Analysis of Bandwidth Gain Over Various Timer_CU of AAL2 for Voice Traffic Multiplexing

Hyun-Jin Lee and Jae-Hyun Kim, *Member, IEEE*

***Abstract*—An asynchronous transfer mode adaptation layer type 2 (AAL2) transmission scheme commonly is used to deliver the voice and tha data traffic between Node-B and the radio network controller on the universal mobile terrestrial device network. To predict the AAL2 multiplexing performance, we analyzed the bandwidth gain and the cell-packing density using discrete Markov chain model for the voice service and validated these results with simulations. We also performed a detailed simulation for the voice and the data services in a concentrator. Based on the analysis, we proposed an engineering guideline for selecting the optimal Timer_CU in a Node-B. We found that there is no major benefit in using the AAL2 multiplexing in a concentrator. The benefit of the AAL2 multiplexing in I_{ub} for the data service was much less than that for the voice service. They also depended heavily on the traffic load.**

***Index Terms*—ATM, multiplexing, UMTS.**

I. Introduction

ASYNCHRONOUS TRANSFER MODE (ATM) has been chosen as the transmission technology to be used in the universal mobile terrestrial system (UMTS) terrestrial radio access network (UTRAN). UTRAN is defined by the Third Generation Partnership Project (3GPP) because of its ubiquitous nature for the heterogeneous traffic types, quality of service (QoS) guarantee, and its widespread deployment in public networks [1]. However, applying ATM to a low data rate mobile voice stream is inefficient because it generates the extra traffic to fill out the payload of an ATM cell. Recognizing the problems, ITU-T standardized the ATM adaptation layer type 2 (AAL2) to efficiently transmit the application that is sensitive to delays. In the 3GPP, an ATM/AAL2 transmission scheme on the I_{ub} interface between Node-B and the radio network controller (RNC) is used to deliver the voice and the data traffic. In the I_{ub} interface, voice and data frame packets with a small size are transmitted according to the short frame length of 10, 20, 40, or 80 ms. Therefore, an AAL2 transmission scheme that is suitable for small packets has been chosen in the 3GPP to increase the efficiency of the bandwidth. The AAL2 protocol suite is mandatory in the first and second release of UMTS and may be applied to both the access and the core network. Hence, the performance of AAL2 multiplexing is one of the most important topics in UTRAN architecture engineering. Several papers have analyzed it by computer simulations or by simple experiments [2]–[5].

Manuscript received April 8, 2004; revised September 17, 2004 and December 18, 2004. This work is supported in part by IITA project. The material in this paper was presented in part at ICOIN2004, Busan, 2004. The review of this paper was coordinated by Prof. D. O. Wu.

The authors are with the Department of Electrical Engineering, Ajou University, Suwon, 443-749, Korea (e-mail: l33hyun@ajou.ac.kr; jkim@ajou.ac.kr).

Digital Object Identifier 10.1109/TVT.2005.851321

Many studies on AAL2 multiplexing showed that the gain obtained by AAL2 multiplexing is significant in terms of the bandwidth [4], [5]. They also pointed out the importance of selecting Timer_CU value because it significantly affects the link efficiency. However, most of these papers did not focus on UMTS network and did not consider UMTS-specific protocol behaviors that are used to evaluate the AAL2 multiplexing efficiency for both the voice and the data traffic. A recent paper [4] included some protocols such as radio link control (RLC) and frame protocol (FP) that has impact on the throughput in I_{ub}. But this study used an approximated protocol overhead and did not consider RLC retransmission and source controlled rate (SCR) operation. RLC retransmission and SCR operation may affect the voice traffic behaviors because RLC retransmission may increase the traffic load, whereas SCR may decrease the traffic load and the average packet size. SCR also may result in reducing the bandwidth utilization for the same traffic load and increasing the cell-packing density. The cell-packing density means the ratio of the average user payload to a cell except ATM overhead. RLC retransmission attempts to recover the corrupted blocks within the air interface before a recovery mechanism from an upper layer protocol such as TCP is activated. The RLC retransmission model has an important role of changing the traffic model within the I_{ub}.

We consider AAL2 multiplexing in a Node-B and a concentrator on the I_{ub}. To evaluate the performance of AAL2 multiplexing, we derive the cell-packing density and the bandwidth gain by using discrete-time Markov chain model and perform detailed simulations, which support the investigation of all connection-, cell-, and bit-level aspects of the UMTS network for the voice and the data services in UTRAN [6]. We use the UMTS-specific traffic model, such as the adaptive multirate (AMR) codec with the SCR feature (UMTS mandatory features) for the voice traffic, and use HTTP1.1 protocol for the web traffic. In addition to the application layer protocol, various UMTS-specific network protocols such as RLC and FP are modeled to evaluate the AAL2-multiplexing effect more precisely.

This paper is organized as follows. Section II introduces the model description. Numerical analysis and simulation parameters are described in Section III. Section IV provides the numerical and simulation results. Finally, we conclude the paper in Section V.

II. Model Description

This section describes a user-plane model of an end-to-end reference connection through a UMTS network. The simulator models all protocol layers from the physical through the

