Traffic Characterization for a UMTS Radio Access Network

Hommad el Allali* and Geert Heijenk†

*University of Twente, P.O.Box 217, 7500 AE, Enschede, The Netherlands
Email: {allali,heijenk}@cs.utwente.nl

†Ericsson EuroLab Netherlands, P.O.Box 645, 7500 AP, Enschede, The Netherlands
Email: geert.heijenk@eln.ericsson.se

Abstract—In this paper we present a UMTS traffic model to evaluate the Measurement Based Admission Control (MBAC) algorithms in a simulator environment. The traffic model describes the traffic aggregation at the Base Station (BS) and Radio Network Controller (RNC) with a Markov Modulated Arrival Process (MMAP). In the model we describe the batch and packet sizes in the Radio Access Network (RAN). Further we identify that the time alignment mechanisms and the Frame Protocol precisely manage the traffic transmission times. As consequence traffic in the RAN consists of batch of packets that are periodically transported over the RAN.

Index Terms—UMTS, RAN, MBAC, QoS, traffic model, MMAP

I. INTRODUCTION

UMTS is the third generation (3G) mobile communication system that is going to support a variety of services and applications using both the circuit and packet switched transport model in its access network. IP is under standardization to be used as transport technology in the Radio Access Network (RAN). Presently, IP has still difficulties to provide reliable service guarantees and fails to meet the stringent RAN requirements. In [1] Resource Management in Diffserv (RMD) is proposed to make IP more suitable for the RAN. RMD is a lightweight resource reservation scheme that can probe the Diffserv domain for resource availability. In RMD two different ways of resource evaluation are defined, namely unit based Admission Control (AC) and Measurement Based Admission Control (MBAC). In our work we investigate the MBAC approach for RMD. This method evaluates the resource availability by measuring the aggregate throughput at the interior nodes in order to have knowledge of the actual bandwidth usage along the transmission path. In order to be able to evaluate alternatives for MBAC, we come up with a characterization of the traffic to be transported in a UMTS RAN.

Up to now several MBAC algorithms have been developed, one group of proposals is: Hoeffding bound [2], Chernoff bound, and Equivalent Bandwidth [3]. These algorithms are based on specific source assumptions. The measured rate sample is used as input parameter in the computed generating function to perform admission control. However, these proposals are not suitable for the IP-based RAN because of the variable bit rate character of the UMTS sources. In UMTS the source behavior is not only dependent on the user behavior but also strongly dependent on the complex RNC management functions. One of the functions is Radio Resource Management (RRM) [4]. This function adjusts the mobile service bit rate every Transmission Time Interval (TTI, e.g. 20 ms) in case the radio interface circumstances changes. This complexity makes it for us very difficult to describe the UMTS source model for the MBAC algorithm. The second group of MBAC proposals is: Measured Sum (MS), Maximum Rate Envelope (MRE) [5], Normal Approximation [6], and Time Scale Composition (TSC) [7]. This group of algorithms uses only knowledge of the measured aggregate behavior to perform AC and are not based on specific source models. From our functional analysis in [8] we concluded that the TSC approach has great potential to perform well in an IP-based RAN. The initial choice for TSC is motivated by the fact that this algorithm is very simple, fast, makes no source assumptions, and is able to track load fluctuations at different time scales.

The next phase of our work is to evaluate the different types of MBAC proposals in a simulation environment to justify our conclusions. In this paper we have a realistic traffic model for an IP-based RAN to improve the TSC approach, and the bandwidth estimation process. Our goal is to use our traffic model in the simulator to have a realistic environment for evaluating MBAC mechanisms in an IP-based RAN. Furthermore, we indicate the important UMTS system parameters to tune the estimation process.

In Section 2 we start with a short description of the UMTS architecture where we discuss the essential system parts. In Section 3 we describe the traffic load in the RAN. Finally, in Section 4 a UMTS traffic model is presented with the belonging UMTS system parameters.

