ABSTRACT
User-perceived application-level performance is very important in the adoption and success of UMTS (Universal Mobile Telecommunication System) services and infrastructure. To predict this performance, a detailed end-to-end UMTS network performance simulator was developed. The simulator allows the modeling of application traffic characteristics, network architecture, network elements, and protocols. It has been used to assess architectural alternatives, identify system bottlenecks, and validate performance requirements.

This paper illustrates the use of this simulator to study the AAL2 switching effect across the interface Iub between a base station and a RNC. In particular, two architectural alternatives are compared. The results indicate that end-to-end application-level performance must be incorporated in all development phases and the benefit of AAL2 switching in Iub depends heavily on the traffic load. Furthermore, the benefit of AAL2 switching in Iub for data services is less than that for voice service and decreases with offered load per base station.

Keywords: UMTS, AAL2, Multiplexing, End-to-end performance, Simulation

1 INTRODUCTION
Wideband code division multiplexing access (WCDMA) has been chosen as the technology to be used on the air interface in 3G wireless network defined by the third generation partnership project (3GPP). The 3GPP is the standardization body for the Universal Mobile Telecommunication System (UMTS). UMTS network will offer mobile telephony, circuit switched service, multimedia and packet-switched services, Internet and Intranet access, entertainment and value-added services. ATM is selected for the UMTS terrestrial radio access network (UTRAN) transport due to its ubiquitous nature for heterogeneous traffic types, quality of service (QoS) guarantee and its widespread deployment in public networks [1]. However, applying ATM to low bit rate mobile voice stream is inefficient due to the delay in filling out the payload of an ATM cell. Recognizing the problems, ITU-T standardized a new adaptation layer type2 (AAL2) for the bandwidth efficient transmission of low bit rate delay sensitive application and in the 3GPP standardization, an ATM AAL2 transmission scheme on the Iub interface between base station (NodeB) and radio network controller (RNC) is used to deliver voice and data traffic. In the Iub interface, voice and data frame packets with a small size are transmitted due to a short frame length of 10 ms, 20 ms, 40 ms, or 80 ms. Therefore, an AAL2 transmission scheme, which is suitable for small sized packet has been chosen in the 3GPP standardization in order to increase bandwidth efficiency using AAL2 packet multiplexing capability.

Since the AAL2/ATM protocol suite is mandatory in the first and second release of UMTS and may be applied to both the access and the core network, vendors are forced to offer AAL2/ATM transmission equipment. Hence, the traffic performance of AAL2 is one of the most important topics in this area and several papers analyzed it by computer simulation or by simple experiment [2]. Many studies on this AAL2 switching show that the gain obtained by AAL2 is significant in terms of bandwidth [2], [3] and point out the importance of selecting Timer_CU value since it significantly affects the link efficiency. However, most of these papers are not UMTS network focused and not considering UMTS specific protocol behaviors for both voice and data traffic to evaluate the AAL2 efficiency. The lately published paper [2] includes some of protocols which impact throughput in Iub such as RLC and Framing Protocol (FP) and focuses on comparison between different scheduling mechanisms in order to chose the suitable algorithm for the CPS multiplexer and briefly mentioned the importance of selecting ATM switch or AAL2 switch within the Iub without quantitative
analysis. [2] still uses simplified models such as approximated protocol overhead, no RLC retransmission because Block Error Rate (BLER) is not explicitly modeled, and no Source Controlled Rate (SCR) Operation, which changes voice traffic behavior and may result in reducing the bandwidth and increasing the ATM cell packing density, the ratio of the average user bytes (excluding ATM and CPS headers) in a cell onto the ATM cell length. All of these are modeled in this study to evaluate the AAL2 protocol in UTRAN more precisely.

In this paper, AAL2 switching in a Concentrator in the Iub is considered, and in order to evaluate the performance (bandwidth efficiency) of AAL2 multiplexing (using Timer_CU parameter) we have extensively used the UMTS network performance simulator, which supports the investigation of both connection-, cell- and bit-level aspects of ATM/AAL2 switching networks [4]. We use the UMTS specific traffic model such as the AMR codec with SCR feature for UMTS voice traffic and use HTTP1.1 protocol for web traffic. In addition to the application layer protocol, various UMTS specific network protocols such as RLC and Framing Protocol (FP) are modeled to evaluate AAL2 switching effect more precisely. RLC retransmission attempts to recover the corrupted blocks within the air interface according to the BLER model before a recovery mechanism from an upper layer protocol such as TCP is in action. The RLC retransmission model has an important role of changing the traffic model within the Iub.

