

# Delay in UMTS Radio Access Networks: Analytical Study and Validation

Xi Li, Su Li, Andreas Timm-Giel and Carmelita Görg  
Institute of Communication Networks  
University of Bremen  
Bremen, Germany  
{xili | suli | atg | cg} @comnets.uni-bremen.de

Richard Schelb  
Siemens AG  
Berlin, Germany  
richard.schelb@siemens.com

**Abstract**—This paper deals with the delay analysis of the traffic within the UMTS Terrestrial Radio Access Network (UTRAN). Delay is a very important Quality of Service aspect in the UTRAN. Especially, the delay over the Iub link is the major limited factor for the dimensioning of the UTRAN. Based on the system analysis, a queuing model is proposed in this paper to derive the delay time over the Iub interface. The arrival process is based on Batch Markovian Arrival Process (BMAP). The suggested queuing model and the related notable results are validated by comparing with simulations. Additionally, we propose a superposition technique applying on the BMAP model to scale the source traffic. Then in this way, not only the delay performance on the Iub interface can be calculated analytically with the proposed queuing model, but also with the suggested superposition technique, the traffic can be scaled to different load ranges based on a basic load, and furthermore the delay times for all various ranges of traffic loads can be also derived with the proposed queuing model.

**Keywords**—UTRAN, Iub, FP PDU delay, BMAP, BMAP/D/I

## I. INTRODUCTION

The Universal Mobile Telecommunication Systems (UMTS) is a third generation mobile communication system based on Wideband Code Division Multiple Access (WCDMA) under standardization at 3GPP [9]. As shown in Figure 1, it consists of three main components: User Equipment (UE), Radio Access Network (UTRAN) and Core Network (CN).

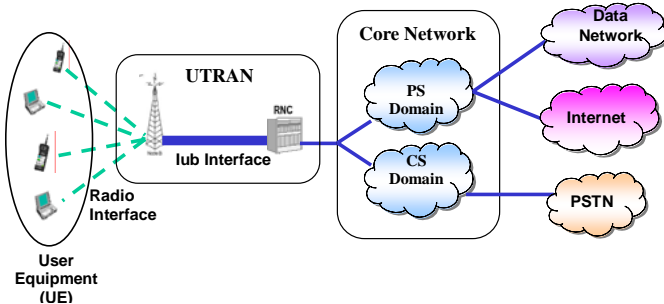


Figure 1. UMTS Network Structure

UMTS supports a wide range of applications and services with seamless and high speed transmissions in a global

mobility environment while granting each of the application or service a certain Quality of Service (QoS). This raises new challenges in designing the UMTS network, especially in the UTRAN where the transmission resources and the operating costs are most expensive. With the extension of the UTRAN to cover more and more suburban and rural areas, the transmission resources in the UTRAN become considerably limited and costly. In order to achieve a cost-efficient design of the UMTS network, a significant amount of research activities have been focused on the evaluating the performance and determination of the minimum required bandwidth for ensuring certain performance level in UTRAN. The most important performance aspects in UTRAN is the stringent delay bounds for the transport of various types of user traffic over the Iub interface, i.e. the interface between NodeB and Radio Network Controller (RNC). This is imposed by the WCDMA radio control functions, i.e. each mobile or User Equipment (UE) is only allowed to send and receive data over the radio air interface every Transmission Time Interval (TTI), e.g. 10, 20, 40, 80ms. This in turn requires that the transport of a radio frame from the RNC to the NodeB should be strictly on time to ensure this radio frame to be able to be delivered to the UE according to the time requirements of the air interface. Whenever a radio frame does arrive at the NodeB too late, i.e. too much behind the scheduled transmission time, then this radio frame is no more able to deliver to the air interface and will be immediately discarded by the NodeB, which will result in a RLC retransmission (if using RLC Acknowledged Mode). One of the main reasons of the late arrival of the radio frame from RNC to NodeB is the long queuing delays due to the congestion on the link of a specific NodeB (Congestions within a more complex network are out of the scope of this article). The retransmissions of the radio frames will waste the limited transmission resources in the UTRAN as well lead to a higher delay and lower QoS for the end user. In order to determine the maximum traffic load allowed on the Iub link ensuring an acceptable user performance (e.g. achieve below certain NodeB discard ratio for late packet arrivals caused by the long delay), an accurate dimensioning method considering these effects is necessary. For this purpose, it is vital to study the delay over the Iub link.

The delay strongly depends on the characteristics of the source traffic. For example, Internet traffic with the nature of bursty, self similarity or long range dependence (LRD) usually experiences much worse delay performance than traditional

voice traffic as a result of series degraded queuing performance caused by the high burstiness of the traffic. Therefore to capture the properties of the source traffic is the key to estimate the delay. Hence, with an accurate source traffic model, according to the queuing theory the delay over the Iub interface can be calculated precisely and in turn to indicate the Iub performance, then based on the required delay QoS requirements of the Iub link, an optimum bandwidth to recommend for the Iub link can be determined.