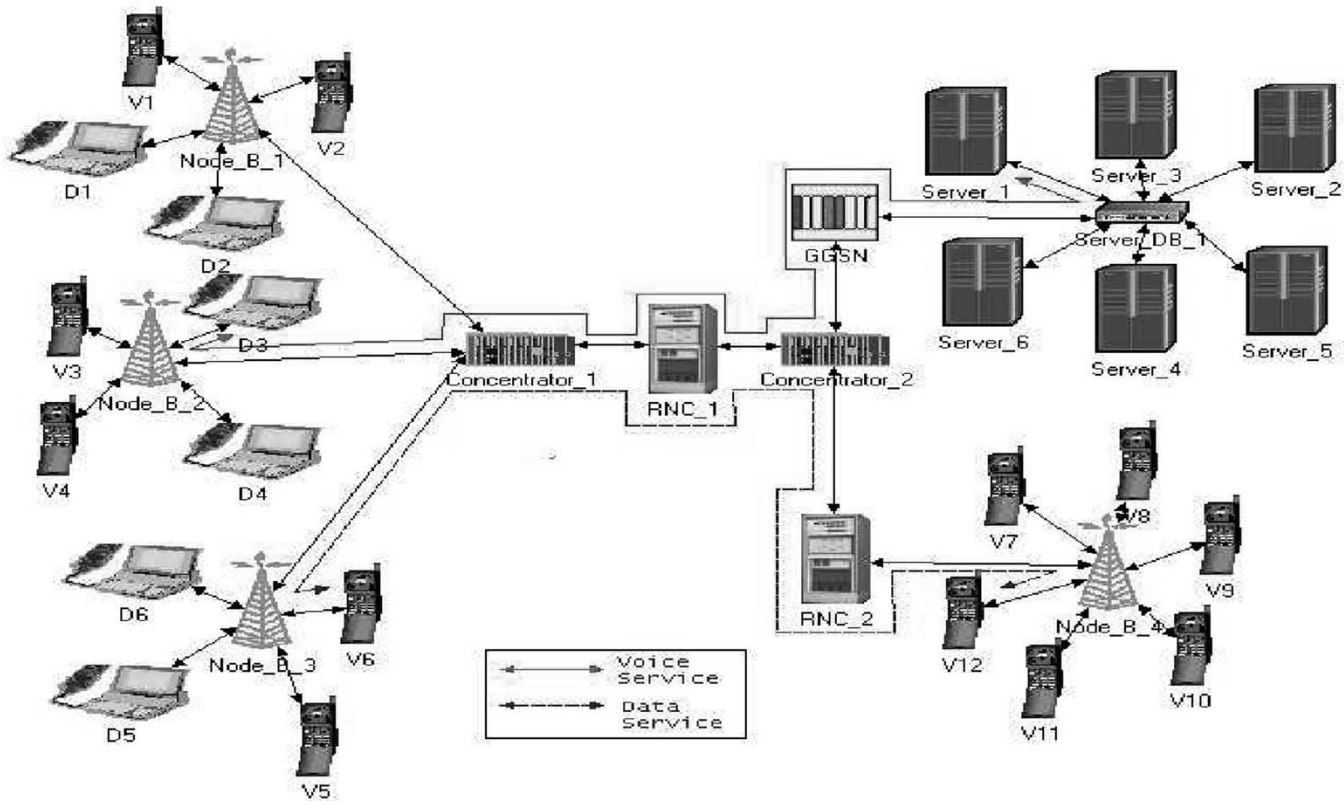

Fig. 1. UMTS network architecture model that supports the voice and data application traffic.

application layer and models details of the packet-handling characteristics of each network element along the path. The system model predicts application-level performance metrics such as response time, packet loss, jitter, delay, and throughput. Refer to [6] for details of the simulator. The reference architecture and connections are based on 3GPP UMTS release 99 standards and the UMTS application models are based on a combination of standards and the published traffic characteristics [7], [8]. A mobile-to-mobile reference connection is considered for the voice traffic model and client-server models are used for the web browsing.

A. UMTS Network Architecture

Fig. 1 shows the network model used in this study. The voice traffic and web-browsing data traffic are offered to the UMTS network, which consists of three Node-Bs, one concentrator, one RNC in the UTRAN, and other core network elements including web servers and voice-called parties. The Node-B-to-concentrator link has a capacity of one E1 (2.048 Mb/s) and the concentrator-to-RNC link capacity is STM-1 (155.520 Mb/s).

B. UMTS Protocol Stack Models for CS and PS Services

Fig. 2 shows the protocol stack modeled in the simulator for the packet switched (PS) service and the circuit-switched (CS) service within UTRAN with AAL2 multiplexing along the I_{ub} interface. The left column in the shaded box is for the PS service and the right column is for the CS service. The nonshaded boxes are for both the PS and the CS services. Fig. 2 does not show I_{ur} interface based on the assumption that the controlling RNC (CRNC) and serving RNC (SRNC) are coincident. The air interface is not modeled explicitly because it would slow down the simulation time too much. Instead, a separate model was used to generate the traced files of block error rate (BLER) under various conditions. This file was fed into the RLC layer for RLC acknowledge-mode operation to be modeled correctly. The detailed AAL2, AMR, RLC, and dedicated channel FP (dchFP) models are described as follows.

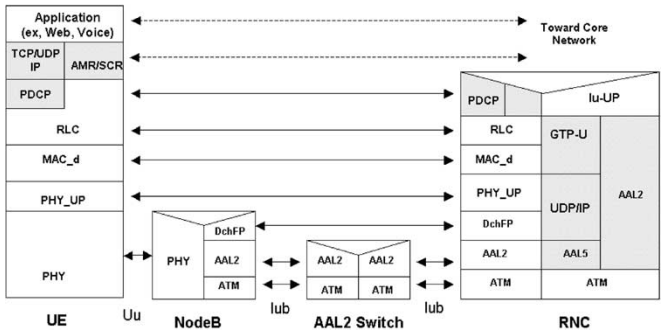

Fig. 2. Protocol stack model for PS and CS service bearer traffic.

1) AAL2 Model: The structure of the AAL2 layer consists of two sublayers, which are called the common part sublayer (CPS) and the service-specific conversion sublayer (SSCS) [9]. The CPS provides multiplexing and demultiplexing of CPS packets over a single ATM virtual channel connection (VCC). Each AAL2 user generates CPS packets with a 3-byte packet header and a variable length payload. AAL2 uses the 8-bit channel ID

(CID) in the CPS packet header to multiplex AAL2 users onto a single VCC. Because of a limited CID size and some reserved values, only 248 individual connections can be differentiated within a single VCC. The CPS sublayer collects CPS packets from the AAL2 users multiplexed onto the same VCC over a specified interval (Timer_CU). If the cell is not completely packed within the period determined by the Timer_CU value, the timer expires and the partially packed cell will be sent. The CPS protocol data unit (CPS-PDU) employs a one-byte start field (STF) followed by 47 bytes payload.

2) AMR Codec Model: AMR codec is the most important vocoder that is a mandatory speech-processing function in UMTS. SCR functionality, which is the part of the standard, is called "discontinuous transmission" in GSM [10], [11]. The AMR codec uses eight source codecs with bit rates of 12.2, 10.2, 7.95, 7.40, 6.70, 5.90, 5.15, or 4.75 kb/s and the coder operates on speech frames of 20 ms corresponding to 160 samples at the sampling frequency of 8000 sample/s. In this paper we consider only one of the specified rates during the ON state (talkspurt) of the AMR codec (for 12.2 kb/s). We also consider the comfort noise during the OFF state (silence) of the AMR codec.

3) RLC Model: The RLC protocol provides segmentation and retransmission services for both user and control data. Each RLC instance is configured by radio resource control (RRC) to operate in one of three modes: transparent mode (TrM), unacknowledged mode (UM), and acknowledged mode (AM). TrM and AM are used for the user plane. UM is not considered in this study because this mode is mainly used for RRC signaling and VoIP. Therefore these two modes are described in this section as modeled in the simulator (see [12] for the detail on the RLC protocol).

TrM RLC entities are defined to be unidirectional mode, whereas AM entities are described as bidirectional mode. In the TrM no protocol overhead is added to the higher layer data. The packet of the streaming type that is transmitted from the higher layer is not segmented. TrM can be used for the CS service such as voice call. In the AM, an automatic repeat request (ARQ) mechanism is used for the error correction. The AM is the normal RLC mode for the PS service, such as web browsing and email downloading.