The results indicate that the benefit of AAL2 switching in Iub depends heavily on the traffic load. Furthermore, the benefit of AAL2 switching in Iub for data services is less than that for voice service and decreases with offered load per base station. We provide an optimal value of the Timer_CU based on ATM cell packing density for voice and web traffic, and we also suggest a break point to make a decision of selecting ATM switch or AAL2 switch within Iub. This result may help on a UTRAN architecture engineering to identify unnecessary locations for AAL2 switches, which requires further complexity in the existing ATM switch.

In Section 2, simulation models used in the simulations are introduced and Traffic models within Iub are described in Section 3. In Section 4, the simulation scenarios and configuration parameters are listed, and simulation results are evaluated in Section 5. Finally, conclusion is presented in Section 6.

2 UMTS SIMULATION MODEL DESCRIPTION

This section describes a user-plane simulation model of an end-to-end reference connection through a UMTS network. The simulator models all protocol layers from the physical through the application layer and models details of the packet handling characteristics of each network element along the path. The simulation model predicts application-level performance metrics such as response time, packet loss, jitter, and throughput and has been used to assess architectural alternatives, identify performance bottlenecks, and validate performance requirements. Refer to [4] for detail on 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 published traffic characteristics [5], [6]. For the voice traffic model a mobile-to-mobile reference connection is assumed and for the web browsing, client-server models are used.

Network Architecture

The network architecture in this study is a simplified from the original UMTS Simulator described in [5]. This will not loose any capability of analyzing AAL2 switching efficiency in Iub. Figure 2-1 shows the network model for the simulation. In this study, voice traffic and web browsing data traffic are offered to the UMTS network, which consist of ten Node-Bs, one ATM switch or AAL2 switch (labeled as Concentrator) and one RNC in the UTRAN and other core network element to support the voice and data application traffic include web server and voice called parties. The NodeB-to-Concentrator link has a capacity of one E1 and the Concentrator-to-RNC link capacity is STM-1.

Figure 2-1 Network Architecture

Protocol Stack Models for CS and PS Service

Figure 2-2 shows the protocol stack modeled in the simulator for packet switched (PS) service and circuit switched (CS) service within UTRAN with AAL2 switch along the Iub interface. The protocol stack with ATM switch can be derived by removing the AAL2 layer in the AAL2 switch in Figure 2-2. The left column in the shaded box is for PS service and the right column is for the CS service. The non-shaded boxes are common for PS and CS service, AMR, SCR, RLC, DchFP and AAL2 models are described in detail in the following sections. Figure 2-2 does not showing Iur interface assuming the Controlling RNC (CRNC) and Serving RNC (SRNC) are co-incident. The air interface is not modeled explicitly as that would slow the simulation down too much. Instead a separate model was used to generate trace files of air interface performance as given by the Block Error Rate (BLER) under various conditions. This trace file of BLER was fed into the RLC layer for RLC acknowledge-mode operation to be modeled correctly.
Payload employs a one octet Start Field (STF) followed by a 47 octet header and a variable length payload. AAL2 uses the 8-bit Channel ID (CID) in the CPS Packet Header to multiplex multiple 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 of time (Timer_CU). If the cell is not completely packed within the time period determined by this Timer_CU value, the timer expires and the partially packed cell will be sent. The CPS-PDU employs a one octet Start Field (STF) followed by a 47 octet payload.

Adaptive Multi Rate (AMR) Codec Model

The Adaptive multi Rate (AMR) codec is the most important vocoder in UMTS and Source Controlled Rate (SCR), called “Discontinuous transmission” in GSM, functionality is also part of the standard [7], [8]. 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 and 4.75 kbit/s and the coder operates on speech frames of 20 ms corresponding to 160 samples at the sampling frequency of 8000 sample/s. The simulation works in this paper consider only one of the specified rates during the ON state (talk spurt) of the AMR codec (for 12.2 kbps) and the comfort noise during the OFF state of the AMR codec.

Radio Link Control Protocol Model

The Radio Link Control protocol (RLC) provides segmentation and retransmission services for both user and control data. Each RLC instance is configured by RRC to operate in one of three modes: transparent mode (Tr), unacknowledged mode (UM) or acknowledged mode (AM). Transparent Mode and Acknowledge Mode are used for the user plane and Unacknowledged mode is not used in this study because the usage of this mode is mainly for RRC signaling and VoIP. Therefore these two modes are described in this section as modeled in the simulator (see [9] for the detail on the RLC protocol). The transparent mode RLC entities are defined to be unidirectional, whereas the acknowledge mode entities are described bi-directional. In the transparent mode (Tr) no protocol overhead is added to higher layer data and the transmission of the streaming type in which higher layer data is not segmented. The Tr mode can be used for circuit service such as voice call. In the acknowledged mode (AM), an ARQ mechanism is used for error correction. The AM is the normal RLC mode for packet-type services, such as Web browsing and email downloading.