This paper focuses on the investigation and analysis of the delay over the UTRAN Iub interface. Based on the sophisticated system analysis, a queuing model is presented in this paper to derive the delay time experienced on the Iub. The arrival process is based on Batch Markovian Arrival Process (BMAP). The BMAP model is an analytically tractable model, which considers different packet lengths and batch arrivals. The main advantage of the BMAP model is that it can capture two important statistical properties of IP traffic, namely burstiness and self-similarity. Nowadays high-speed data transferring becomes the major trend in UMTS network. More and more mobile subscribers access UMTS network to request Internet services such as email, file downloading or web browsing. Therefore, UMTS is not a voice-only network any more but an integrated services IP network. Therefore, the simple Poisson traffic model is not applicable in the UMTS since the Poisson model is typically used for the classical telephone networks and it can not capture any burstiness or self-similarity properties of the IP traffic. In the earlier work [1], the BMAP model has been demonstrated to accurately capture the characteristics of the aggregated IP traffic in the UTRAN. In this paper, we mainly focus on the IP traffic. The proposed queuing model and related notable results are validated by comparison with simulations. And the network performance in terms of delay, resource usage (e.g., queue length) can be derived from the queuing model. The other contribution of this paper is that we propose a superposition technique on the BMAP model to scale the source traffic. Then from the simulation point of view, instead of running different simulation scenarios with different traffic loads to provide different traces for constructing the BMAP model, the traffic load can be scaled to various ranges based on a single simulation trace.

The remainder of the paper is organized as following: In section II, a detailed analysis of the delay over the Iub is given. Section III presents the system model. In this section, the BMAP model and the queuing model for calculating the delay over the Iub interface is introduced. Section IV presents the main results of validating the queuing model by comparing with simulations. Section V discusses the results of scaling the traffic with BMAP. Section VI concludes the paper.

## II. DELAY ANALYSIS IN UTRAN

### A. UTRAN Overview

The protocol structure of the UTRAN is shown in Figure 2.

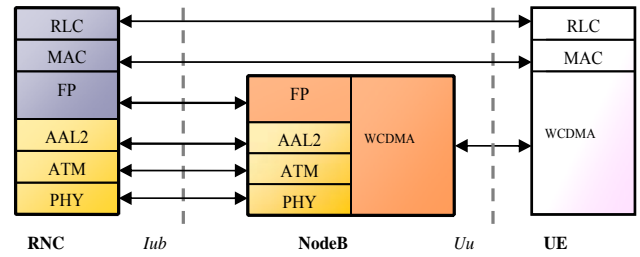


Figure 2. UTRAN Protocol Stack

After the data arrives at the RNC from the network, the Radio Link Control (RLC) layer will segment the higher layer packet, e.g. IP packet, into appropriate RLC Packet Data Units (PDU) and then pass them to the Medium Access Control (MAC) layer. The MAC layer maps the logical channels to the transport channels, and forms sets of transport blocks and schedules them according to the timing requirements of WCDMA, i.e. every transmission time interval (TTI). TTI defines the scheduled period, which is the time between two consecutive radio frames sent to the air interface. The TTI value and the amount of user data per TTI are determined by the Radio Access Bearer (RAB). Each RAB provides a certain peak data rate to transfer user data between UE and CN for a certain traffic class. The selection of an appropriate RAB for a user connection is determined by the Radio Resource Management (RRM) in the RNC.

Each user traffic stream (voice, video or data) is delivered by a so-called Dedicated Transport Channel (DCH) between RNC and Node B, which is similar to a classic modem connection. The Frame Protocol (FP) is defined at the Iub interface responsible for the relaying of transport channels between the UE and the RNC via the NodeB. It extends the radio transport channels in the UTRAN. The DCH traffic stream is handed over from the FP layer to the Transport Network Layer (TNL) in the form of FP PDUs. Due to the MAC scheduling, the FP PDUs are transmitted or received every TTI and each FP PDU carries the amount of user traffic for the RLC blocks to be sent in one TTI. In UMTS release 99, ATM constitutes the current transport network in the UTRAN. As AAL2 protocol is chosen for transporting the user plane data at the Iub interface, FP PDUs are segmented into AAL2 packets. These AAL2 layer packets are then packed into ATM cells before being transmitted over the Iub link.

