets including the HOL packet and the subsequent ones is set to the one from the HOL packet. Once that flow is selected for transmission, the subsequent packets may enjoy a free ride and the flow may get more bandwidth than necessary. To correct this issue, the actual finish time is calculated by taking into account the total amount of bits transmitted over the single physical packet. Thus the actual finish time for n packets to be transmitted over a single physical packet given that the kth packet of the ith flow is at HOL is

$$\begin{aligned} F_i(t) &= max(t_{a,(i,k)}, F_{i,k-1}(t)) \\ &+ \sum_{j=k}^{k+n-1} \frac{L_{i,j}}{R_i} + \sum_{j=k}^{k+n-2} (t_{a,(i,j+1)} - F_{i,j}(t)) \cdot 1_{\{t_{a,(i,j+1)} > F_{i,j}(t)\}} \end{aligned} \tag{7}$$

where 1_x is the indicator function whose value is 1 if the condition x is satisfied and 0 otherwise.

Figure 1 shows dominance of weighting function $((\overline{D_i(t)}+1)^3+1)$ as the delay of HOL packet approaches the prescribed target. Here, $C_{wd}(t)$ is set to 63 for illustration. As shown in the figure, the rule is designed so that the flows are scheduled on the PF rule when the current delay has a sufficient margin from the target, but the delay-based priority part in the metric gradually overrides the PF rule as the delay approaches to the target. By scheduling this way our proposed rule can utilize temporal channel variation and consequently maximize the channel capacity. In



〈Figure 1〉 Weight function $((\overline{D_i(t)}+1)^3+1)$ versus normalized waiting time $\overline{D_i(t)}$

contrast, the Exponential rule (2) suppresses selection of the HOL packets whose weighted delays are less than mean value regardless of the channel conditions.

The weighting factor $w_i(t)$ is another parameter configured at the connection setup phase, according to service requirements of traffic. The values of $w_i(t)$ for each traffic class is set as follows:

$$w_i(t) = \begin{cases} 1 & \text{for } \mathit{Class\ I} \\ \mathit{variable}\,(0-1), & \text{for } \mathit{Class\ II} \\ 0, & \text{for } \mathit{Class\ III} \end{cases} \tag{8}$$

The $w_i(t)$ is fixed at 1 for Class I. In this case the resulting rule behaves like a delay-based priority scheme for lagging flows $(\overline{D_i(t)} > 0)$, whereas leading or in-sync flows $(\overline{D_i(t)} \leq 0)$ still runs on the PF to maximize the throughput given equal time share to those flows. For Class III, $w_i(t)$ is fixed at 0, suppressing the weighting function and equalizing the rule to the PF scheme. For Class II, $w_i(t)$ varies depending on how strictly the bandwidth guarantee is required.

3. Numerical Results and Simulations

In this section, we present simulation results for the scheduling policy described in the previous section. We consider a CDMA cell in which several MTs run different applications under time-varying channel conditions. The channel models are Rayleigh fading, which is further approximated into finite-state Markov channel model following the approach in [8]. The Markov channel model has 10 channel states that correspond to the DRC rates listed in Table 1. The E_c/N_t thresholds for DRC selection in Table 1 are used to determine the ranges of E_c/N_t mapping onto the states of the discrete Markov channel model [9]. The speed of a MT determines how fast the channel varies, which will be characterized by the transition probability of the Markov channel model.

We simulate a scenario of mixed traffic in which MTs with RT traffic or with NRT traffic are uniformly positioned in a cell, competing for downlink bandwidth. We assume that NRT traffic generates packets following a Poisson process, whereas RT source traffic is modelled by the real-time video streaming model in [10]. In this model, a video streaming session consists of a sequence of frames with the period of 100 msec. The number of packets generated in each frame is fixed at 8, and the packet inter-arrival time and packet size is distributed by truncated Pareto distribution:

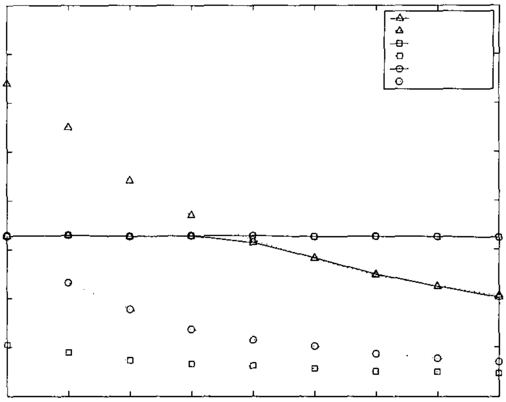
$$F(x) = \begin{cases} 1 - \dfrac{K^s}{x^s}, & x < m, \\ 1, & x \geq m, \end{cases}$$

with s = 1.2, K = 2.5 msec, m = 12.5 msec for the inter-arrival time, and s = 1.2, K = 20 bytes, m = 125 bytes for the packet size, respectively. With these parameters the mean rate of a single video traffic is 32 Kbps. For the sake of simplicity, no packet error is assumed. Since DRC rate is selected to target at 1 % packet error rate [1], we believe the impacts of packet error on system performance can be safely ignored. All the simulation results are reported based on 1000 sec run, equivalent to 1,670,000 slots. Table 2 summarizes the values of the various parameters in simulation.