Fig. 3 shows a simplified block diagram of an AM RLC entity. Fig. 3 shows only how an acknowledge mode data (AMD)-protocol data unit (PDU) can be constructed in the simulator. RLC service data units (SDUs) received from higher layers via AM service access point (SAP) are segmented and/or concatenated to payload units (PU) of fixed length. For concatenation or padding purposes, bits carrying information on the length and extension are inserted into the beginning of the last PU where data from an SDU are included. If several SDUs fit into one PU, they are concatenated and the appropriate length indicators are inserted into the beginning of the PU. An RLC AMD PDU is constructed by taking one PU from the transmission buffer and adding a header to it. The receiving side of the AM entity receives RLC AMD PDUs through one of the logical channels from the MAC sublayer. If the received PDU was a control message or if the status information was piggybacked to an AMD PDU, the control information (STATUS message) is passed to the transmitting side, which will check its retransmission buffer against the received status information. Once all PUs belonging to a complete SDU are in the receiving buffer, the SDU is reassembled. After this, the checks for sequence delivery and duplicate detection are performed before the RLC SDU is delivered to the higher layer.

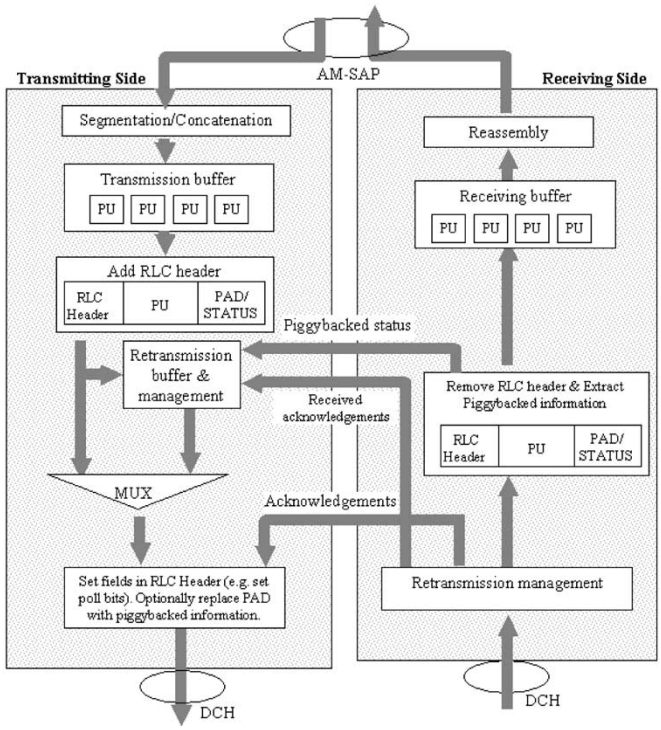

Fig. 3. A simplified block diagram of an RLC AM entity.

4) DchFP Model: The purpose of the DchFP is to transport the transport blocks transparently between Node-B and SRNC. This protocol allows that the coordinated dedicated transport channel is multiplexed onto one transport bearer with the same transmission time interval (TTI). The blocks transported from all the coordinated dedicated channels (DCHs) for TTI are included in one frame [13]. There are two types of frames for DchFP (indicated by the frame type field): DCH data frame and DCH control frame. We modeled the DCH data frame because the user plane is the interest in this paper. See [13] for the detail on the structure of the uplink and downlink data frame.

C. Application Service Traffic Model

It is assumed that each user traffic related to the radio bearers over I_{ub} will be carried within FP. The user traffic is defined as each of the physical channel types over radio. To consider the traffic flow over I_{ub}, it is necessary to consider the various protocol overheads over I_{ub} interface. It is also assumed that all radio bearers are carried on dedicated physical data channels (DPDCH) using DchFP over I_{ub}. To determine the bandwidth efficiency, voice and web-browsing traffic models are used.

1) Voice Traffic Model: The voice traffic is generated by two hierarchical structures as shown in Fig. 4(a): call and packet level. The call-level model is composed of a sequence of ON

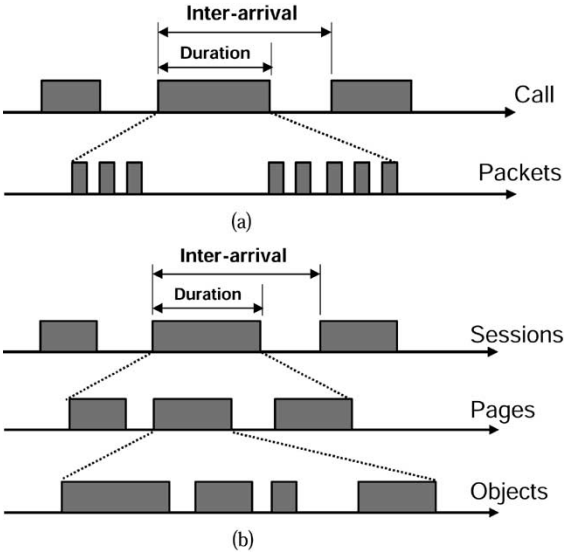

Fig. 4. Voice and data service traffic model.

TABLE I
VOICE APPLICATION TRAFFIC MODEL PARAMETER

Category	Value
packet size	talkspurt, 244 bit, silence, 39 bits
mean call arrival rate in busy hours per user	exponential, 0.6 call/h
call session holding times	exponential, 100
uplink/downlink activity factor	0.5

and OFF periods. Packet levels generate a pattern of talkspurt and silence states by means of a voice activity detector (VAD). Therefore, the call level and the packet level can be modeled as a two-state Markov chain. The voice packets are generated at 12.2 kb/s for the talkspurt state. The comfort-noise packets are generated for the silence state. It has been found that the length of the talkspurt and the silence states is exponentially distributed [7]. Each state has the mean of 3 s according to [14]. And the application packet length of each state is 244 bits and 39 bits, respectively. It is added to the overhead when the packet goes through the lower layer. Finally, the packet length, including all the overhead, is 42 bytes and 15 bytes at AAL2 layer. Table I presents the voice application traffic model in detail.