II. UMTS RADIO ACCESS NETWORK

Fig. 1 depicts the network architecture of UMTS. The architecture model is composed of three parts, the Core Network, RAN, and the User Equipment (UE). The UE is connected with the RAN through the Uu interface (WCDMA radio link), whereas the Core Network is connected with the RAN with the Iu interface. The Iu interface has two separated parts: the
Iu-ps (Iu packet-switched), and the Iu-cs (Iu circuit-switched). Each separate Iu part connects the RAN with the PSTN or the Internet. The main network elements in the CN are the Mobile Switching Centers (MSCs) and the GPRS Support Nodes (GSNs). MSCs perform the switching functions of the cellular network and GSNs perform routing of the IP traffic from and to the Internet. The RAN itself consists of two components, the Radio Network Controller (RNC) and the Base Station (BS). The RNC is centralized in the RAN and is linked with a number of BSs via the Iub interface. Moreover, the RNCs are also interconnected with each other via the Iur Interface, which is needed to deal with UE mobility (e.g. soft handover). In the RNC, management functions are implemented to optimize the transport efficiency of the UMTS system. Each RNC in the RAN is responsible for managing and coordinating up to several hundreds of BSs that are located in different cells. The BS, in turn, is in charge of radio frequency and physical radio channel processing. Since packet delay in the RAN is the most important QoS aspect. Both the RNC and BS have the responsibility to transport the received packets in time to the next node. This implies that the packet transportation between the BS and the RNC requires strict packet transmission and receiving times. In UMTS, time alignment techniques are used in the RNC and the BS to support the precise timing between the two elements. This approach will directly result in time synchronization between Uu interface and the RAN.

![Fig. 1. UMTS network architecture](image)

III. UMTS TRAFFIC

The traffic to be transported in an IP-based RAN is the traffic between the RNC performing MAC and higher layer functions, and the BS performing physical layer functions. It is determined by the traffic source, and the MAC and higher layer functions in the UE and its counterpart at the network side, as discussed in Section 3A. Further, the traffic load at an IP-based RAN is heavily influenced by the Frame Protocol (FP), which runs on top of IP in the RAN, which will be discussed in Section 3B.

A. UE Traffic

In the UE, traffic is generated for voice calls, or data calls, or both. Traffic in the UE is processed as follows; the source delivers data to the RLC layer where it will be segmented in fixed portions to fit in the radio packets size. After segmentation, the RLC sends the RLC PDUs to the MAC layer over logical channels. The MAC, in turn, performs scheduling and transport format selection. After this MAC processing the produced MAC PDUs, also called Transmission Blocks (TBs), are sent at every Time Transmit Interval (TTI) frequency over dedicated channels (DCHs) to the physical layer. In more detail, the MAC appends all the TBs of the same Transport Channel (TCh) in one packet and forwards it per TTI to the physical layer. In UMTS, this packet is called Transmission Block Set (TBS) and the TTI is defined as the inter-arrival time of a TBS. The composition of TBs is shown in Fig. 2. TBs for a data service are standardized with a size of 320 bit and the possible TTIs are 10, 20, 40, and 80 ms. It is important to note that the TTI and the number of TBs per TBS must be chosen in such a way that the service rate of the requested service can be supported. Small TTI values are chosen for the real time traffic with stringent delay requirements and a large TTI for best effort traffic. It is interesting to find that the TTI characterizes the inter-arrival-time of MAC PDUs and gives the periodicity of delivering TBS from the MAC to the Physical layer. In addition to DCHs, user data can be sent over the Shared Channel (SCH). It should be noted that the TB and TTI for SCH are standardized with 360 bit and 20 ms, respectively.

