Cross-Layer Design for Mobile Internet Services in Cellular Communications Systems

Dong Geun Jeong
Hankuk University of Foreign Studies

Recently, cross-layer design approach has been greatly attracting researchers' attention as an alternative for improving the performance of wireless data networks. The main reason why cross-layer approaches are particularly well suited for wireless networks is that there exists direct coupling between physical layer and upper layers. Therefore, with cross-layer approach, the protocol designers try to exploit the interaction between layers and promote adaptability at all layers, based on information exchange between layers. In this article we focus on the cross-layer engineering for high data-rate mobile Internet services through cellular networks. First, the general considerations in cross-layer engineering are outlined. Then, we discuss the common approach in literatures, which mainly deals with adaptability in physical and medium access control layer. Finally, we show that the cross-layer engineering taking account of all layers is more adequate for the mobile Internet services through cellular networks.

As the boundary of cellular mobile communications services has been continuously expanded, the cellular systems will provide a variety of multimedia services e.g., high-speed Internet access, streaming and interactive video/audio, and mobile game. The diverse quality of services (QoS) requirements from these services should be satisfied by using physically unreliable radio spectrum. In addition, since the radio is very scare resource, the most important engineering issue is to improve the radio resource efficiency.

Recently, an approach to improve radio efficiency remarkably has attracted researchers' attention, which is the "cross-layer design and protocol engineering" [1], [2]. Layered architecture is widely used in protocol implementation, because the modularization of protocol stack leads to many advantages, for example, the scalability and the ease of standardization. The traditional approach to protocol design and engineering for layered protocol stack is the "layered-design approach." With this approach, each layer performs a well defined function that is separated from others. The boundaries of layers are chosen so as to minimize the information flow across the interface between layers. The most important merit of this approach is that the change in one layer does not require changes in other layers. Thus, it is flexible to deploy a new protocol for specific layer(s). However, with this approach, some redundancies are inevitable, which are distributed more than one layer. In mobile communications with portable devices, the redundancy should be avoided to minimize

battery power consumption. In mobile communications, the time-varying radio channel largely affects on the operation of other layer protocols. That is, there exists direct coupling between physical layer and upper layers. In this case, it is inadequate to optimize each layer independently.

With the cross-layer protocol engineering, the interactions between layers are seriously taken into account. It is noted that the repeal of layered protocol structure and the integration of all layers are not practical, considering protocol implementation, upgrading, and standardization. In addition, for internetworking with existing wired Internet, layered structure of wireless protocol is inevitable. The cross-layer approach does not avoid the interactions between different layers but exploits them and aims at joint optimization of two or more layers, while maintaining the layered structure.

There are various types of wireless and mobile communications systems in operation or under development, such as 3G/4G cellular networks, wireless local area networks (WLANs), and ad-hoc networks. Although all these systems support multimedia services, each type of systems has their own main application in a specific environment. Since the cross-layer design and protocol engineering aims at the performance improvement from the standpoint of the *whole system*, it is difficult to develop a general methodology that can be applied to any type of networks. For example, in cross-layer engineering for ad-hoc networks, the interaction of network layer with the medium access control (MAC) and physical (PHY) layer is of importance since routing is a core engineering issue in ad-hoc networks. On the other hand, the cross-layer engineering with network layer is not important in the cellular systems, where most functions of this layer are implemented not in the wireless network but in the core network. In this article, we focus on cellular systems, e.g., cdma2000, WCDMA, and WiBro (or IEEE 802.16e). Specifically, we give our attention to the cross-layer engineering for *high data-rate mobile Internet services*.

First, we outline the general consideration in cross-layer engineering. Then, we review a common approach in most literatures that deal with cross-layer engineering for cellular systems and discuss the limitation in the approach. That is, they propose the schemes for the cross-layer engineering between a part of layers in protocol stack and/or analyze their performance. We demonstrate that this approach cannot be optimal from user' s viewpoint. Finally, we show why the cross-layer engineering not between a part of layers but *among all layers* is important. For this purpose, we exemplify cross-layer engineering for web browsing and streaming video services that respectively generate the best-effort (BE) traffic and real-time (RT) traffic.