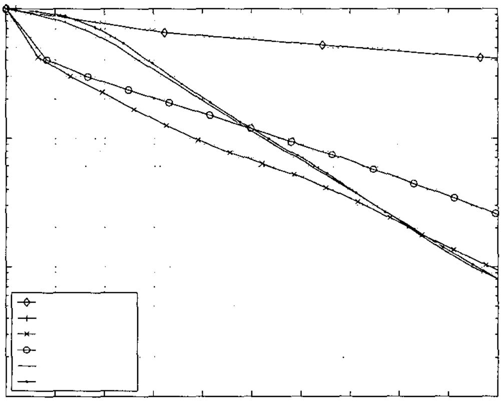
Figure 2 shows throughput and delay CDF (Complementary Distribution Function) curves for scheduling policies: PF, EXP (Exponential rule in [4]), and Adaptive rule proposed in this paper. All the MTs are under homogeneous channel conditions, i.e., $E_c/N_t = 0$ dB and v = 3 Km/h. In terms of throughput, the PF policy shows the best performance due to its property of placing absolute priority on the MT seeing the best channel. However, throughput of RT traffic reduces proportionally as the system gets loaded with more traffic. Obviously it is because RT traffic is equally treated with other NRT traffic. In contrast, EXP and Adaptive policies guarantee a constant throughput (i.e., 32 Kbps) irrespective of traffic con-

〈Table 2〉 Parameter values used in the simulations

Parameters	Symbol	Value
Number of RT mobiles	N_{rt}	3-5
Number of NRT mobiles	N_{nrt}	10-50
Maximum inter-arrival time of NRT traffic	-	10-50
Maximum inter-arrival time of NRT traffic	-	0.01 sec
Mean packet size of NRT traffic	-	640 bits
Video frame period	-	100 msec
Numbr of packets in a video frame	-	8
Mean rate of video traffic	-	32 Kbps
Mean inter-arrival time of video packets	-	12.5 msec
Mean packet size of video traffic	-	50 bytes
Maximum packet size of video traffic	-	125 bytes
Mobile speed (Km per hour)	v	3-120 Km/h

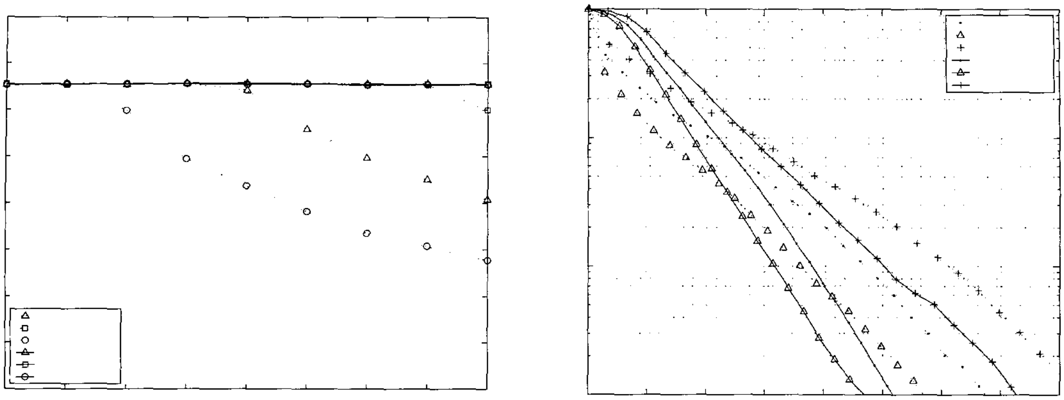


(a) Throughput (bps)



(b) Delay CDF

〈Figure 2〉 Throughput and delay CDF of MTs with RT and NRT traffic ($v = 3$ Km/h)



(a) Throughput (bps)

(b) Delay CDF

⟨Figure 3⟩ Throughput and delay CDF under heterogeneous channel environments ($v = 3$km/h, N_{nrt} = 30).

ditions. In particular, the Adaptive policy yields higher throughput in NRT traffic than the EXP policy. For N_{nrt}= 10, the Adaptive policy achieves twice the throughput of NRT traffic that EXP policy does. In the curves of delay CDF, the PF policy incurs an extremely long delay relative to others. In EXP and Adaptive policies, the delay CDF is somewhat similar, although Adaptive policy shows less variation as N_{nrt} increases from 10 to 50.

Figure 3 shows throughput and delay CDF of RT traffic under heterogeneous channel conditions in which mean E_c/N_t of 0 dB, -3 dB, and 3 dB is given to three MTs, respectively. As show in Fig. 3, throughput decreases proportionally with worse channel conditions under the PF policy, whereas the Adaptive policy offers a steady throughput irrespective of channel conditions. In delay CDF, both EXP and the Adaptive policies miss 0.2 sec delay bound. However, the EXP policy fails at 0 and -3 dB, but the Adaptive policy does only at -3 dB. A similar behavior is observed for v = 50 Km/h. One difference is both schemes now meet 0.2 sec delay bound with a significant margin.