For the voice service the traffic is handled differently. The AMR voice codec delivers different type of information bits with different priorities to the RLC. In more detail, dependent on the error sensitivity, the AMR sorts the information bits in three different classes namely, A (TB=TBS=81 bits), B (TB=TBS=103 bits), and C (TB=TBS=60 bits). Class A, is the most sensitive class. To support service differentiation in the air for the different types of classes, the UE simultaneously establishes three coordinated DCHs, like shown in Fig. 2 b. For each of the coordinated DCH a separate Transport and Physical channel is provided. In contrast to the data service the AMR codec delivers voice packets with the proper size to the MAC layer. The MAC traffic processing for voice service is the same as for data service. Since, the voice service has a more stringent delay requirement than the data service, the TTI for the voice service is standardized to 20 ms. Thus, the AMR codec delivers voice packets to the MAC layer per 20 ms. The MAC layer in turn transports also the TBS at each 20 ms time period to the

![Fig. 2. UMTS traffic profile for data (a) and voice (b) service](image)
Physical layer. This means that the MAC maps one TBS over two radio packets. It is important to note that in case the AMR codec changes its bit rate state another compression code setting is chosen while the size of the TBS is kept the same. Only if the AMR codec is in the silent state, it generates noise packets, which result in generation of smaller packet sizes (TBS=39 bits). Thus, during a voice connection the source can switch between two different TBS sizes and the TBS is always filled with information bits.

B. Traffic transport in RAN

Traffic handling in the RAN is mainly characterized by the different network interface requirements. For instance, the main objectives of the RAN is to avoid queuing at the RAN nodes and at the same time fully utilize the scarce radio resources. In UMTS, these goals are achieved by synchronizing the traffic transport channels at each interface. The BS for example is situated between the wireless interface Uu and the wired interface Iub. As a consequence, the BS is the one that takes care of the time synchronization between the two interfaces. Besides the BS, the RNC needs also to be synchronized with the Uu interface in order to deliver the downlink packets to the BS in time. Since the radio packets are transmitted at fixed time intervals of 10 ms, the RAN components must assure that the packets in the downlink direction arrive exactly in time to fully utilize the Uu interface. The synchronization of the BS and RNC with the Uu interface in UMTS is realized by using the so-called Frame Protocol (FP) in combination with time alignment procedures.

Since DCH channels and Common Channels (CCHs) require different transport services in the RAN, two different types of FP packets are defined namely, DCH FP and CCH FP. In general, the DCH FP maps the TBS traffic from one specific DCH into one FP packet. However, for voice connections, the FP maps three coordinated DCHs into one FP packet. The other FP type, the CCH FP maps one TBS from the CCH (e.g. the SCH type) into one FP packet. The FP packet length depends on the number of TBSs, so mobile services with different bit rates result in FP packets with different lengths. After composing the FP packet, the FP-layer sends the FP packet per TTI into the Iub or Iur link.

1) Uplink Traffic:  In the uplink direction (from the UE to RNC), all the traffic from the active UEs in a particular cell aggregates at the BS. The ideal situation at the BS is to receive the TBS of each UE one after the other such that the load over the Iub and Iur is spread out without having to delay the TBSs. To achieve this, the BS uses a frame offset timing mechanism to control the transmission times of the UEs. Each UE in the cell is provided a frame offset that is computed from the load distribution algorithm. The frame-offset value specifies precisely the time when one of the active UEs is permitted to start sending the radio packets over the Uu interface.

Stringent delay requirement force the BS to transmit the TBS per TTI on the Iub interface without a significant queuing delay. More specifically, the incoming radio packets are stored in a small FIFO queue until the whole TBS is received. The BS multiplexes the TBS by using a time framing mechanism with a certain time period T. For each UE one of the T/K slots is assigned during the call establishment. The traffic from the UE is only served in the assigned slot period. After receiving the entire TBS in the allocated slot, the BS maps one TBS into one FP packet. After that, the FP-layer transmits the FP packet per TTI into the Iub link. FP packets that are delayed too much in any of the RAN nodes would immediately be dropped because the receiving codec accepts only TBSs in time. Finally, all the FP traffic from the BSs aggregates at the RNC.