Internet protocol stack consists of PHY, data link, network, transport, and application layer (see Fig. 1). Cross-layer engineering exploits the interaction between layers to improve channel efficiency, while guaranteeing QoS of users.

Since QoS is essentially a matter of application services, the "QoS measures" also should represent service quality from the viewpoint of application users. For example, in the web browsing, the web page downloading delay can be the ultimate QoS measure. Nevertheless, the QoS measures at lower layers are also helpful to grasp the performance of each layer to support given services and optimize the layer operation.

The role of transport layer is reliable data transmission between two processes (in this article, we will refer to this as end-to-end transmission). At a transport layer, the end-to-end delay and the end-to-end throughput can be the

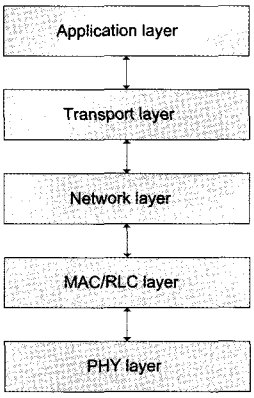


Fig. 0. Internet protocol stack.

measures. For the cellular systems, network layer cannot be of interest since most functions of this layer are implemented in the wired core network. For the ad-hoc networks, the parameters related to the routing performance can be the measures. Especially for the sensor networks, the network life time also can be an important QoS measure. For the data link layer, our concern is on the MAC and the radio resource management (RRM) sub-layers. At MAC/RRM layer, the access delay and throughput can be important for the BE traffic, while the loss probability is a major parameter for the RT traffic. At PHY layer, signal-to-interference ratio (SIR), bit error rate (BER), and frame error rate (FER) can be used as the QoS measure. There can be also some QoS measures that are identified across several layers. The examples include the blocking or dropping probability of new connection (including circuit call) or handoff connection by connection admission control mechanism.

In cross-layer design of a scheme, even if the scheme locates at a specific layer, the design should aim at improving the QoS at application layer (i.e., QoS of user) rather than that at any other lower layers.

In mobile communications systems, not only traffic but also (physical) channel is time-varying. QoS and system performance can be improved by promoting adaptability to the traffic and channel variation. Since there exists tight interdependence between layers in wireless systems, the adaptability of the system can be improved by exploiting the interactions between layers.

System can respond to the traffic variation relatively easily since it is an input parameter. For a variety of traffic types with different QoS requirements, traffic adaptation is performed at several layers. For example, a scheduler at MAC/RRM layer determines the transmission priority of traffic, while adaptive spreading gain is used at PHY (e.g., cdma2000 or WCDMA).

On the other hand, link adaptation is referred as the adaptive mechanism to the channel variation, which includes power control, adaptive modulation and coding (AMC), retransmission, and hybrid automatic repeat request (HARQ). The AMC is a task of PHY, whereas the retransmission is performed by link layer or transport layer. Note that the link adaptation mechanisms can be implemented at more than one layer. Therefore, in cross-layer design, it should be determined which layers take the responsibility of adaptation to channel variation.

A feature of cross-layer protocol engineering is the joint optimization of several layers. In general, the whole layers need not to be jointly optimized together, although it ultimately aims at improving QoS of user. The group of layers that can be jointly optimized effectively are different for each system or application. For example, in ad-hoc networks, since the routing algorithm in network layer and the multiple access schemes in MAC layer are closely affected with each other, the joint optimization of these two layers may improve system performance. For the mobile Internet services in cellular systems, we will show later that

QoS is improved remarkably when all the layers together are optimized jointly.

Since the cross-layer engineering is a matter of joint optimization, it can be regarded as the convergence process between the QoS requirements from application layer (the top of protocol stack) and link adaptation for channel from physical layer (the bottom of protocol stack). Thus, an engineering issue is which layer has the main role of convergence. For example, in the TCP operation over wireless networks, some researches suggest modification of TCP, whereas other researches propose a new packet transmission scheduler at radio interface, which manages resource allocation to adapt the TCP operation to channel variance. Obviously, the TCP at transport layer plays a main convergence role in the former approach, while the scheduler at MAC layer plays a role for convergence in the latter approach.