Delay performance under heterogeneous QoS requirements is shown in Fig. 4. Five MTs having RT traffic set maximum delay bound at 0.1, 0.2, 0.3, 0.4, and 0.5 sec, respectively. In the Adaptive policy, those bound are set to the parameter of $D_{i,max}$. In EXP policy, the weight factor is set following the formula $a_i = - log(\delta_i)/T_i$ suggested in [4]. Here δ_i and T_i is derived from the following delay requirement:

$$P[Delay > T_i] \le \delta_i.$$

Thus, δ_i is set to 0.01 and T_i is set to from 0.1 to 0.5 as above. For the case of v = 3 Km/h, both policies did not deliver successfully the specified QoS requirements. However when it comes to the level of deviation, the EXP policy far exceeds the desired delay bound, whereas the Adaptive policy offers a delay bound relatively close to the specified values. For v = 50 Km/h, both policies meet the maximum delay bound as specified, but in a somehow conservative manner. In particular, the EXP policy exercises too much bandwidth as shown in Fig. 4 (b). From those results, the Adaptive policy can offers a maximum delay bound more accurately relative to the EXP policy.

4. Conclusions

In this paper, we proposed a new scheduling policy that provides QoS guarantees to various traffic types demanding diverse service requirements. In our proposal, data traffic is categorized into three classes, depending on sensitivity of its performance to delay or throughput. And the primary components of our policy, namely, the PF, WFQ, and delay-based prioritized scheme are intelligently combined to satisfy QoS requirements of each traffic type. Our policy changes the level of emphasis on each of those ingredient policies in an adaptive manner, taking into account the channel conditions and QoS requirements. Such flexibility, as shown in numerical results, leads to offering QoS guarantees effectively and, at the same time, maximizing the throughput.

References

[1] P. Bender, P. Black, M. Grob, R Padovani, N Sindhushayana, and A. Viterbi, "CDMA/HDR: a bandwidth-efficient high-speed wireless data service for nomadic users," IEEE Commun. Mag., vol. 38, no. 7, pp. 70-77, July 2000.

[2] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR: a high efficiency, high data rate personal com munication wireless system," in Proc. IEEE VTC '00, May 2000, vol. 3, pp. 1854-1858.

[3] F. P. Kelly, "Charging and rate control for elastic traffic," European Transactions on Telecommunications, vol. 8, pp. 33-37, 1997.

[4] S. Shakkottai and A. Stolyar, "Scheduling algorithms for a mixture of real-time and non-real-time data in HDR," in Proc. International Teletraffic Congress '2001, Sept. 2001.

[5] P. Liu, R. Berry, and M. L. Honig, "Delay-sensitive packet scheduling in wireless networks," in Proc. IEEE WCNC 2003, Mar. 2003, vol. 3, pp. 1627-1632.

[6] C. Y. Huang, H. Y. Su, S. Vitebsky, and P. Chen, "Schedulers for 1xEV-DO: third generation wireless high-speed data systems," in Proc. IEEE VTC '03, Apr. 2003, vol. 3, pp. 1710-1714.

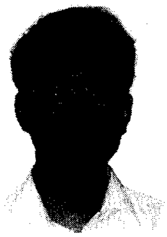
[7] H. Zhang, "Service discipline for guaranteed performance service in packet-switching networks," Proc. IEEE, vol. 83, no. 10, pp. 1374-1396, Oct. 1995.

[8] H. S. Wang and N. Moayeri, "Finite-state Markov channel-a useful model for radio communication channels," IEEE Trans. Veh. Technol., vol. 44, no. 1, pp. 163-171, Feb. 1995.

[9] D. Paranchych and M. Yavuz, "Analytical expression for 1xEV-DO forward link throughput," in Proc. IEEE WCNC '03, Mar. 2003, vol. 1, pp. 508-513.

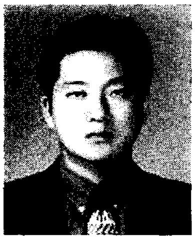
[10] 3GPP2, 3GPP2 1xEV-DV Evaluation Methodology, July 2001.

◐ 저 자 소 개 ◑



김 정 근

1990년 연세대학교 전기공학과 졸업(학사)
1992년 연세대학교 대학원 전기공학과 졸업(석사)
2000년 University of Arizona, 전기공학과(박사)
2000~2003 QUALCOMM, Senior Systems Engineer
2003~현재 경희대학교 정보통신대학원 교수
관심분야 : 통신 네트워크, 네트워크 성능분석
E-mail : jg_kim@khu.ac.kr



홍 인 기

1989년 연세대학교 전기공학과 졸업(학사)
1991년 연세대학교 대학원 전기공학과 졸업(석사)
1995년 연세대학교 대학원 전기공학과 졸업(박사)
1995~1999 SK Telecom, 중앙연구원
1997~1998 NTT DoCoMo, Mobile Network Lab
1999~현재 경희대학교 전자정보학부 교수
관심분야 : CDMA, 이동통신
E-mail : ekhong@khu.ac.kr