Data Flow through the RLC Layer (AM): This section takes a closer look at how data packets pass through the RLC layer. Figure 2-3 shows a simplified block diagram of an AM-RLC entity. The figure shows only how an AMD PDU can be constructed in the simulator. Data packets (RLC SDUs) received from higher layers via AM-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 is 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, adding a header for 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 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 is in the receiving buffer, the SDU is reassembled. After this, the checks for in-sequence delivery and duplicate detection are performed before the RLC SDU is delivered to the higher layer.

Figure 2-3 A simplified block diagram of an RLC AM entity

DCH Frame Protocol Model

The purpose of the user data frames is to transparently transport the transport blocks between NodeB and Serving RNC. The protocol allows for multiplexing of coordinated dedicated transport channels, with the same transmission time interval, onto one transport bearer. The transport blocks of all the coordinated DCHs for one transmission time interval are included in one frame [10]. The header contains a CRC checksum, the frame type field and information related to the frame type. There are two types of DCH FP frames (indicated by the Frame type field): DCH data frame and DCH control frame. We modeled DCH data frame protocol since the user plain is the interest in this paper. See [10] for the detail on the structure of the UL and DL data frame.
3 SERVICE TRAFFIC MODELS AND TRANSPORT OVER Iub

It is assumed that each user traffics related to Radio Bearers over Iub will be carried within frame protocol (FP), defined for each of the types physical channel over radio. To consider traffic flow over Iub, it is necessary to consider the various protocol overheads over the Iub interface. In this paper it is assumed that all Radio Bearers are carried with dedicated physical channels (DPDCH) within DCH frame protocol over Iub. To determine the bandwidth efficiency, voice and web browsing traffic models are used.

Voice Traffic Model

Voice call activities generate a pattern of talk spurt and silence (or ON and OFF) intervals by means of a speech activity detector so that it can be modeled as a two state Markov chain. It has been found that length of talk spurt and silence period are exponentially distributed [5]. In the voice traffic model 3 sec for both talkspurt and silence period is used followed by [11].

Web Browsing Traffic Model

The distributions of the parameters for the web browsing traffic model is determined in [6] and an application session is divided in ON/OFF periods representing page web downloads and the intermediate reading times. These ON and OFF periods are a result of human interaction, which represents a user’s request information and the reading time identifies the time required to digest the web page. The initial HTML page is referred as the “main page” and the each of the constituent objects referenced from the main page are referred to 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 widely used protocols, is used in the web browsing simulation.

Here we model the arrival of sessions, characterize the arrival of page requests within a session, and the number of objects and their sizes for each page. It is assumed that maximum data rate is 64kbps for uplink and 144 kbps for downlink, and we do not assume compression on user plane TCP/IP header for data packets in this simulation.

4 Simulation Scenarios and Parameters

It is assumed that each of the simulation scenarios contains only voice or web traffic and consists of multiple voice traffic generators or data traffic generators in UEs. Each traffic generator can generate multiple sessions of voice or web browsing traffic. To evaluate AAL2 switching/multiplexing effectiveness at the concentrator in Iub, the bandwidth utilization and the packing density at the link between NodeB and Concentrator (ATM switch or AAL2 switch) and between the Concentrator and RNC have been monitored. In the evaluation, only uplink traffic was monitored for voice traffic since uplink and downlink traffic for voice would be similar, and for the web traffic both uplink and downlink are monitored because of the asymmetric traffic behavior. In the uplink, the packet size would be smaller and there would be less traffic compared to the downlink. Through the simulations, AAL2 multiplexing in a NodeB and RNC is used with Timer_CU value from 1 msec to 9 msec as a default since AAL2 is a mandatory between these two network elements while ATM switch or AAL2 switch with Timer_CU value from 1 msec to 9 msec is used as a concentrator in Iub. The maximum number of application sessions generated to a NodeB from UEs attached in the NodeB is the “Max [number of sessions generate traffic = E1 link, 248 sessions]”. The first NodeB is used until the E1 link between itself and the adjacent Concentrator reaches the link capacity and more application sessions are added to the next NodeB until the total number of application sessions is 248. Thus the maximum number of application sessions from the Concentrator to the RNC is 248 through the simulation scenarios. The transmission time interval (TTI) for voice session is set to 20 ms and 40 ms for web browsing session. The 3GPP defined the maximum allowable block error rate (BLER) for voice should be < 10-2 and for data should be < 10-1. In this paper, 1% BLER for voice and 4% BLER for 64 kbps uplink (UL) and 5% BLER for 144 kbps downlink (DL) for data traffic are used. The maximum allowable retransmission to recover the block error between RLC layers in the UTRAN is limited to three, and those packets could not recovered by the RLC layer are rely on the recovery mechanism in TCP protocol.