### B. Delay Analysis in the UTRAN

In the UTRAN the most important delay is the FP PDU delay, i.e. the delay of a radio frame across the Iub between RNC and NodeB. From the system point of view, the FP PDU delay comprises all delays accumulated on the Iub interface (see Figure 3): i. the segmentation and reassembly delay; ii. the queuing delay in the AAL2 buffer; iii. the queuing delay in the ATM buffer; and iv. the transmission delay or processing delay over the Iub link.

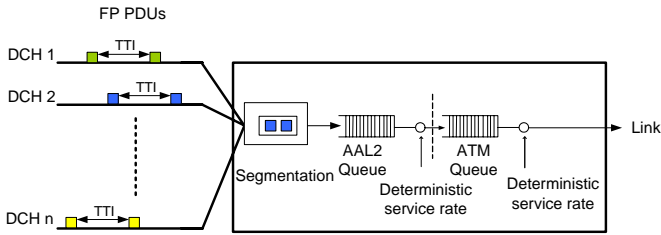


Figure 3. Delay in the UTRAN

i. The segmentation and reassembly delay includes the delay of segmenting FP PDU into AAL2 Packets and AAL2 packet to ATM cells, and reassembly for the other way around. This delay is usually neglected due to a very low value.

ii. As shown in Figure 3, the DCH traffic from each user is aggregated at the AAL2 layer, the arrival traffic of AAL2 queue can be seen as superposition of arriving FP PDUs streams of all users. In order to avoid overload on the ATM link, usually the aggregated AAL2 traffic (or flow) is limited by the available peak cell rate, called AAL2 shaping in this paper. With this shaping at the AAL2 layer, the service rate of the AAL2 queue is deterministic. For simplification, the AAL2 buffer is set to unlimited capacity within this paper, i.e. there is no packet discarding due to buffer overflow. The resulting queuing delay in the AAL2 buffer clearly depends on the arrival rate and the FP PDU size. Due to the stochastic arrival process, traffic can be very bursty in nature and the AAL2 queuing becomes significant. Estimation of the AAL2 queuing delay is the main focus of this paper.

iii. For the ATM buffer, the service rate is the link rate, which is also deterministic. Due to traffic shaping on the AAL2 layer, the data rate on the ATM layer is below the configured peak cell rate. Therefore there will be no congestion or overload situation on the ATM link in general. As a result, the queuing delay in the ATM buffer is quite small and stable.

iv. The transmission delay over the Iub link is rather low due to a high transmission speed. For transmitting an ATM cell on a 2Mbps E1 line it is 0.212ms. The node processing delay such as multiplexing, microwave coding, etc, is also very small, which can be neglected from the system point of view.

Following gives an example of UTRAN FP PDU delay, AAL2 queuing delay and ATM cell delay from an OPNET simulation. The simulation is set up for one NodeB connected with RNC via a 2Mbps Iub line, transmitting the web traffic with an average web page size of 15kbyte (Pareto page size distribution). In this example, the FP PDU delay is shown for the downlink, i.e. the delay of transporting FP PDUs from the RNC to the NodeB, measured at the NodeB side. The AAL2 queuing delay is the queuing delay of the AAL2 buffer at the RNC side. The ATM cell delay consists of (iii) the queuing delay in the ATM buffer and (iv) transmission delay on the Iub link. Table I gives the values of mean and variance of the ATM cell delay, AAL2 queuing delay and FP PDU delay from the simulation.

TABLE I  
UTRAN Delay Components

	Mean (s)	Variance
ATM Cell Delay	0.521e-3	1.836e-8
AAL2 Queuing Delay	0.09076	0.07549
FP PDU Delay	0.091789	0.075

The results from the above table show that the ATM cell delay is much smaller than the AAL2 queuing delay and the FP PDU delay, while the AAL2 queuing delay is quite close to the FP PDU delay. The gap between the AAL2 queuing delay and the FP PDU delay is mainly caused by the ATM cell delay and additional segmentation and reassembly delay. Furthermore, the variance of ATM cell delay is rather small, which means that the ATM cell delay is kept quite stable. This implies that the main component of FP PDU delay is the AAL2 queuing delay. This can be further verified by Figure 4 which compares the FP PDU delay, AAL2 queuing delay and ATM cell delay.