2) Web Browsing Traffic Model: We hierarchically model the web browsing traffic consisting of session, page, and object, which is presented in Fig. 4(b). The distributions of each of the parameters for the web browsing traffic model are determined in [8] and an application session is divided into ON/OFF periods representing web page downloads and the intermediate reading times. These ON and OFF periods are a result of human interaction, i.e., these periods are represented by the user's request information and the reading time identifies the time required to digest the web page. The initial HTML page is referred to the "main page" and each of the constituent objects referenced from the main page is referred to as an "embedded object." The web traffic will depend on the version of HTTP used by the web browsers and servers. HTTP 1.1, one of the

TABLE II
WEB BROWSING TRAFFIC MODEL PARAMETER

Active Data Session		
session interarrival time distribution		exponential
session duration		531
number of subsessions page		log normal distribution, mean = 17 standard deviation = 22
reading time		exponential distribution, mean = 30 s; $\lambda = 0.033$ s
Objects		
Link Direction	Downlink	Uplink
number of packets	pareto distribution, mean (43)	Pareto distribution, mean (22)
packet size (bytes)	576 (24%) 1500 (76%)	40 (85%) 360 (15%)
interarrival time of packets	geometric distribution, mean (71 ms)	geometric distribution, mean (153 ms)

widely used protocols, is used in the web browsing simulation. In HTTP 1.1, persistent TCP connections are used to download the objects, which are located at the same server, and the objects are transferred serially over a single TCP connection. The TCP overhead of the slow-start and congestion control occur only once per persistent connection. We model the arrival of sessions, the characteristics of the arrival of page requests within a session, and the number of objects. Their sizes for each page are summarized in Table II. It is assumed that maximum data rate is 64 kb/s for the uplink and 144 kb/s for the downlink. We do not assume compression on user-plane TCP/IP header for data packets.

III. PERFORMANCE ANALYSIS AND SIMULATION

A. Packing Density and Bandwidth Gain Analysis

To understand how many AAL2 packet bytes are packed into an AAL2 packet, we derive the cell-packing density by

$$\psi(\%) = \frac{B(\tau, N) + 1}{48} \times 100, \quad (1 \le B \le 47) \tag{1}$$

where N equals total voice users, $B(\tau, N)$ means the average user payload which is multiplexed during Timer_CU (τ). In this paper we defined the user payload as only CPS-PDU without STF at ATM layer. The average user payload in an ATM cell depends on whether there is a remainder from the previous cell, how many cells are totally transmitted, and how many of them are full of payload. Therefore, the user payload can be modeled by using a Markov chain. Each state of Markov chain is presented by the remainder of a cell. L_{talk} and L_{silence} denote the number of packet in the talkspurt and the the silence states. The average number of packet is given by

$$E[L_{\mathrm{talk}}] = \frac{E[T_{\mathrm{talk}}] + H}{\mathrm{TTI}} \tag{2}$$

$$E[L_{\mathrm{silence}}] = \frac{E[T_{\mathrm{silence}}] - H}{\mathrm{TTI}} \cdot \frac{1}{8} \tag{3}$$

where TTI means the transmission time interval, T_{talk} and T_{silence} are the talkspurt and the silence duration, respectively. To ensure the correct estimation of the comfort-noise parameter

at a receiver, the first seven frames (140 ms) after the reset or after enabling the SCR operation will always be regarded as the talkspurt packets, even if VAD decide silence packets [11]. This is a hangover procedure. We assumed that the talkspurt packets are generated during the hangover period ($H = 140$ ms). In our analysis model, every eighth packet includes frame information and the others do not contain any information during the silence state [15]. Therefore we divide the number of the silence packet by 8 in (3). Let P_{talk} and P_{silence} be the probabilities of packet arrivals during Timer_CU (τ) for user equipment (UE) in the talkspurt and the silence states respectively. The probability of a packet arrival from a user is given by

$$P_{\text{talk}} = \frac{E[L_{\text{talk}}]\tau}{E[T_{\text{talk}}] + E[T_{\text{silence}}]}, \tag{4}$$

$$P_{\text{silence}} = \frac{E[L_{\text{silence}}]\tau}{E[T_{\text{talk}}] + E[T_{\text{silence}}]}. \tag{5}$$

Voice sources are assumed to be independent. After the receipt of the first packet in a cell, the probability of no arrivals (p_{00}) from all the other sources within τ can be derived in

$$\begin{aligned} p_{00} &\triangleq P[N_{\text{silence}} = 0, N_{\text{talk}} = 0] \\ &= (1 - P_{\text{talk}} - P_{\text{silence}})^{N-1} \end{aligned} \tag{6}$$

where N_{talk} and N_{silence} are the number of concurrent users for talkspurt and silence state, respectively. The probability of n packet arrivals from all the other sources in τ after one packet is arrived is given by

$$p = \binom{N-1}{n} (1 - P_{\text{talk}} - P_{\text{silence}})^{N-n-1} \times (P_{\text{talk}} + P_{\text{silence}})^{n}. \tag{7}$$

Given that n packets arrive, the probability of the talkspurt packets to be i and the silence packets to be j is given by

$$\begin{aligned} p_{ij} &= P[N_{\text{talk}} = i, N_{\text{silence}} = j] \\ &= \binom{N-1}{n} \binom{n}{i} (1 - P_{\text{talk}} - P_{\text{silence}})^{N-n-1} \\ &\quad \times (P_{\text{talk}})^{i} (P_{\text{silence}})^{j}, \quad i + j = n. \end{aligned} \tag{8}$$

The remainder that can be carried over to the next ATM cell varies from 0 to 41 because the maximum packet size is 42 bytes (talkspurt packet). If r_k is the remainder bytes carried over from the $(k-1)$th cell, the event of $r_k = 0$ can occur for following cases.

1) The Timer_CU in the $(k-1)$th cell expires.
2) $r_{k-1} = 2$, and 3 silence packets are arrived in the $(k-1)$th cell.
3) $r_{k-1} = 5$, and 1 talkspurt packet is arrived in the $(k-1)$th cell.
4) $r_{k-1} = 17$, and 2 silence packets are arrived in the $(k-1)$th cell.
5) $r_{k-1} = 32$, and 1 silence packet is arrived in the $(k-1)$th cell.

r_k $(1 \le r_k \le 40)$ can be calculated by iteration. The event $r_k = 41$ happens only if $r_{k-1} = 4$ and 2 talkspurt packets are arrived. r_k can be modeled by using the state of Markov chain. Considering the stationary state where all $\{r_k\}$ have the same probability distribution, and let r denote the random variable for the remainder length. The stationary state probability is given by

$$\pi_i = P[r = i], \quad i = 0, \ldots, 41. \tag{9}$$

To calculate stationary state probability, we need to obtain a transition probability matrix ($\mathbf{T}$). Because $\mathbf{T}$ is composed of p_{ij} and the probability that Timer_CU expires, we need to find the probability that Timer_CU expires. We define the probability that Timer_CU expires as follows:

$$\begin{aligned} Q_j &\triangleq P[\text{Timer\_CU expires} \,|\, r = j] \\ &= \begin{cases} p_{00} + p_{01} + p_{02} + p_{03} + p_{10}, & 0 \le j \le 2 \\ p_{00} + p_{01} + p_{02} + p_{10}, & 3 \le j \le 5 \\ p_{00} + p_{01} + p_{02}, & 6 \le j \le 17 \\ p_{00} + p_{01}, & 18 \le j \le 32 \\ p_{00}, & 33 \le j \le 41 \end{cases} \end{aligned} \tag{10}$$

And we define the normalization coefficient (C_j) to have the irreducible property as follows.