5 SIMULATION RESULTS

There is a very important assumption (setting Timer_CU = 1 msec in NodeB) through the simulations in this paper. To understand the effectiveness of the additional AAL2 switching in the Concentrator, a set of simulations has been performed with various Timer_CU values and number of concurrent users for the voice and web browsing traffic.

Voice Traffic Scenarios

The ATM cell scenarios and the link utilization results with various Timer_CU values and number of concurrent users for the voice traffic are summarized in Figure 5-1 through 5-4. Since AAL2/ATM has to be implemented in NodeB and RNC by 3GPP standards, setting the Timer_CU to 0 msec and 1 msec or other values is up to the network operator. This study shows that about 170 concurrent voice calls can be connected in a NodeB with a Timer_CU = 0 msec because of the link type (E1) between Node-B and Concentrator. By setting Timer_CU = 1 msec in the Node-B, however, more than 200 concurrent voice calls can be served in a Node-B without any other system or network changes (Figure 5-3). This gives a strong reason to use AAL2 multiplexing in NodeB.

Figure 5-1 ATM cell packing density with various Timer_CU and number of concurrent voice users
In Figure 5-1 and Figure 5-2, we measured the link throughput between Node-B and Concentrator, and the payload in a cell (packing density) in NodeB. In this scenario, we set up one NodeB and vary the number of voice calls in a Voice UE from 1 to 170 users. Timer_CU in Node-B, Concentrator and RNC are set to 0 through 9 msec. To understand how many AAL2 packet bits are packed into an ATM cell, we measured the payload packing density in a cell. Payload density in cell is calculated by Eq. (1)

\[ \text{Payload}(\%) = \frac{B}{47 \times 8} \times 100 \]  

Where, \( B \) is the payload size in bytes and 47 bytes is the maximum bytes for an AAL2 packet per ATM cell. Based on the simulation results, the minimum packing density, which obtained without AAL2 multiplexing (Timer_CU=0msec), is about 83.5% and AAL2 payload density reaches 96% when the number of voice calls is 80 with Timer_CU=1 msec and it reaches 96% when the number of voice calls is 40 with Timer_CU=2msec. This high minimum packing density is the result from the SCR feature. The average packing density without AAL2 multiplexing can be verified as following.

\[ \text{PktCount}_{\text{talk}} = \frac{\text{Duration}_{\text{talk}} (3sec)}{\text{TTI}(20ms)} + \frac{\text{Hangover}(7)}{8} = 157 \]

\[ \text{PktCount}_{\text{comfort\_noise}} = \frac{\text{Duration}_{\text{silence}} (3sec)}{\text{TTI}(20ms)} \]

\[ B = \text{TalkPktSize}(42B) \times \left( \frac{\text{PktCount}_{\text{talk}}}{\text{PktCount}_{\text{talk}} + \text{PktCount}_{\text{comfort\_noise}}} \right) + \frac{\text{SilencePktSize}(15B) \times \text{PktCount}_{\text{comfort\_noise}}}{\text{PktCount}_{\text{talk}} + \text{PktCount}_{\text{comfort\_noise}}} = 39.2 \text{Bytes} \]

The average packet size (B) without AAL2 multiplexing is 39.2 bytes and it gives 83.5% of packing density according to the Eq. (1). To assess the bandwidth (BW) efficiency of the AAL2 in Iub, the traffic load and bandwidth gain were measured in Iub. Bandwidth gain is defined by Eq. (2)

\[ \text{BW Gain}(\%) = \frac{\sum \lambda_{\text{Iub Node-B}} - \lambda_{\text{Iub Concentrator}}}{\sum \lambda_{\text{Iub Node-B}}} \times 100 \]  