For example, consider a packet transmission scheduler at MAC layer, which has been designed so as to be optimal under the assumption of Poisson traffic arrival process (or the full load assumption). However, if TCP is used as transport layer protocol, due to the congestion control mechanism of TCP, the traffic arrival process to MAC layer is not Poisson (or full load). In this case, the scheduler cannot be optimal. A similar argument can be made about the interaction between application layer (or human interaction) and TCP layer.

This example shows the importance of grasping the interaction between layers. However, it is hard to characterize fully the interactions between protocols at different layers. Table 1 shows a brief summary of interaction between layers.

The performance evaluation is not avoidable when designing or modifying a protocol or scheme. As the methodology, to choose between the simulation and the theoretical analysis is tradeoff between complexity and accuracy. The performance evaluation based on cross-layer engineering is usually very complicate since the operations of all layers are considered together. An excessive simplification of model for mathematical tractability could not provide practically sound results. In usual case, the system level simulation, where all of PHY, MAC/RRM, network, transport, and application layer are implemented, can be the better method for performance evaluation although it requires time and cost [3]-[5].

Table 1. Interaction between layers.

	Protocol Parameters	Interaction with (layers)
Application layer	Traffic pattern QoS requirement	MAC/RLC PHY Transport
Transport layer (TCP)	Congestion window Retransmission timer	PHY Network*
Network layer*	Routing strategy (energy aware or not) Dynamic rerouting Topology (link status)	PHY MAC/RLC
MAC/RLC layer	Retransmission number, Sleep interval Scheduling	PHY Network* Application
Physical layer	Node location* Radio transmission range* Link adaptation scheme Power control SNR	MAC/RLC Network* Transport Application

*for ad-hoc networks

In near future, the services generating massive traffic in cellular systems will be Internet services including web browsing and video-on-demand (VOD). Fig. 2 shows the model of mobile Internet services using cellular system [3]. Up to date, most researches about the cross-layer engineering for mobile Internet services have considered only a part of layers. We hereafter examine two such cases,

one is cross-layer engineering between PHY and MAC layer and the other is that between PHY layer and TCP (transport layer).

There have been many studies on cross-layer engineering between PHY and MAC layer, e.g., [6]. Usual strategy is the adaptive resource management in MAC layer, taking account of link adaptation mechanisms, or at least, channel variation at PHY layer. We first outline this approach and then discuss a solution to overcome its limit.

PHY provides link adaptation mechanism, which is the process of modifying the transmission parameters to compensate for the changing channel condition. Power control is a traditional and fundamental link adaptation technique for wireless system. Advanced cellular systems exploit AMC and HARQ. With AMC, a system adjusts the modulation and coding scheme (MCS) level for a unit of data (say, frame) of a user according to the channel quality. Since the channel quality for a user is a time-varying parameter, the adequate MCS level and the corresponding data rate also vary over time.

The design objective of MAC (RRM) layer is optimal resource sharing among users with different classes of services. MAC layer algorithms based on cross-layer design concept exploit the information from PHY. For example, with the proportional fair scheduling scheme [7] designed for 1xEV/DO systems, the transmission priority for user i is determined as $r_i(t)/X_i(t)$, where $r_i(t)$ is an instantaneous data rate of user i at frame t and $X_i(t)$ denotes the throughput of user i measured at frame t. In this scheduling algorithm, $r_i(t)$ is determined by the AMC scheme at PHY layer, reflecting physical channel condition, while $X_i(t)$ is used for fairness among users.

However, this approach does not reflect the interaction between MAC/PHY and upper layers. The traffic arrival process at MAC layer is a result of interaction between layers and is a complex one as will be shown later. Accordingly, the cross-layer engineering only between MAC and PHY layer may fail to achieve high efficiency in many situations.