$$C_j = \begin{cases} p_{04} + p_{11} + p_{20}, & 0 \le j \le 1 \\ p_{03} + p_{11} + p_{20}, & 2 \le j \le 4 \\ p_{03} + p_{10}, & 5 \le j \le 16 \\ p_{02} + p_{10}, & 17 \le j \le 31 \\ p_{01} + p_{10}, & 32 \le j \le 41 \end{cases} \tag{11}$$

From (8), (10), (11) and the cases of r_k, we find $\mathbf{T}$ in (12). (Please see the equation at the bottom of the page.) The stationary state

$$\mathbf{T} = \begin{pmatrix} Q_0 & 0 & 0 & 0 & 0 & \cdots \\ Q_1 & 0 & 0 & 0 & 0 & \cdots \\ Q_2 + \frac{(1-Q_2)p_{03}}{C_2} & 0 & 0 & 0 & 0 & \cdots \\ Q_3 & \frac{(1-Q_3)p_{03}}{C_3} & 0 & 0 & 0 & \cdots \\ Q_4 & 0 & \frac{(1-Q_4)p_{03}}{C_4} & 0 & 0 & \cdots \\ Q_5 + \frac{(1-Q_5)p_{10}}{C_5} & 0 & 0 & \frac{(1-Q_5)p_{03}}{C_5} & 0 & \cdots \\ Q_6 & \frac{(1-Q_6)p_{10}}{C_6} & 0 & 0 & \frac{(1-Q_6)p_{03}}{C_6} & \cdots \\ Q_7 & 0 & \frac{(1-Q_7)p_{10}}{C_7} & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{12}$$

probability is uniquely determined through (13) because $\mathbf{T}$ has an irreducible and a periodic homogeneous ergodic property.

$$\sum_{i=0}^{41} \pi_i = 1$$

$$\pi_i = \sum_{j=0}^{41} \pi_j t_{ji} \tag{13}$$

where t_{ji} means the elements of $\mathbf{T}$. The solution is obtained by simultaneously solving the two equations that are given by (13) and (14). (Please see the equation at the bottom of the page.) (14) represents the probability of a remainder of the kth cell according to the probability of a remainder of the previous cell.

Using $\{\pi\}$ and p_{ij}, the average of user payload bytes in an ATM cell is given by

$$\begin{aligned} B(\tau, N) = {} & p_{00} \sum_{i=0}^{41} i\pi_i + p_{01} \sum_{i=0}^{32} (i+15)\pi_i \\ & + p_{02} \sum_{i=0}^{17} (i+30)\pi_i + p_{03} \sum_{i=0}^{2} (i+45)\pi_i \\ & + p_{10} \sum_{i=0}^{5} (i+42)\pi_i, \quad \tau > 0 \\ & + 47(1 - p_{00} - p_{01} - p_{02} - p_{03} - p_{10}). \end{aligned} \tag{15}$$

We define the bandwidth gain $[\xi(\tau, N)]$ as the ratio of the difference between the ingress traffic load and the egress traffic load to the ingress traffic load as follows:

$$\begin{aligned} \xi(\tau, N) &= \frac{\sum_{\text{all}} \lambda_{\text{ingress}} - \lambda_{\text{egress}}}{\sum_{\text{all}} \lambda_{\text{ingress}}} \times 100 \\ &= \frac{B(\tau, N) - B(0, N)}{B(0, N) + 6} \times 100 \end{aligned} \tag{16}$$

where λ_{ingress} means the ingress traffic load from previous network elements and λ_{egress} means the egress load to the next network element measured in the link. In (16), 6 is included due to the existence of ATM cell overhead and STF. $B(0, N)$ is the average user payload without using AAL2 multiplexing as follows.

$$B(0, N) = \frac{15 \times E[L_{\text{silence}}] + 42 \times E[L_{\text{talk}}]}{E[L_{\text{silence}}] + E[L_{\text{talk}}]} = 39.2 \tag{17}$$

In this paper we derive the bandwidth gain at a Node-B and a concentrator using the numerical analysis and the simulation.

TABLE III
SIMULATION PARAMETER

Category	Voice Traffic	Web Traffic
monitoring	uplink	both uplink and downlink
session	248	
TTI (ms)	20	40
BLER (%)	1	uplink: 4 downlink: 5
the maximum retransmission in RLC	no use	3
HTTP version	v1.1	
TCP parameters	windows 2000-based parameters	
window size	16 kb	
TCP header compression	no	

B. Simulation Scenarios and Parameters

It is assumed that each of the simulation scenarios contains only voice or web browsing traffic. Each traffic generator can generate multiple sessions of voice or web browsing traffic [16]. To evaluate AAL2 multiplexing effectiveness at the Node-B (the concentrator) in $\mathrm{I_{ub}}$, the bandwidth gain and the cell-packing density at the UE-to-Node-B link (Node-B-to-concentrator link) and the Node-B-to-concentrator link (concentrator-to-RNC link) have been measured. Table III presents the simulation parameters used in this paper. In the evaluation, only the uplink traffic is measured for the voice traffic because the uplink and the downlink traffic for the voice would be similar, and for web browsing traffic both the uplink and the downlink are measured because of the asymmetric traffic behavior. In the uplink, the packet size would be smaller and there would be less traffic compared with the downlink. The maximum number of application sessions generated is the "$\mathrm{Max}\{\text{number of sessions generate traffic} = \mathrm{E1\,link}, 248\,\text{sessions}\}$." The first Node-B generates the traffic session until the E1 link between itself and the adjacent concentrator reaches the link capacity. Then, more application sessions are added to the next Node-B until the total number of application sessions is 248. The maximum number of application sessions from the concentrator to the RNC is 248 in every simulation scenario. The transmission interval is set to be 20 ms for the voice session and 40 ms for the web browsing session. The 3GPP defined that the maximum allowable BLER for the voice should be less than 10^{-2} and for the data should be less than 10^{-1}. We use 1% BLER for the voice and 4% BLER for the 64 kb/s uplink, and 5% BLER for the 144 kb/s downlink for data traffic. The maximum allowable retransmission to recover the block error between RLC layers in the UTRAN is limited