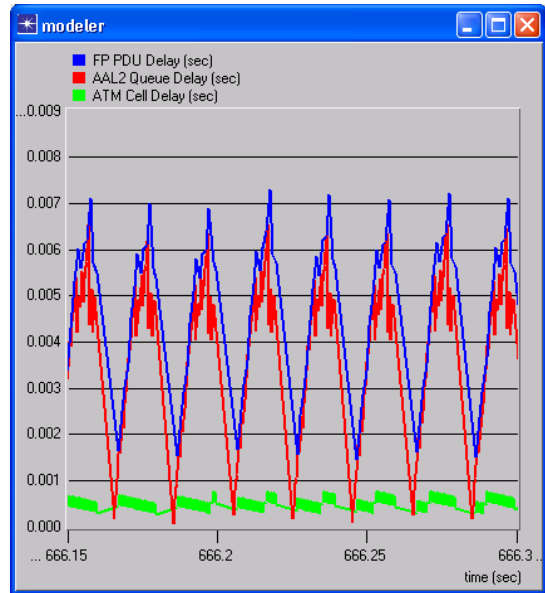


Figure 4. FP PDU delay vs. AAL2 queuing delay vs. ATM cell delay (s)

From the Figure 4, it can be seen obviously that the queuing delay in the AAL2 buffer contributes the major part to the FP PDU delay. The FP PDU delay is very important as it represents the delay of transporting a radio frame over the Iub link. Whenever a FP PDU delay is longer than a certain predefined boundary on the Iub interface, then the carried radio frame in this FP PDU can no longer be delivered to the air interface in time and will be immediately discarded by the NodeB. Hence, in order to evaluate the delay performance in the UTRAN, it is important to set up a queuing model to analytically estimate the AAL2 queuing delay, which is the major part of the FP PDU delay. The following section is going to introduce a queuing model proposed within this paper to calculate the AAL2 queuing delay.

### III. SYSTEM MODEL

The aim of this paper is to evaluate the AAL2 queuing delay, which is the major delay experienced at the Iub interface.

#### A. Modeling the Arrival Process

As already shown in Figure 3, the FP PDUs from each DCH are sent to the common AAL2 buffer every TTI and the traffic of all DCHs is aggregated at the AAL2 layer, i.e. the arriving traffic of the AAL2 queue is seen as superposition of arriving FP PDUs streams of all DCHs. The arrival process strongly depends on the aggregated arrival rate and the FP PDU size, and it is evidently not a simple Poisson process as the FP PDUs of each DCH arrives periodically (every TTI), but rather bursty in case of the IP traffic. UMTS provides a set of RAB types for transmission the user data. For different RAB types, the TTI and the FP PDU size are different.

Therefore, Batch Markovian Arrival Process (BMAP) model is chosen as the arrival process model since it considers different lengths of packets and batch arrivals. For example, if there are three different RAB types in the UMTS network, correspondingly the BMAP will consider three different packet lengths, i.e. FP PDU sizes. In this way BMAP can provide a rather accurate model for characterizing the aggregated traffic, specifically on the self-similarity and burstiness properties of the IP traffic.

#### B. BMAP

The Batch Markovian Arrival Process (BMAP) is the generalization of Phase-type (PH) distribution [7]. PH distribution is usually used to analyze an absorbing Markov chain, i.e. a Markov chain that includes at least one absorbing state. For example, an absorbing Markov chain with states  $(1, 2 \dots N, N+1)$ , where the states  $(1, 2 \dots N)$  are transient states and state  $(N+1)$  is the absorbing state. It is possible to transit from each non-absorbing state to the absorbing state in one or more time-steps. The distribution of time from transient state  $i$  to absorbing state  $(N+1)$  is called PH distribution. The infinitesimal generator of the PH distribution is given by:

$$Q = \begin{pmatrix} T & T^0 \\ 0 & 0 \end{pmatrix} \quad (1)$$

Matrix  $T$  is an  $N \times N$  matrix which represents the transitions among the transient states. Matrix  $T^0$  is a column vector of size  $N$  which represents the transitions from the transient states to the absorbing state  $(N+1)$ .

BMAP is characterized by a finite and absorbing Markov chain [4]. Considering a two-dimensional Markov process  $\{P(t), J(t)\}$  with state space  $\{(i, j); i \geq 0, 1 \leq j \leq N\}$ , here  $P(t)$  counts the number of arrivals in the interval  $(0, t)$ , and  $J(t)$  represents the underlying Markov chain. An infinitesimal generator  $Q$  has the following structure:

$$Q = \begin{pmatrix} D(0) & D(1) & D(2) & D(3) & \dots \\ 0 & D(0) & D(1) & D(2) & \dots \\ 0 & 0 & D(0) & D(1) & \dots \\ 0 & 0 & 0 & D(0) & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix} \quad (2)$$

Here,  $D(0)$ , similar to matrix  $T$  in PH distribution, is an  $N \times N$  matrix (the dimension equals to the number states of  $J(t)$ ), which has negative diagonal elements and nonnegative off-diagonal elements.  $D(0)$  represents a transient state transitions without packet arrivals.  $D(m)$  ( $m \geq 1$ ), similar to matrix  $T^0$  in PH distribution, is also an  $N \times N$  matrix, which has nonnegative elements representing a state transition with a packet arrival of batch size  $m$ . In case  $m = 1$ , it is a Markovian Arrival Process (MAP), in which there is only one kind of batch arrivals. If  $\pi$  is defined as the initial probability vector of the underlying Markov chain  $\{J(t)\}$ ,  $\pi$  should satisfy:

$$\pi \cdot D = 0, \pi \cdot 1 = 1 \quad (3)$$

$$D = \sum_{m=0}^{\infty} D(m)$$

with  $1$  is a column vector of ones.