Where, \( \lambda_{\text{Iub Node-B}} \) means the carried load from NodeB to the Concentrator and \( \lambda_{\text{Iub Concentrator}} \) means the carried load from the Concentrator to RNC measured in cells per second. The results are plotted in Figure 5-3. We fixed the total number of voice calls to 248, which is a maximum number of different users that can be multiplexed in an AAL2 packet since the 8 bits Channel Identifier (CID) identifies the individual user channel. We changed the number of voice calls in a NodeB from 10 to 170. The number of voice calls in the last NodeB was padded to make 248 calls. For example, 100 voice calls are generated by 2 NodeBs and the remaining 48 voice calls are generated by the third NodeB. As you can see in the graph, increasing traffic load per NodeB from 15% to 33% (the number of voice calls are increased from 30 to 60 per NodeB) results in the drastic drop of BW gain from 11 to 3%. Figure 5-3 indicates that there is no significant AAL2 switching benefit (less than 5%) in terms of bandwidth gain when the traffic load from a Node-B is above 30%.

Figure 5-4 shows the BW gain and the packing density in a Concentrator while the number of NodeB varies 1 to 8 and the each NodeB multiplexes 20 voice calls. As the number of NodeB increases from 1 to 4, the BW gain and payload density shows big change while there is only 2% increase for both between 4 to 10 Node-Bs. We found two important impacts from the results. First, the maximum AAL2 switching gain in a Concentrator is less than 4 % when the traffic in a NodeB is about 20%. Second, there is no big increase of AAL2 switching benefit when the number of NodeBs is greater than 4 with 20 users/NodeB.
Data (Web browsing) Traffic Scenario

The ATM cell packing density and the link utilization results with various Timer_CU values and number of concurrent users for the voice traffic are summarized in Figure 5-5 and Figure 5-6. These two figures represent downlink channel only. The traffic load generated by 40 simultaneous web-browsing sessions is about 600 Kbps (~28% of E1) at the link between NodeB and Concentrator after AAL2 multiplexing with Timer_CU of 1 msec at each NodeB. The traffic on this link includes DCHFP and ATM protocol overhead as well.

Figure 5-5 shows that the ATM cell packing density is larger than 93% for every scenario even only one user with Timer_CU of 0.5 msec and bandwidth gain is not observed in Figure 5-6 either. Since IP packet is already long enough to fill an ATM cell except the last fragment of the IP packet loaded in an ATM cell, the average ATM cell packing density is high even with only one user and this implies that the AAL2 multiplexing gain is very minimal even at the first multiplexing place (NodeB). Thus, the bandwidth gain with an additional AAL2 switching at the concentrator would be negligible.

![Figure 5-5 ATM cell packing density with various Timer_CU and number of concurrent users (Web Browsing)](image)

![Figure 5-6 Link Utilization VS. Number of concurrent users and Timer_CU values (Web Browsing)](image)

6 Conclusions

This paper provides analysis of the AAL2 switching effect in terms of bandwidth gain due to Timer_CU value in the Iub link in a UTRAN. UMTS standards specify two options for placing AAL2 switching functionality at Iub and Iur. In this paper, AAL2 switching at a Concentrator at the Iub in addition to the AAL2 multiplexing in a NodeB is considered, and in order to evaluate the performance (bandwidth efficiency) of AAL2 multiplexing (using Timer_CU parameter) we have used the UMTS network performance simulator with UMTS specific traffic model such as the AMR codec with SCR function for UMTS voice traffic and used HTTP1.1 protocol for web traffic. In addition to the application layer protocol, various UMTS specific network protocols such as RLC protocol and Framing Protocol (FP) are modeled to evaluate AAL2 switching effect more precisely.

The results indicate that the benefit of the additional AAL2 switching in Iub depends heavily on the traffic load. The bandwidth gain at the Concentrator according to the additional AAL2 switching becomes negligible when Timer_CU at the NodeB is 2 msec and 40 concurrent voice application users exist in the NodeB since the ATM cell packing density at the NodeB is already larger than 98%. Furthermore, the benefit of the AAL2 switching in Iub for data services is less than that for voice service since most of the ATM cell’s packing density is 100% except the last fragment of the IP packet loaded in to an ATM cell and decreases with offered load per base station. If a service provider or UMTS network architecture designer had an expected user traffic profile, it can be used with this paper to make a product selection decision for the ATM concentrator.

7 ACKNOWLEDGEMENT

We would like to thank the Department of Computer and Information Science, University of Pennsylvania for their financial support in presenting this paper at the CCCT conference.

8 REFERENCES

[8] 3G TS 26.093: “Mandatory speech codec speech processing functions AMR speech Codec; Source Controlled Rate Operation (Release 1999).
[10] 3G TS 25.427: “UTRAN Iub and Iur interface user plane protocols for DCH data streams”