$$\pi_i = \begin{cases} \sum_{j=0}^{41} Q_j \pi_j + \frac{(1-Q_2)p_{03}\pi_2}{C_2} + \frac{(1-Q_5)p_{10}\pi_5}{C_5} + \frac{(1-Q_{17})p_{02}\pi_{17}}{C_{17}} + \frac{(1-Q_{32})p_{01}\pi_{32}}{C_{32}}, & i = 0 \\ \frac{(1-Q_{i+2})p_{03}\pi_{i+2}}{C_{i+2}} + \frac{(1-Q_{i+5})p_{10}\pi_{i+5}}{C_{i+5}} + \frac{(1-Q_{i+17})p_{02}\pi_{i+17}}{C_{i+17}} + \frac{(1-Q_{i+32})p_{01}\pi_{i+32}}{C_{i+32}}, & 1 \le i \le 9 \\ \frac{(1-Q_{i+2})p_{03}\pi_{i+2}}{C_{i+2}} + \frac{(1-Q_{i+5})p_{10}\pi_{i+5}}{C_{i+5}} + \frac{(1-Q_{i+17})p_{02}\pi_{i+17}}{C_{i+17}} + \frac{(1-Q_{i-10})p_{11}\pi_{i-10}}{C_{i-10}}, & 10 \le i \le 12 \\ \frac{(1-Q_{i+2})p_{03}\pi_{i+2}}{C_{i+2}} + \frac{(1-Q_{i+5})p_{10}\pi_{i+5}}{C_{i+5}} + \frac{(1-Q_{i+17})p_{02}\pi_{i+17}}{C_{i+17}} + \frac{(1-Q_{i-10})p_{11}\pi_{i-10}}{C_{i-10}} + \frac{(1-Q_{i-13})p_{04}\pi_{i-13}}{C_{i-13}}, & 13 \le i \le 14 \\ \frac{(1-Q_{i+5})p_{10}\pi_{i+5}}{C_{i+5}}, & 15 \le i \le 36 \\ \frac{(1-Q_{i-37})p_{20}\pi_{i-37}}{C_{i-37}}, & 37 \le i \le 41 \end{cases} \tag{14}$$

to three. Those packets that could not be recovered by the RLC layer rely on the recovery mechanism in TCP protocol.

IV. Performance Analysis Results

To evaluate the bandwidth gain in a Node-B and in a concentrator, a set of simulations has been performed with various Timer_CU values and the number of concurrent users for the voice and web browsing traffic.

A. Voice Traffic Scenarios

Figs. 5 through 8 represent the cell-packing density and the bandwidth gain results for the various Timer_CU values and the number of concurrent users for the voice traffic. In Figs. 5 and 6, we present the bandwidth gain and the cell-packing density in a Node-B using mathematical analysis and simulations. The curves mean analytic results and the symbols represent simulation results. It can be seen that the analytic results are closed to the simulation ones. In this scenario, we set up one Node-B and vary the number of concurrent voice users from 1 to 170 because the capacity of Node-B-to-concentrator link is restricted by E1 capacity. Timer_CU in the Node-B is set to be 0 through 4 ms.

Fig. 5 presents the cell-packing density versus the number of concurrent voice users and Timer_CU. Based on the analytic results, the minimum cell-packing density, which is obtained without AAL2 multiplexing ($\mathrm{Timer\_CU} = 0$ ms), is about 83.5%. The cell-packing density reaches to 96% when the number of concurrent voice users is 80 with $\mathrm{Timer\_CU} = 1$ ms. Meanwhile, the cell-packing density reaches same value (96%) with only 40 concurrent voice users and $\mathrm{Timer\_CU} = 2$ ms. It can be seen from Fig. 5 that the cell-packing density is strongly affected by Timer_CU value. Fig. 6 shows the average bandwidth gain of AAL2 multiplexing in a Node-B. The result indicates that the maximum bandwidth gain with AAL2 is about 18% higher than the bandwidth gain without AAL2. If the link type between the Node-B and the concentrator is E1, 170 concurrent voice users can be connected simultaneously in the Node-B with a $\mathrm{Timer\_CU} = 0$ ms. By setting $\mathrm{Timer\_CU} = 1$ ms, however, more than 200 concurrent voice users can be served in a Node-B without any other system or network changes. This gives a strong reason to use AAL2 multiplexing in a Node-B.

Using the mathematical analysis, Fig. 7 shows the Timer_CU value required to get the specific bandwidth gain according to the number of concurrent voice users. When the number of concurrent voice users is small, the Timer_CU value, at which the given bandwidth gain is achieved, is decreased rapidly. But it is closed to asymptotic value over 80 concurrent voice users. So we propose the engineering guideline that can be used to select the optimal Timer_CU based on Fig. 7.

Fig. 8 shows the bandwidth gain in a concentrator with various numbers of concurrent voice users by using only the simulation results. We changed the number of concurrent voice users in a Node-B from 10 to 170. As you can see in the graph, increasing traffic load per Node-B from 15% to 33% (when Timer_CU value is 2 ms) results in the drastic drop of the bandwidth gain from 11% to 3%. This result indicates that there is no significant

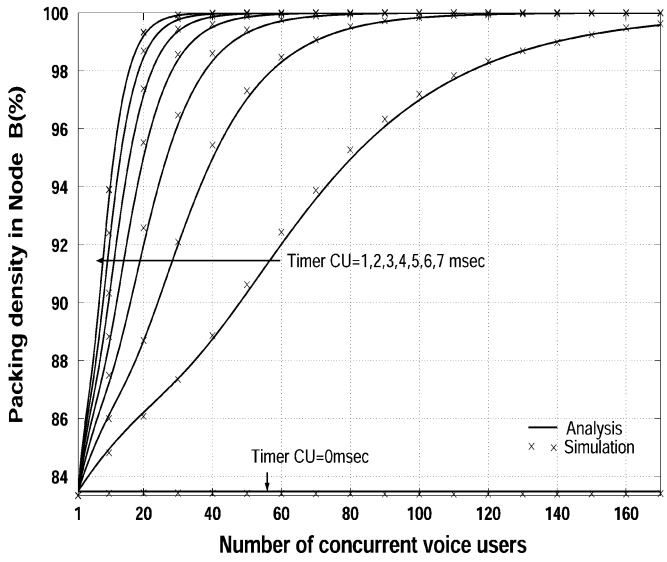

Fig. 5. The cell-packing density with various Timer_CU and number of concurrent voice users by using simulation and mathematical analysis.

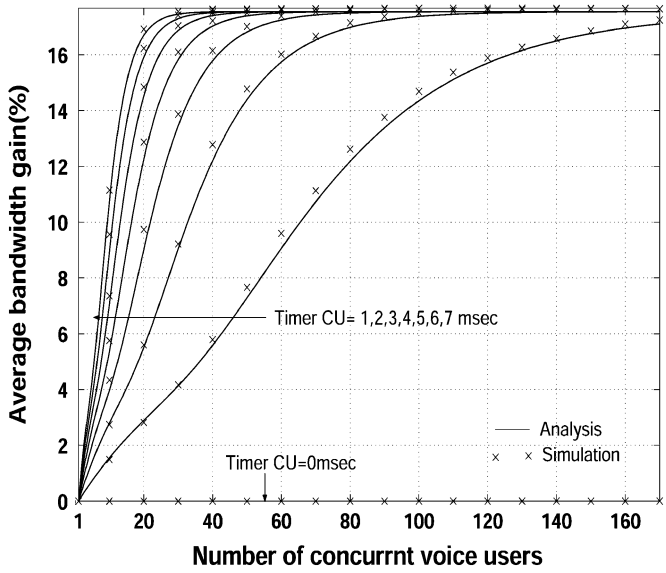

Fig. 6. Effectiveness of AAL2 multiplexing on bandwidth gain with various Timer_CU values.