The cumulative distribution function of the inter-arrival time for the batch size  $m$  is [5]:

$$F(t) = \pi \left( 1 - e^{D^{(0)}t} \right) \left( -D(0) \right)^{-1} D(m) 1 \quad (4)$$

The BMAP model is characterized by parameters  $D(0)$ , which represents the transition probability between the transient states without arrivals, and  $D(m)$ , which is the transition probability within an arrival of packet with batch size  $m$ . Normally, the BMAP traffic arrival process is difficult to calculate analytically, and its important metrics are often calculated numerically based on Expectation Maximization (EM) algorithm for estimating the parameters to fit to the measured data (mentioned in [1]). In this paper, the BMAP results are generated using a software package IP2BMAP which is provided by [2]. In previous researches, the BMAP model has been applied to analyze the aggregated traffic modeling of IP networks [8]. And in [1], the BMAP model has been demonstrated to model the aggregated IP traffic in the UTRAN, but only for a single RAB scenario. For a further validation, the results of the multiple RABs scenario are presented here.

Following gives an example of three Packet Switched RAB types being used at the same time, i.e. RAB 64kbps, 128kbps and 384kbps. The application is web browsing with Pareto page size distribution and an average page size of 15kbyte. The HTTP requests are equally distributed among these three RAB types. The Iub link is a 2Mbps E1 line and the Permanent Virtual Circuit (PVC) on the ATM layer is set to 1.600 Mbps.

Figure 5 compares the Cumulative Distribution Functions (CDF) of FP PDU inter-arrival time obtained from the measured simulation trace and from BMAP. It is obvious that the CDF of BMAP model matches the CDF of the measured traffic very well. The average FP PDU inter-arrival time measured from the simulation is 5.8ms while from IP2BMAP is around 5.7ms.

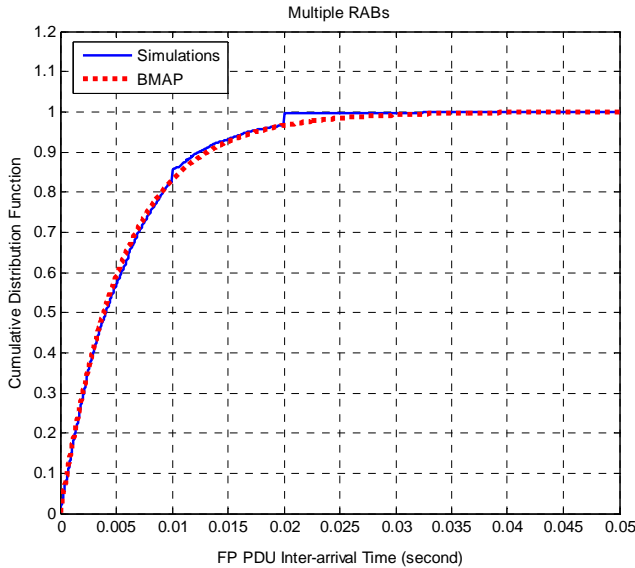


Figure 5. CDF of FP PDU Inter-arrival time (simulation vs. BMAP)

This example proves that the BMAP model can accurately capture the characteristics of the aggregated traffic of UMTS for combination of different traffic types as well as various RAB types. The main benefit of using BMAP model is that it can not only capture the properties of the aggregated traffic on the Iub interface in UTRAN including the self-similar and busty IP traffic, but also to use the BMAP model to generate the trace for simulations can avoid the correlations of the simulation sequences by the same random number seed and at the same time greatly increase the simulation speed and efficiency.

### C. BMAP/D/1

As analyzed in the previous section, the arrival process of the AAL2 queue is modeled using the BMAP. The service rate of AAL2 queue is deterministic due to the AAL2 shaping, and the AAL2 buffer size is unlimited, hence the AAL2 queue can be modeled as BMAP/D/1 queuing system. With this queuing system, the queuing performance given for example in queuing delay or queue length can be derived.