In wireless and mobile Internet access, TCP is widely used as an end-to-end transport layer protocol. However, since TCP is originally designed for the wired networks, it cannot

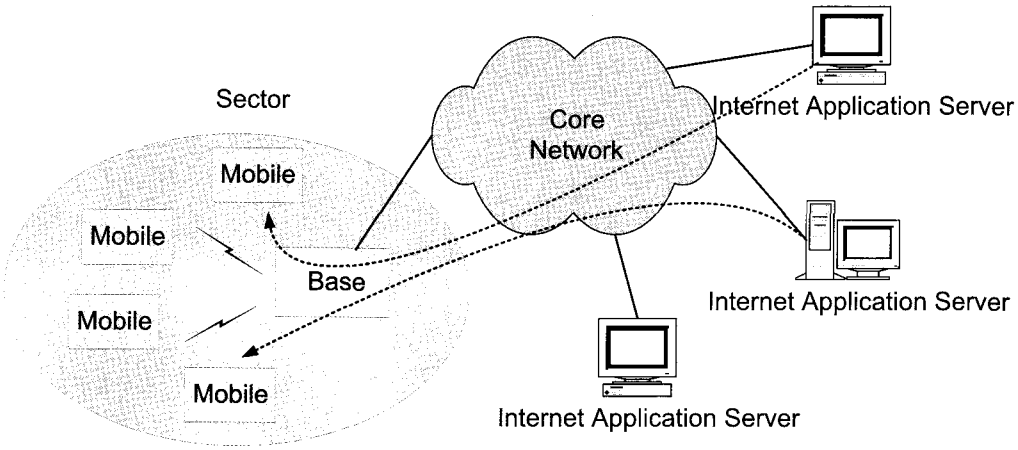


Fig. 0. Mobile Internet system model [3].

operate optimally over the wireless channel with the poor link quality or highly time-varying bandwidth. The target systems of most studies on TCP over wireless channel are WLAN. However, since some studies consider cellular environments, to address this topic herein is worthy.

TCP provides transparent segmentation and reassembly of user data. In the wired networks with highly reliable transmission medium, the segment losses (or packet losses at IP layer) occur mostly because of congestion. TCP adopts the congestion control mechanisms, e.g., the slow-start and the fast-recovery.

After sending a segment, a sender TCP starts the retransmission timer for the segment and waits for acknowledgement (ACK) from the receiver TCP. The value of a retransmission timer is updated, based on measurement of the end-to-end round trip delay. If the ACK is not received until the retransmission timer expires, the sender TCP knows the segment loss occurs (usually, due to congestion in case of wired networks). Then, the TCP retransmits the segment and sets the size of the sending window (so-called the congestion window) to one. Since the number of outstanding data segments that can be sent without receiving ACK cannot exceed the sending window size, reducing the sending window size results in the decrease of transmission rate of sender and can lessen the network congestion. This mechanism is called the slow-start.

Let us consider the case that a TCP connection is set up via wireless (radio) link. Because of high-error probability of radio channel, an exclusively long delay can occur at radio interface. The longer transmission delay of a segment on wireless link may lead to more frequent expiration of the retransmission timer in the sender TCP, which results in slow-start and decreases TCP segment transmission rate.

Many schemes for improving TCP performance in wireless environment, assuming a high bit error rate, are found in literature. They can be greatly categorized into two groups. One is to split a single TCP connection between a fixed host (Internet server) and a mobile into two separate TCP connections of wired part and wireless part, such as I-TCP [8]. The other is TCP-aware link layer schemes such as Snoop [9], where a base provides local retransmission of data on wireless link and local filtering of ACK, for preventing the server from invoking congestion control. Both I-TCP and Snoop have the advantages that they do not require changes to TCP in server and improve TCP performance. But, end-to-end semantics are violated in I-TCP, that is, ACK may be delivered to the sender TCP at server before data is delivered to the receiver TCP in mobile station. Snoop also has the disadvantage that a TCP-aware agent is needed in a base.

Strictly say, this kind of approaches were not based on the cross-layer design concept. However, the transport layer protocol is designed in consideration of wireless channel characteristics at PHY layer. Thus, we can classify this approach as a type of cross-layer protocol engineering. From the cross-layer point of view, these schemes only consider TCP and channel characteristics; they do not think over the intermediate layers and interaction between those layers. Moreover, these schemes have assumed the wireless part of TCP link provides the high error rate. However, with AMC and HARQ, the wireless link can be regarded as the time-varying rate channel, having as good quality as the wired part of TCP link. Therefore, a base may not need either an agent in Snoop or TCP module of I-TCP.