2) Downlink Traffic: An important requirement of the RAN is to achieve full Uu interface utilization without delaying the received packets at the BS. This requires very accurate timing between the BS and RNC and creates stringent queuing requirement on the RAN nodes in the downlink path. The BS uses a time alignment mechanism to synchronize the transmission times of the RNC with the radio timing structure. In case a mistiming takes place, the BS sends a negative acknowledgment to indicate the RNC how early or late the packet was received. The RNC in turn adapts the transmission time so that it assures that RNC FP packets arrive in time at the BS such that they can be transported directly over the allocated radio transmission slot.

For our traffic model, it is interesting to notify that the produced FP packets are transported from the FP-layer across the Iub or Iur link per TTI time. The transmission time of FP traffic over the RAN is thus directly related to the timing characteristics of the Uu link. In our previous explanation we have defined TTI as the transmission frequency of TBS from the UE to Uu interface. Since the FP and time alignment mechanism synchronizes each RAN component with the Uu timing structure, the TTI definition can be redefined in: TTI is the frequency of TBS within the UMTS system (Uu interface, Iu-cs, RAN).

IV. UMTS TRAFFIC MODEL

The traffic description study shows that due to the management functions in the BS and RNC, the BS and RNC can be modeled as a $D^{[X]}/G/\infty$ system. We have seen that the TBSs arrive in groups at this system, where the random batch size X is distributed according to a stochastic process. Each state of the batch size X should correspond to the type of service and the active number of UE’s in the cell. Furthermore, due to the radio framing, the inter-arrival time of the batches would be deterministic. The service times are assumed to be random due to the variable TBS sizes. Since all arriving TBSs are immediately served and experience no buffering delay (only serialization delay) the BS and RNC can be modeled as an infinite server. The next step is to define the arrival process at the RAN Nodes in more detail.

A. Connection model

From the traffic description we have seen that UMTS traffic aggregates at the RAN nodes and consists of various traf-
fic sources each with its own character. To include the behavior of each traffic level, we model the aggregated traffic with a Markov Modulated Arrival Processes (MMAP).

In our model, several connections in a cell may be active. Each connection has a type $i$, where $i \in \{v, d\}$, represents a voice or data connection, respectively. We assume that the arrival of new connections follows a Poisson process with rate $\lambda_i$ and that the holding time of connections is exponentially distributed with mean $1/\mu_i$. Since we consider two different type of connections, the aggregation at the BS can be modeled with two variables $Y=\{n_v, n_d\}$. Where $n_v$ is the number of voice connections and $n_d$ the number of data connections in the system. $n_i$ is limited by $0 \leq n_i \leq N_i$, where $N_i$ is the maximum number of connections of type $i$ that is allowed in the system. In Fig. 3 the connection model is illustrated by a state transition diagram.

In our model we assume that voice users have priority over data users in the sense that the decision to accept voice connections depends only on the number of voice connections that are currently active in the system (acceptance rule: $n_v \leq N_v$). Data connections, besides having some dedicated capacity are allowed to use the residual voice capacity. However, the acceptance of data connections is not limited by the residual capacity. The reason for this is that the data connection has the opportunity to use the shared channel capacity. As consequence $n_v$ and $n_d$ are independent of each other although the actual traffic load of data depends on the number of voice connections $n_v$ as we will see later. Due to this independency the probability of being in state $Y=\{n_v, n_d\}$ of a given number of active voice and data connection is $Pr\{Y_1=n_v, Y_2=n_d\}=Pr\{Y_1=n_v\}Pr\{Y_2=n_d\}$, where [9]

$$Pr\{Y_i=n_i\} = \frac{(\lambda_i)^{n_i} \cdot 1}{\sum_{n_{i=0}}^{N_i} \frac{\lambda_i}{\mu_i} n_i! \cdot n_i \cdot 1}, \quad n_i \leq N_i \quad (1)$$