The BMAP trace arrival process is in fact fairly difficult to apply analytically and its important metrics are often calculated numerically and require parametrization when fitted to the measured trace data. The waiting time distribution in the BMAP/D/1 queue can be found in [3], which is very complicated and involves many separate techniques and theories. In this paper, rather than solving the analytic equations for the waiting time, the BMAP/D/1 queue was simulated using the OPNET simulator with a traffic source, a queue and a traffic sink, as shown in Figure 6. The source

traffic of the queue is a trace generated using the BMAP parameterized with real data using IP2BMAP. In the simulated queue system, there is only one server and the service rate is set to deterministic. The queuing performance is presented in the next section.

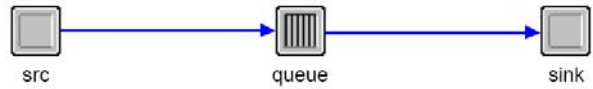


Figure 6. BMAP/D/1 Queue in OPNET

## IV. NUMERICAL RESULTS AND VALIDATION

In this section, we present the numerical as well as simulation results for validation the BMAP/D/1 to determine the AAL2 queuing delay. In the shown example scenario, only a PS RAB 64kbps is used for transmitting the web applications with an average web page size of 25kbyte (Pareto page size distribution). The Iub link is a 2Mbps E1 line and the PVC on the ATM layer is set to 1.55 Mbps Peak Cell Rate. Figure 7 depicts the cumulated distribution function of the FP PDU inter-arrival time for both the simulation and the BMAP trace. It can be seen clearly in this figure that the CDF curve of the BMAP model accurately matches the CDF of the simulation trace. That means, the bursty property of the simulated traffic is correctly represented by the BMAP model.

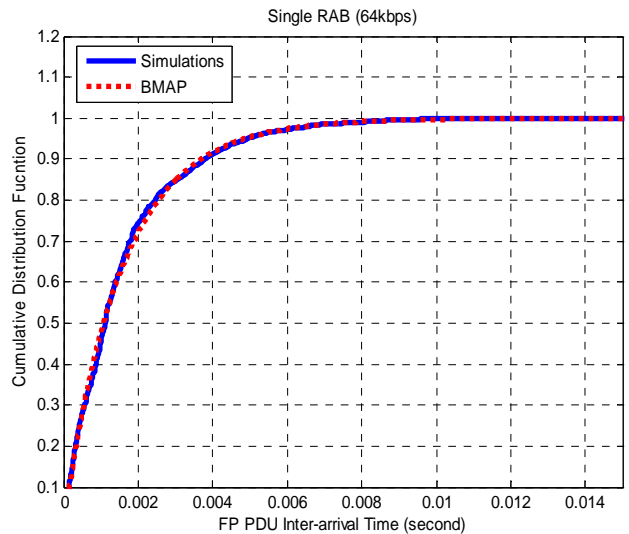
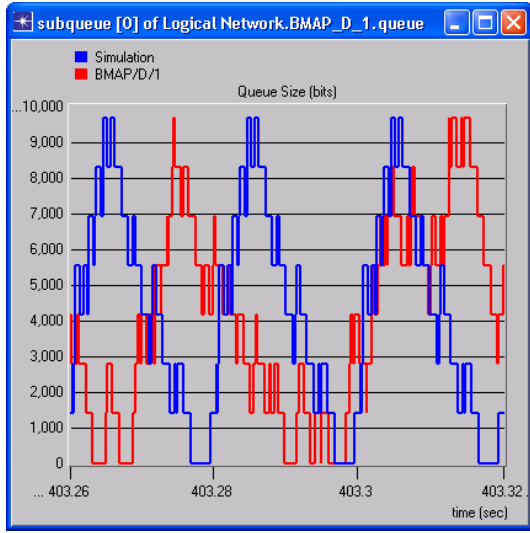
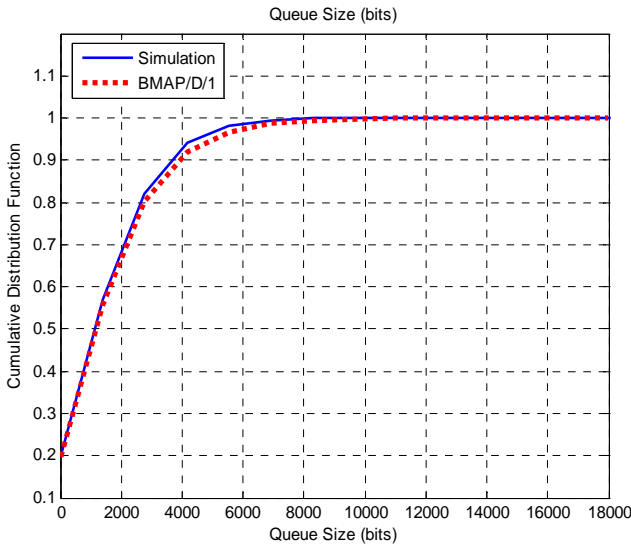


Figure 7. CDF of FP PDU Inter-arrival time (simulation vs. BMAP)

Figure 8 shows the AAL2 queue length in bits obtained from the simulation and from the BMAP trace and their CDF. It is obvious that the BMAP/D/1 queue shows a similar behavior in terms of queuing occupancy. And the obtained cumulative distribution function of the queue size from the BMAP/D/1 model matches correctly with that from the simulation.