As mentioned before, a protocol entity or a scheme at a layer should contribute not to the optimality of a part of protocol layers but to the system-wide optimality. To accomplish the system-wide optimality for mobile Internet

services, the cross-layer design/engineering has to take account of all the layers together. Now, let us investigate the interactions between layers for Internet services and how they could be taken into account in design. For this purpose, we exemplify web browsing as the representative service that generates the best-effort (BE) traffic and streaming video as the representative one that generates real-time (RT) traffic.

Let us observe the downlink (from base station to mobile station) data flow in web browsing service using the Hypertext Transfer Protocol (HTTP) at application layer, for example. One can get similar insight on cross-layer design for other traffic classes.

As a typical web browsing session, an HTTP session on downlink is divided into on/off periods. These "on" and "off" periods are a result of human interaction and are referred to as the "packet call" and the "reading time," respectively [10]. A packet call represents web-page download by a user' s request, which consists of multiple packets, where the reading time identifies the time required to digest the web-page. Note that the term "packet" herein means that of HTTP protocol. One should distinguish it from an Internet Protocol (IP) packet at network layer. From the user' s point of view, the "packet call delay" is the most important performance measure since it is the delay from the request for a web page to the error-free delivery of the whole page.

The HTTP (web) server locates usually in the core (fixed) network (see Fig. 2). HTTP uses TCP as the transport layer protocol. An HTTP packet is divided into one or more TCP segments. The web traffic is transferred in the unit of TCP segment and a TCP segment again forms an IP packet. The core network delivers IP packets from server to the base station of the cell (or sector) of interest by using the IP address and proper routing algorithms. Then, the base station transmits the IP packets (TCP segments) to mobile stations through the radio interface. When a base has more than one packet that should be transmitted to one or more mobile stations, the scheduler in the base determines the transmission priority among packets.

To understand the overall interaction between HTTP, TCP, MAC/RRM, and PHY layer, let us suppose a scenario that a user suffers a deep fading and recovers from it. As the channel quality gets worse, the available data rate for the user becomes lower by AMC and HARQ mechanism. If

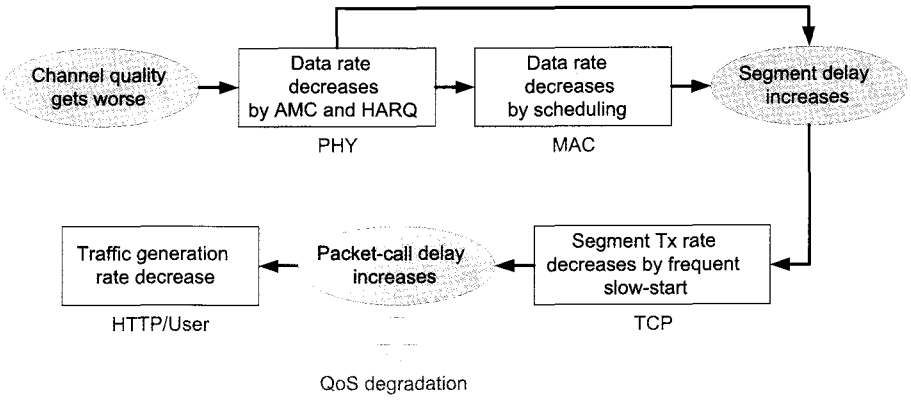


Fig. 0. Cross-layer interaction when channel quality gets worse.

there are other users competing with our user to be assigned radio resource, the situation may get much worse. A scheduler using AMC gives usually lower transmission priority to the users with worse channel quality, in order to improve system-wide spectral utilization. This scheduling policy again decreases the effective data rate of our user.

The segment transmission delay on the radio link is the sum of the time from the segment arrival at a base to the first transmission trial (that is defined as the queuing time) and the time from the first transmission to the successful reception at the destination mobile station. The former (queuing time of a segment) largely depends on the load and the scheduling scheme, whereas the latter is dependent on the wireless channel quality and retransmission scheme (including HARQ).

The worse channel quality causes the longer transmission delay of a segment on wireless link and this, in turn, results in the longer segment transmission delay at TCP layer. The long segment delay may lead to more frequent expiration of the retransmission timer in the sender TCP, which, in turn, results in frequent slow-start. Due to the slow-start, the transmission rate of sender TCP decreases. As TCP transmission rate decreases, the packet call delay at HTTP layer, which is the QoS measure for the end user at application layer, is increased. Fig. 3 shows the cross-layer interaction described above.