AAL2 multiplexing benefit in a concentrator on $\mathrm{I_{ub}}$ in terms of the bandwidth gain. This is a reason that the bandwidth gain is not meaningful for small traffic load and 3% bandwidth gain is negligible at heavy traffic load in a concentrator.

B. Data (Web Browsing) Traffic Scenario

Figs. 9 and 10 show the cell packing density and the link utilization results with various Timer_CU values and the number of concurrent users for the data traffic. These two figures represent the uplink traffic only. The traffic load generated by 40 simultaneous web-browsing sessions is about 600 kb/s (28% of E1) at the Node-B-to-concentrator link after AAL2 multiplexing with Timer_CU of 1 msec at each Node-B. We use 4% BLER for the wireless channel. It causes to retransmit a page request packet so

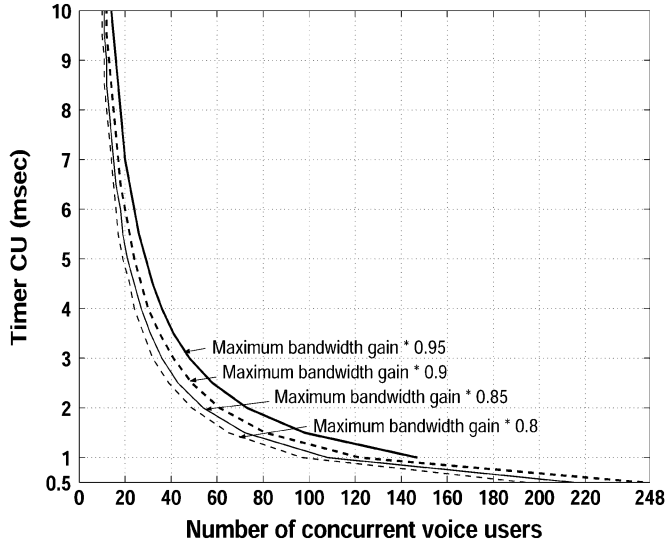

Fig. 7. Timer_CU to reach specific value (from 80% to 95% of the maximum bandwidth gain) vs. number of concurrent voice users.

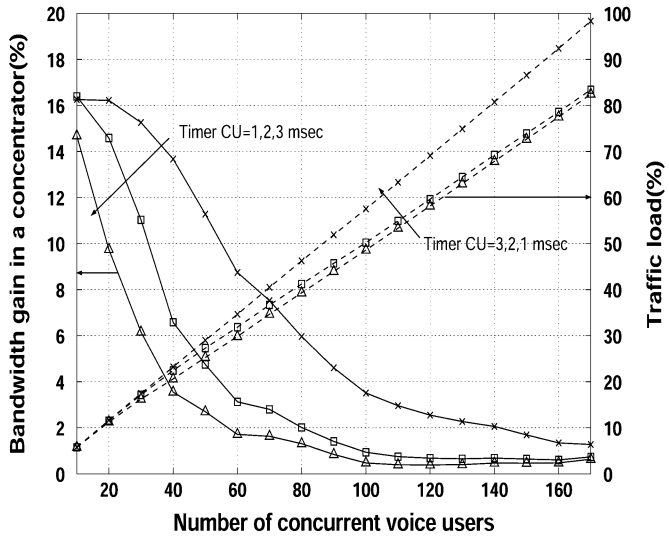

Fig. 8. Bandwidth gain in a concentrator on $\mathrm{I_{ub}}$ interface (Timer_CU of Node-B fixed 1 ms and concentrator $\mathrm{Timer\_CU} = 1, 2,$ and 3 ms; $\mathrm{total\ voice\ users} = 170$).

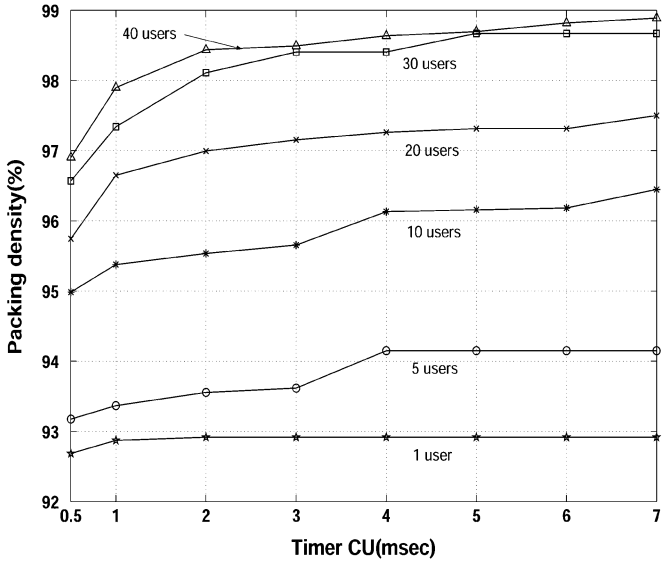

Fig. 9. ATM cell-packing density with various Timer_CU values (web browsing).

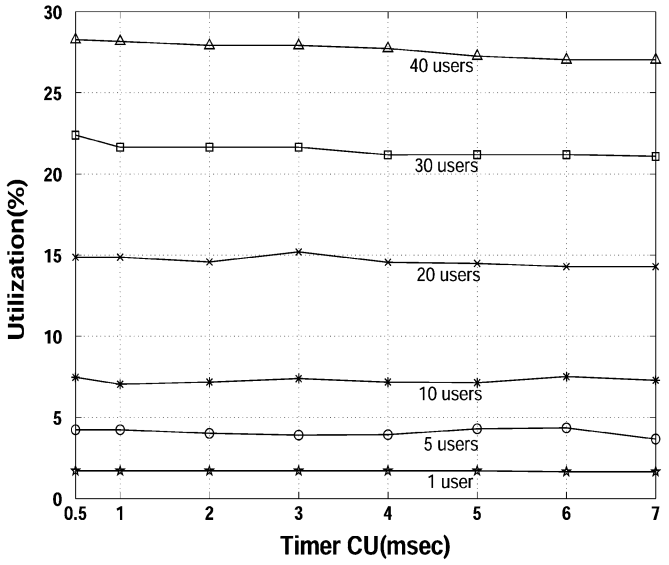

Fig. 10. Utilization vs. number of concurrent users and Timer_CU values (web browsing).

that the upstream traffic increases. This upstream traffic includes RLC, DchFP and ATM protocol overhead as well. Fig. 9 shows that the cell packing density is higher than 92% for every scenario with Timer_CU of 0.5 msec even though there is one user. This is the reason that IP packet is already long enough to fill an ATM cell except the last fragment of the IP packet loaded in an ATM cell. It can be seen from Fig. 10 that the link utilization is hardly changed according to varying Timer_CU values. This result implies that the bandwidth gain is not obtained even at the first multiplexing place (Node-B) for web browsing traffic. Thus, the bandwidth gain with an additional AAL2 multiplexing at the concentrator would be negligible.