B. Channel model

After we have modeled the connection level, the next step is to describe the type of channels that are used by $n_v$ and $n_d$. Each time when a new connection arrives, the connection gets a particular channel type assigned. The type of channel can be a DCH or a SCH. A DCH is a channel that is dedicated to one connection and a SCH channel is assigned to one cell, and shared by multiple connections. In UMTS, voice connections use only DCHs, data connections on the other hand can use DCHs or a SCH channel. Since each channel type transports traffic differently, it is essential to distinguish between the DCH and SCH channel connections in the traffic model. The number of voice connections that use a DCH type of channel is equal to $n_v^{DCH} = n_v$. Furthermore, suppose we use for all the voice connection a maximal data rate $S_v$ and for all the data connections that use a DCH a rate of $S_d$. Then the number of data users that use a DCH will be $n_d^{DCH} = \min\{\text{round}\left(\frac{C_{total} - n_v \times S_v}{S_d}\right), n_d\}$, where $C_{total}$ is the system capacity. In case there are no more DCHs available, the other data connections get the shared channel assigned. The number of SCH channel users is $n_d^{SCH} = n_d - n_d^{DCH}$. Let $I_{SCH}=1$ in case at least one of the data connections use the SCH channel ($n_d^{SCH} > 0$) and otherwise $I_{SCH}=0$.

C. Activity model

The activity model, characterizes the active and inactive periods for each channel type. We model this behavior with a two state Markov process. This process alternates between the active and inactive state, the lengths of the two states are assumed to be exponentially distributed with mean $1/\alpha_i$ and $1/\beta_i$, respectively. The active and inactive states represent for voice the talk and silence periods during a conversation. For data the active and inactive states represent periods of user/application (in)activity. The probabilities of being in one of the states is $Pr(0) = \frac{\beta_i}{\alpha_i + \beta_i}$ and $Pr(1) = \frac{\alpha_i}{\alpha_i + \beta_i}$, where 0 and 1 are the inactive and active state, respectively. Given the number of aggregated sources $n_d^{DCH}$ and $n_v^{DCH}$ we define $n_i^{DCH,1}$ and $n_i^{DCH,0}$ as the number of active and inactive sources of type $i$, respectively. The steady state probability distribution of $n_i^{DCH,1}$ is binomial

$$\pi_i(m) = \binom{n_i^{DCH}}{m} \left(\frac{\alpha_i}{\alpha_i + \beta_i}\right)^m \left(1 - \frac{\alpha_i}{\alpha_i + \beta_i}\right)^{n_i^{DCH} - m} \quad (2)$$

D. Packet model

In UMTS, data service bit rates are categorized in the classes: $S[k] = \{16, 32, 64, 128, 384, 768\} \text{[kbps]}$. Each service class can be parameterized with pair $S_d[k] \rightarrow \{j_d[k], t\}$, where $j_d[k] = \{1, 2, 4, 8, 12, 24\}$ is the number of TBs in the TBS and $t_c \{1, 2, 4, 8\}$ is the number of radio frame time intervals within the TTI that are needed to support service $S_d[k]$. Thus, the pairs $\{j_d[k], t\}$ for a service type $k$ are chosen such that the function $r_d(j_d[k], t) = j_d[k] \times TB$ is equal to service rate $S_d[k]$. For data the services bit rates are parameterized with $S_d[0, \ldots, 3] \rightarrow \{j_d[0, \ldots, 3], t=2\}$ and for $S_d[4, 5] \rightarrow \{j_d[4, 5], t=1\}$. The SCH
channel on the other hand uses one type of service character, which is $S_d^{SCH} \rightarrow \{ j_d^{SCH} = 1, t = 2 \}$. Furthermore, voice is standardized in UMTS with the AMR bit rates $S_v[k] = \{ 4.75, 5.15, 5.9, 6.70, 7.4, 7.95, 10.2, 12.2 \}$ [kbps]. In contrast with data services the TBS and TTI values for the $S_v[k]$ bit rates are fixed. The $S_v[k]$ rates are therefore only parameterized with pair $S_v[k] \rightarrow \{ j_v = 1, t = 2 \}$. The next step is to obtain for each type of channel the TBS packet size.