Accordingly, the next web page request is also delayed. Because the web page download request rate is low and TCP window size is small, the TCP transmission rate becomes low. That is, the segment arrival rate at the base becomes low.

In the mobile communications, the period that radio channel of a user is in deep fading is usually shorter than the time that is required for TCP window size to be recovered to the normal values from slow-start. Assume this is the case. Since the channel quality is recovered and becomes good, the available data rate at PHY is high and the scheduler assigns high transmission priority to the user. However, there can be no more traffic to be transmitted at the base because of low TCP transmission rate. Note that there can be still a lot of traffic to be transmitted in the server. In summary, QoS at application layer is still poor in spite of good channel condition. Fig. 4 depicts this situation.

This unfortunate situation is caused by the fact that the convergence speed in TCP window size control mechanism (i.e., slow-start and fast-recovery) is much slower than the changing rate of physical channel condition in mobile communications. The scheme(s) to solve this problem can locate at any layer(s) theoretically. However, there can be only a strict choice for us practically. The AMC and HARQ at PHY layer are applied not only to HTTP services but to all services. The modification of AMC or HARQ is not easy since it can affect all services. On the other hand, the change in TCP operation itself is also not practical since in most

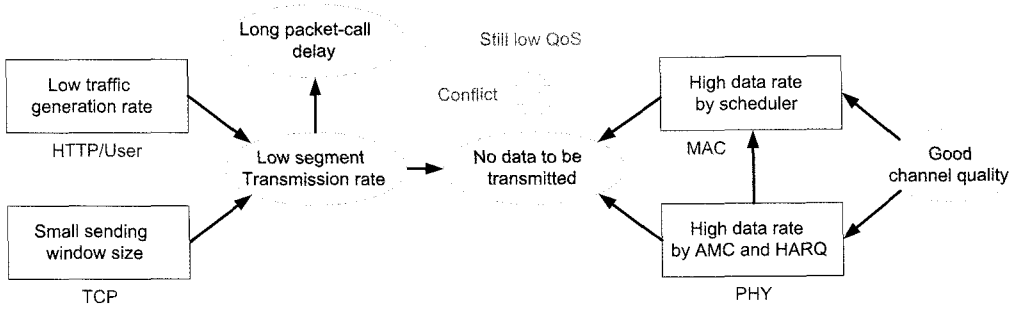


Fig. 0. Situation of problem.

cases the TCP sender is a server that serves wired users as well as mobile users. Thus, the most realistic solution seems to be the design of resource management schemes, including scheduler and (wireless) connection management mechanisms. We discuss some examples of this approach.

As mentioned above, most researches have not taken into account of all the layers but only a part of them. An exception can be found in [3]. In [3], a scheduler for high speed downlink packet transmission has been proposed. The scheduler considers all the Internet protocol layers including HTTP, TCP, scheduler, retransmission, HARQ, AMC, and fading channel model. It is designed so that its performance is maximized from the standpoint of end-to-end application level rather than from the standpoint of radio interface. To do so, the scheduler aims at avoiding excessive packet delays that may invoke the slow-start on the TCP connections.

As stated before, the transmission delay of a packet over wireless link is the sum of queuing delay at a base and the time from the first transmission trial to the successful reception at the destination mobile station. Therefore, the downlink packet delay depends on several factors including the packet arrival rate, the scheduling policy, the packet error rate, and the efficiency of retransmission scheme. For given retransmission scheme, channel environment, and load condition, if we adopt a queuing discipline (scheduling) to give higher priority to the packets that suffer the long queuing time until the first transmission trial, the overall occurrence rate of the TCP slow-start may be decreased. To avoid the long queuing delay until the first transmission trial, the scheduler proposed in [3] tries to meet a "virtual" delay constraint of packets, by best-effort.

We also examine mobile Internet systems supporting RT services as well as BE services. Suppose streaming video and web browsing as representative RT and BE services, respectively. RT and BE segments can be transported from server to the base by User Datagram Protocol (UDP) and TCP, respectively. At the base, a scheduler assigns priority to both RT and BE packets, for transmitting the packets.

