Transport Capacity Estimations for Over-provisioned UTRAN IP-based Networks

H. Galeana, R. Ferrús, J. Olmos
Signal Theory and Communications Department
Universitat Politècnica de Catalunya (UPC)
c/ Jordi Girona, 1-3, Campus Nord, 08034, Barcelona, Spain.
email: [hiram.galeana, ferrus, olmos]@tsc.upc.edu, http://www.tsc.upc.edu/

***Abstract*—Within the context of Beyond 3G (B3G) networks, the usage of IP-based transport to support specific Radio Access Networks (RANs) deployments is becoming a feasible solution as an alternative to ATM or TDM networks. In this paper we address the impact, in terms of the required transport bandwidth, of using dedicated and high speed air interface channels on an IP-based UTRAN. In particular, we estimate the transport capacity requirements, for different mean traffic loads, over the Iub interface so that delay bounds imposed to the transport network layer (TNL) are satisfied. We claim that the provided results can be an useful estimation of an over-provisioning solution for the transport capacity required in an IP-based UTRAN.**

Keywords-IP-based UTRAN; over-provisioning; dedicated channels; high speed channels

I. Introduction

Third generation (3G) mobile communication systems continue with an intensive growth, increasing the number of operative networks and subscribers around the world. In particular, the Universal Mobile Telecommunications System (UMTS) represents an evolution in terms of capacity, transmission rate and new service capabilities from second generation (2G) mobile networks. During early stages of 3G deployment, mobile operators reuse as much as possible the existing infrastructure in order to minimize the cost of 2G-3G transition. This is the case of the GSM/GPRS radio access network which consists of a vast infrastructure of point-to-point connections (typically E1) from base stations (BTS) to base stations controllers (BSC). In this context, microwave radio and leased lines are the most used transmission media [1]. Over such a basis, an important upgrade of the capacities provisioned in the transport network is expected in order to handle 3G traffic between the equivalent elements in the UMTS Terrestrial Radio Access Network (UTRAN), namely, the Node B and the Radio Network Controller (RNC).

The transport capacity increment for 3G access is expected to come along with the progressive migration of current 2G traffic to 3G networks. Meanwhile the traffic will continue to increase with data applications like video streaming and high speed web surfing. Another fact that will directly impact on resource utilization of backhaul transmission is the introduction of high speed channels such as HSDPA (High Speed Downlink Packet Access). HSDPA was defined and included in release 5 specifications, and has been designed to support a peak user data rate of over 10 Mb/s. As a result, HSDPA will consequently require large bandwidth in the radio access network [2][3].

Moreover, significant progress has been done in the transport network layer (TNL) of UTRAN [7]. The inclusion of IP as a transport technology facilitates the integration of different radio access technologies operating over a unique backbone and therefore enables the development of heterogeneous networks. On the other hand, it also represents a challenge to the TNL in order to fulfill strict timing requirements, and particularly because IP by itself does not offer any Quality of Service (QoS) guarantees. Furthermore, advanced radio control functions require that the transport of the user traffic over the UTRAN must satisfy stringent delay bounds, regardless if the traffic is real-time or non real-time [4][5]. The IP-based transport network in UTRAN, dubbed as IP-RAN, should meet these requirements in a cost-effective way in terms of efficiency and maximal resource utilization.

Nowadays, when operators are more and more aware that their capital and operational expenses are mostly on the radio access network, bandwidth estimation is an essential strategy to properly dimension UTRAN facilities. With respect to this, it has been argued that over-provisioning of transport resources in the access network is not an economically viable solution [1][2], but to the best of authors' knowledge, there are not references devoted to quantify an over-provisioning solution for the dimensioning of an UTRAN IP-RAN. This paper tries to provide a useful insight into this issue. It is important to remark here that most of the studies dealing with over-provisioning planning in IP networks have been mainly addressed so far to backbone networks [6] and therefore, it is deemed mandatory to have a specific analysis in the context of an IP-based UTRAN due to the particular conditions found there (e.g. different levels of traffic aggregation, characteristics of the applications using the transport, delay restrictions imposed by the radio applications).