(a)



(b)

Figure 8. (a) - AAL2 queue length (simulation vs. BMAP/D/1);  
 (b) - CDF of AAL2 queue length (simulation vs. BMAP/D/1)

As a consequence, the distribution of the AAL2 queuing delays obtained from the BMAP/D/1 model, as shown in Figure 9, is also pretty close to the distribution of the measured AAL2 queuing delays from the simulation. From this example, the BMAP/D/1 model is proven to be an accurate analytical model to calculate the queuing delay in the AAL2 buffer on the Iub link.

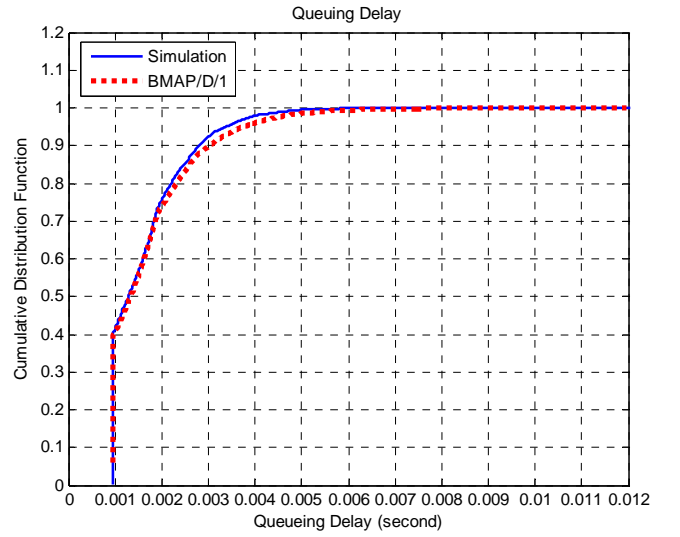


Figure 9. CDF of AAL2 queuing delay (simulation vs. BMAP/D/1)

## V. TRAFFIC SCALING

In this section, we present a scaling technique on the BMAP model to construct a superposition of the source traffic. [6] introduces a scaling method for MMPP (Markov-Modulated Poisson Process). Since MMPP is a special case of BMAP, the same approach is applied by the authors on the BMAP model for constructing a superposition of the traffic. The superposition can be described as follows:

$$\begin{aligned}
 D(0) &= D_1(0) \oplus D_2(0) \oplus D_3(0) \oplus \dots \\
 D(m) &= D_1(m) \oplus D_2(m) \oplus D_3(m) \oplus \dots
 \end{aligned} \tag{5}$$

Where  $D(0)$  represents in BMAP a transient state transitions without packet arrivals and  $D(m)$  ( $m \geq 1$ ) represents a state transition with an packet arrival of batch size  $m$ . Here the operation  $\oplus$  means the Kronecker sum. The following gives an example of how to scale a two - state BMAP model with above equation. The simulation scenario uses a PS RAB 64kbps for transmitting the web applications with an average web page size of 15kbyte. The obtained BMAP parameters  $D(0)$  and  $D(1)$  parameterized from the simulation trace are given below:

$$\begin{aligned}
 D(0) &= \begin{bmatrix} -148.189931 & 46.574224 \\ 87.204874 & -185.366900 \end{bmatrix} \\
 D(1) &= \begin{bmatrix} 62.762956 & 38.852751 \\ 60.902377 & 37.259649 \end{bmatrix}
 \end{aligned}$$

If doubling the amount of traffic from this trace, the new BMAP parameters becomes a four state BMAP model, where the corresponding parameters of  $D(0)_m$  and  $D(1)_m$  are calculated with equation (5):

$$D(0)_m = D(0) \oplus D(0) = \begin{bmatrix} -296.3799 & 46.5742 & 46.5742 & 0 \\ 87.2049 & -333.5568 & 0 & 46.5742 \\ 87.2049 & 0 & -333.5568 & 46.5742 \\ 0 & 87.2049 & 87.2049 & -370.7338 \end{bmatrix}$$

$$D(1)_m = D(1) \oplus D(1) = \begin{bmatrix} 125.5259 & 38.8528 & 38.8528 & 0 \\ 60.9024 & 100.0226 & 0 & 38.8528 \\ 60.9024 & 0 & 100.0226 & 38.8528 \\ 0 & 60.9024 & 60.9024 & 74.5193 \end{bmatrix}$$

With these new parameters  $D(0)_m$  and  $D(1)_m$  the BMAP trace are generated. Figure 10 shows the CDF of the FP PDU inter-arrival time from the regenerated BMAP trace for the doubled amount of the traffic, compared with the CDF of FP PDU inter-arrival time from the simulation trace using the same amount of traffic. It is observed that the scaled BMAP traffic matches the traffic properties of the real simulation trace. This proves that the method of using the Kronecker sum of two traces to scale the source traffic on the BMAP model is accurate.