1) Voice: The activity model for voice modulates between two packet sizes, we describe the packet size for each state separately. In case voice is in the active state the packet size is $3TBS = j_v \times TB = 1 \times 244$ bits=244 bits. Note that the voice service uses three coordinated DCHs, which will be combined into one packet. Thus, the voice source produces at every time period of TTI= $t \times 10$ ms=20 ms packets of size 244 bit. In case the connection is in the inactive state, the source produces packets with size $TBS = j_v \times TB = 1 \times 39$ bits=39 bits at every TTI= $t \times 10$ ms=20 ms.

2) Data: In the data packet model it is important to make distinction between the DCH and SCH users. Each type of channel has its own service characterization. Suppose that all the data connections use DCH use service $S_d[k] = \{ j_d[k], t \}$. In the active state, the data source generates traffic with size: $TBS = j_d[k] \times TB = j_d[k] \times 320$ bits at every TTI= $t \times 10$ ms. The other data sources that use the SCH channel produce a total packet size of $TBS = j_d^{SCH} \times TB = 1 \times 360$ bits=360 bits at every TTI= $t \times 10$ ms=20 ms. In contrast with voice, the data source produces no packets in the inactive state.

In Fig. 4 we show that the connections can be grouped in the type of connection, type of channel used, and the activity state. Each individual connection in a particular group has the same traffic size and TTI time (e.g. 320 bit, 20 ms) like we have described in Section D. Because of this, each group can be seen as a deterministic source that generates batches with fixed packet sizes. In other words, the group where data connections use DCHs and are in active state $n_d^{DCH,1}(320,20)$ produces every 20 ms a batch of $n_d^{DCH,1}$ packets of size 320 bits.

E. Aggregate traffic

In the previous sections, we discussed how connections are generated, how they use channels, how modulate between active and inactive periods, and how they generate packets. The resulting traffic is a heterogeneous mix of batch sizes and packet sizes. The traffic experienced by the IP-based RAN is the aggregation of this mix. We can make certain observations on the aggregated traffic. For voice traffic, the total batch size only depends on the number of voice connections, only packet size changes with activity or inactivity. On the other hand, each data connection that alternates from active to the inactive state stops generating TBS packets. As consequence, the batch size changes with the number of active data channels. Another important effect that must be included in the batch size model is the modulation through the switching of data connections between the DCHs and the SCH channel. The third effect is the switching of the service bit rate. In UMTS, it is possible that one data connection uses multiple DCH to support higher service rates, which leads to an increase of the batch size. Due to the complexity we have not taken the adjustment of data bit rates at TTI intervals into account. Finally we can describe the batch size for all the groups. The batch size $X$ that is contributed by voice and data connections can be written as $X = n_v^{DCH} + n_d^{DCH,1} + I_{SCH}$. Finally, the mean traffic rate $R$ is given by

$$E[R] = \frac{244 \text{bits}}{0.02} \sum_{m=0}^{n_v} m \pi_v(m) + \frac{j \times 320 \text{bits}}{t \times 0.01} \sum_{m=0}^{n_d} m \pi_d(m)$$

$$+ \frac{39 \text{bits}}{0.02} (n_v - \sum_{m=0}^{n_v} m \pi_v(m)) + \frac{360 \text{bits}}{0.02} I_{SCH} \quad (3)$$

V. CONCLUSION

We have presented a traffic model that describes the influence of connection arrivals, channel selection, and source activity on the batch and packet size. The resulting traffic can be described as a $D^X/G/\infty$ system. Future work includes the evaluation of various MBAC algorithms for IP-based RANs, using the proposed traffic model.

REFERENCES