V. Conclusion

In this paper, we analyzed the performance of AAL2 multiplexing in a Node-B and in a concentrator on $\mathrm{I_{ub}}$, with the consideration of UMTS-specific protocols such as SCR, FP, and RLC. To derive the cell-packing density and the bandwidth gain in a Node-B for the voice service, we applied the discrete Markov chain model. We considered not only the talkspurt packet but also the comfort noise packet in the silence period which has impact on traffic load and packing density. The mathematical results are validated by using the detailed simulation, which considered UMTS-specific protocol behavior. We also evaluated the benefit of additional AAL2 multiplexing in a concentrator by using simulations.

We concluded that the bandwidth gain of AAL2 in a Node-B and a concentrator on I_{ub} depended heavily on the traffic load and the maximum bandwidth gain with AAL2 is about 18% higher than the bandwidth gain without AAL2 in the node-B. There is also no significant AAL2 multiplexing benefit in the concentrator in terms of the bandwidth gain. Based on the analytic results, we propose the engineering guideline that can be used to select the optimal Timer_CU in a Node-B. For the data traffic, the benefit of the AAL2 multiplexing in I_{ub} is less than that for the voice service. The main contributions of this paper are threefold:

1) The performance of AAL2 in a Node-B is derived analytically, with the consideration of UMTS-specific protocol overhead.
2) The basis for selection optimal Timer_CU is proposed.
3) The decision-making criterion for the use of the AAL2 feature in a concentrator on I_{ub} is suggested.

If a service provider or UMTS network architecture designer have an expected user traffic profile, the results of this paper can help to select the optimal Timer_CU and make a fair product selection decision for a concentrator.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their comments and suggestions, which helped to improve the presentation of the paper. The authors are grateful to B. H. Kim. D. J. Houck, and I. S. Lee for initiating of mathmetical analysis in this paper.

REFERENCES

[1] UTRAN I_{ub} Interface., Dec., 2000. 3GPP TS 25.430 v.3.4.0, Release 1999.
[2] G. Eneroth, G. Fodor, G. Leijonhufvud, A. Racz, and I. Szabo, "Applying ATM/AAL2 as a switching technology in third-generation mobile access networks," *IEEE Commun. Mag.*, vol. 37, pp. 112–122, Jun. 1999.
[3] S. Nananukul, Y. Guo, M. Holma, and S. Kekki, "Some issues in performance and design of the ATM/AAL2 transport in the UTRAN," in *Proc. IEEE WCNC 2000*, vol. 2, Chicago, IL, Sep. 2000, pp. 742–746.
[4] R. Makke, S. Tohme, J.-Y. Cochennec, and S. Pautonnier, "Performance of the AAL2 protocol within the UTRAN," in *Proc. ECUMN 2002*, Colmar, France, Apr. 2002, pp. 92–100.
[5] C. Liu, S. Munir, R. Jain, and S. Dixit, "Packing density of voice trunking using AAL2," in *Proc. GlobeCom99*, vol. 1(B), Rio de Janeiro, Brazil, Dec. 1999, pp. 611–615.
[6] D. J. Houck, B. H. Kim, and J. H. Kim, "End-to-end UMTS network performance modelling," in *Proc. Netw. 2002*, Germany, Jun. 2002, pp. 133–138.
[7] P. T. Brady, "A model for on-off speech patterns in two-way conversation," *Bell Syst. Tech. J.*, vol. 48, pp. 2445–2472, Sep. 1969.
[8] HTTP and FTP Traffic Models for 1xEV-DV Simulations.. 3GPP2-C50-EVAL-2001022-0xx.
[9] B-ISDN ATM Adaptation Layer Specification: Type 2 AAL., Nov., 2000. ITU-T Recommendation I.363.2.
[10] AMR Speech Codec Frame Structure., Dec., 1999. 3GPP TS 26.101, Release 1999.
[11] Mandatory Speech Codec Speech Processing Functions AMR Speech Codec: Source Controlled Rate Operation., Mar., 2003. 3G TS 26.093, Release 1999.
[12] Radio Link Control (RLC) Protocol Specification., Mar., 2003. 3G TS 25.322.
[13] UTRAN Iur and I_{ub} Interface User Plane Protocols for DCH Data Streams., Jan., 2003. 3G TS 25.427.
[14] ETSI TR 101.112 V3.2.0ETSI, Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS (UMTS 30.03)., Apr., 1998.
[15] Mandatory Speech Codec Speech Processing Functions; AMR Speech Codec; Comport Noise Aspects., Mar., 2001. 3G Ts 26.092, Release 4.
[16] H. J. Lee, J. H. Kim, and B. H. Kim, "Decision point of AAL2 multiplexing for voice and data services in 3G WCDMA network," in *Proc. ICOIN2004*, Pusan, Korea, Feb. 2004, pp. 1313–1322.

Hyun-Jin Lee was born in Gyeongsangnam-do, Korea, in 1979. He received the B.S. degree from Ajou University, Suwon, Korea, in 2004, and is working toward the M.E. degree and Ph.D. degree in electrical engineering at Ajou University.

His research interests are in QoS, especially network optimization and wireless packet scheduling.

Mr. Lee was awarded 10th Samsung HumanTech Thesis Prize, Mar. 2004.

Jae-Hyun Kim (S'89–M'99) was born in Seoul, Korea. He received the B.Sc, M.Sc., and Ph.D. degrees, all in computer science and engineering, from Hanyang University, Ansan, Korea, in 1991, 1993, and 1996, respectively.

In 1996, he was with the Communication Research Laboratory, Tokyo, Japan, as a Visiting Scholar. From April 1997 to October 1998, he was a Post-Doctoral Fellow at the Department of Electrical Engineering, University of California, Los Angeles. From November 1998 to February 2003, he worked as a Member of the Technical Staff in the Performance Modeling and QoS Management Department, Bell Laboratories, Lucent Technologies, Holmdel, NJ. He has been with the Department of Electrical Engineering, Ajou University, Suwon, Korea, as an Assistant Professor since 2003. His research interests include QoS issues and cross layer optimization for high-speed wireless communication.

Dr. Kim was the recipient of the LGIC Thesis Prize and Samsung HumanTech Thesis Prize in 1993 and 1997, respectively. He is a Member of the Korean Institute of Communication Sciences (KICS), Korea Institute of Telematics and Electronis (KITE), and Korea Information Science Society (KISS).