Most existing packet schedulers give higher priority to the RT packets over BE packets unconditionally. However, when a lot of RT packets are waiting for transmission, the BE packets can suffer long delay with this policy, even though there is enough margin to the deadline of RT packets. The excessively long delay at radio interface may incur the slow-start in TCP layer and, as a result, degrade the system performance.

To overcome this problem, a dynamic priority scheme for streaming video is proposed in [4]. The proposed scheduler does not give higher priority to RT packets unconditionally. Instead, only the RT packets being close to the deadline are given the higher priority. The other RT packets are given the lower priority, which is the same as the priority of BE packets. With this scheme, the performance of the BE services can be improved at the small cost of RT packets, which does not become a problem in practice.

The dynamic priority scheme in [4] requires the cross-layer engineering in two-folds: the priority of RT packet depends on the deadline defined by application layer; the transmission of RT packets having the lower priority is affected by the BE traffic generation process that is again a result of interaction between layers as shown above.

We have examined the cross-layer engineering for high data-rate mobile Internet services in cellular systems. The cross-layer approach improves the radio efficiency and promotes QoS by exploiting the interactions between protocol layers. General considerations in cross-layer engineering have been addressed. Recent approach to cross-layer design has been analyzed. The main contribution

of this article is to demonstrate the reason why the cross-layer engineering taking account of all the layers, including PHY, MAC/RRM, network, transport, and application layer, is important. The discussion in this article can be used as qualitative guide-lines in designing advanced cellular systems for high data-rate mobile Internet services.

[1] Special Issue on "Cross-layer protocol engineering for wireless mobile networks: Part I," *IEEE Commun. Mag.*, Dec. 2005.

[2] Special Issue on "Cross-layer protocol engineering for wireless mobile networks: Part II," *IEEE Commun. Mag.*, Jan. 2006.

[3] W. S. Jeon, D. G. Jeong, and B. Kim, "Packet scheduler for mobile Internet access using high speed downlink packet access systems," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1789-1801, Sept. 2004. An earlier version was presented at IEEE *VTC 2002-Spring*, May 2002.

[4] W. S. Jeon and D. G. Jeong, "Combined connection admission control and packet transmission scheduling for mobile Internet services," *IEEE Trans. Veh. Technol.*, vol. 55, no. 5, pp. 1582-1593, Sept. 2006.

[5] E. Hossain, and V. K. Bhargava, "Cross-layer performance in cellular WCDMA/3G networks: Modelling and analysis," in *Proc. PIMRC 2004*, Sept. 2004.

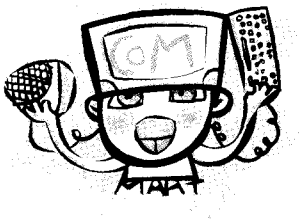
[6] L. Alonso and R. Agusti, "Automatic rate adaptation and energy-saving mechanisms based on cross-layer information for packet-switched data networks," *IEEE Commun. Mag.* pp. S15-S20, Mar. 2004.

[7] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proc. IEEE VTC2000-Spring*, Tokyo, Japan, May 2000.

[8] A. Bakre and B. Badrinath, "I-TCP: Indirect TCP for mobile host," in *Proc. 15th International Conference on Distributed Computing Systems* (ICDCS), May 1995.

[9] H. Balakrishnan, S. Seshan, E. Amir, and R. H. Katz, "Improving TCP/IP performance over wireless networks," in *Proc. ACM Mobicom95*, Nov. 1995.

[10] 3GPP2, C50-20010507-004R2, "1xEV-DV Evaluation Methodology (Rev.26)," May 2001.



정 동 근

1983년 서울대학교 공과대학 제어계측공학과 학사
1985년 서울대학교 대학원 제어계측공학과 공학석사
1993년 서울대학교 대학원 제어계측공학과 공학박사 (통신)
1986년 ~ 1990년 (주)데이콤 정보통신연구소 주임연구원
1994년 ~ 1997년 (주)신세기통신 기술연구소 책임연구원
1997년 ~ 현재 한국외국어대학교 전자정보공학부 교수
관심분야 : 무선/이동통신