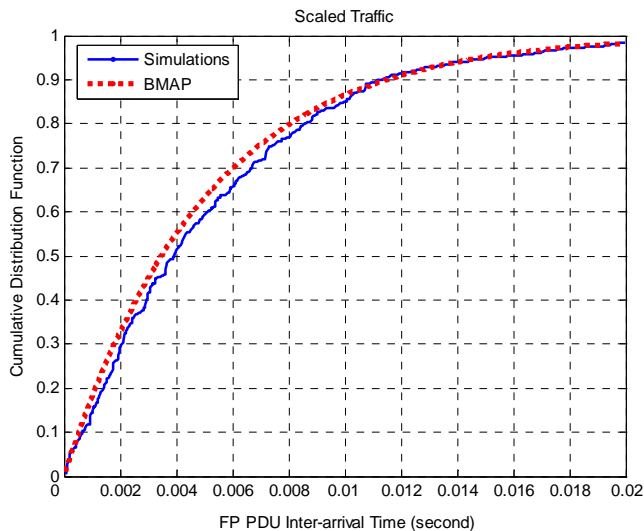


Figure 10. CDF of FP PDU Inter-arrival time (simulation vs. BMAP)

## VI. CONCLUSION

This paper investigates and analyzes the delay over the UTRAN Iub interface. The Iub delay can be represented by the FP PDU delay, which measures the delay of transporting a radio frame across the Iub interface. It consists of all delays that appeared at the Iub interface. It is found that among all delay components, the AAL2 queuing delay contributes the major part of the FP PDU delay, while the ATM cell delay and the additional segmentation and reassembly delay is quite small compared to the AAL2 queuing delay. Based on this analysis, the BMAP/D/1 queuing model is presented in this paper to estimate the AAL2 queuing delay. The proposed BMAP/D/1 model is validated by comparing the calculated results to the simulation results. The other contribution of this paper is the

proposal of a superposition technique on the BMAP model to scale the source traffic. The benefit of scaling the traffic with BMAP is to generate a wide range of traffic loads analytically based on a single simulation trace. The further work is to investigate the network performance of the scaled traffic.

As a summary, based on the results of one simulation, the delay performance on the Iub link can be approximately determined analytically with the BMAP/D/1 queuing model. Furthermore with the proposed superposition on the BMAP model to scale the source traffic, we can generate a wide range of traffic loads based on only a single simulation trace, and the AAL2 queuing delay for all different ranges of traffic loads can be also calculated with the BMAP/D/1 queuing model.

## ACKNOWLEDGMENT

The simulation model was developed in cooperation with Siemens in Berlin, Germany.

## REFERENCES

- [1] Xi Li, Su Li, Carmelita Görg and Andreas Timm-Giel, "Traffic Modeling and Characterization for UTRAN", 4th International Conference on Wired/Wireless Internet Communications, 2006.
- [2] C. Lindemann and M. Lohmann. IP2BMAP software package. Available online: [www.ip2bmap.de](http://www.ip2bmap.de).
- [3] David M. Lucantoni. "The BMAP/G/1 queue", A tutorial. Models and techniques for Performance Evaluation of Computer and Communications Systems, pages 330-358, 1993.
- [4] D. M. Lucantoni, "New Results on the Single Server Queue with a Batch Markovian Arrival Process", Comm. in Statistics: Stochastic Models 7, pp.1-46, 1991.
- [5] SH Kang, YH Kim, DK Sung, and BD Choi, "An application of Markovian Arrival Process (MAP) to modeling superposed ATM cell streams", IEEE Trans. Commun., vol.50, no. 4, pp. 633-642, 2002.
- [6] Tadafumi Yoshihara, Shoji Kasahara, Yutaka Takahashi "Practical Time-Scale Fitting of Self-Similar Traffic with Markov-Modulated Poisson Process"
- [7] Alma Riska, "Aggregate Matrix-analytic Techniques and their Applications"
- [8] A. Klemm, C. Lindemann, and M. Lohmann, "Modeling IP Traffic Using the Batch Markovian Arrival Process (extended version)", Performance Evaluation, 54, pp. 149-173, 2003
- [9] [www.3gpp.org](http://www.3gpp.org)