In the present work we aim to study the capacity requirements when using an IP-based transport network in UTRAN. In particular, the analysis is carried out for two different scenarios. In the first scenario, traffic is mainly supported by means of Dedicated Channels (DCHs) in the radio interface. In the second scenario, traffic is supported over HSDPA channels. Each scenario imposes quite different conditions and restrictions to the transport network. The remainder of this paper is organized as follows. Section II describes the overall framework of the analysis, as well as the performance metrics. In section III we detail the components and requirements for each scenario. Section IV specifies both traffic and transport network reference models used in simulation setups. Simulation results are presented in section V, and finally in section VI we give some conclusions.

II. STUDY FRAMEWORK

This section provides some key aspects related to the introduction of IP transport in the RAN and elaborates on the dimensioning approach used in this work to assess the required transport capacity for an over-provisioning solution.

A. *IP-RAN topologies*

In our framework, we consider that the transport in the UTRAN is entirely based on IP technology, which means that network nodes (i.e. Node B and RNC) are connected through an IP network responsible for transporting user and control plane, as well as data and O&M information in the UTRAN. Thus, Iub interface [8] between RNC and NodeB is supported over the IP transport. Since standardization of IP transport option is intended to be layer 2 independent, IP transport architecture is limited to nodes implementing an IP layer. So, in an IP-based transport, we can distinguish between end nodes (hosts) and intermediate nodes or routers, responsible for forwarding IP packets. In this sense, a Node B will be usually equipped with an IP host but, in case this Node B serves as an intermediate node in the transport network topology, it will be integrated with an IP router.

With respect to the UTRAN transmission network implementation, different topologies are normally used to interconnect Nodes-B and RNCs [4]. Fig. 1 shows an example of commonly used deployments where we can realize the existence of last mile links, reaching Nodes-B, interconnected in daisy-chain, tree and star topologies, as well as the existence of a high speed backbone network dealing with higher traffic aggregation levels. However, since UMTS specifications do not limit operators to a specific physical infrastructure, other configurations can be implemented, depending on specific operators' requirements. From an economical point of view, topologies allowing a high degree of traffic concentration could result more attractive.

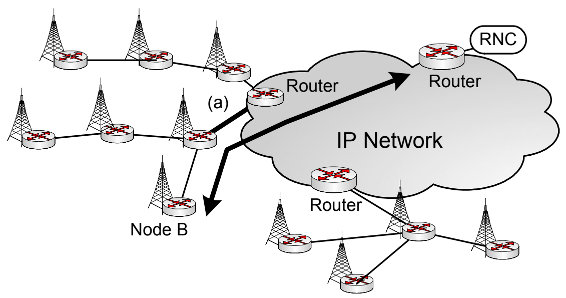

Figure 1. Typical UTRAN interconnection topologies

B. *Dimensioning Approach*

The dimensioning of an IP transport network leads to different approaches depending on the timeframe under consideration. Thus, a long-term analysis (i.e. days/weeks) requires the identification and characterization of those periods (i.e. hours, minutes) with highest traffic peaks (similar to the definition of the busy hour concept in the telephone networks). Then, over those periods, a short-term approach is used to analyze the system dynamics (i.e. the presence of micro-bursts) so that it is possible to estimate the required capacity to prevent queue build-up or excessive delays.

In our study, the long-term characterization relies on the knowledge of the number, and traffic characteristics, of concurrent connections that can be supported, in a given instant, in each Node B of a given UTRAN deployment. Then this traffic information, jointly with network routing information, can be used to calculate the amount of aggregated traffic traversing each link of the transport network. In particular, if we consider any sub-5-minute period, we can state that the *mean* traffic rate supported in a given link can be obtained by the sum of the *mean* values of the traffic generated by concurrent connections traversing that link. Then, for a given number of concurrent connections in a link, or equivalently, for a given mean aggregated bit rate, we can obtain the minimum link capacity required in order fulfilling a given delay bound. The required capacity will be expressed in terms of a parameter known as the "over-provisioning factor", β, which relates the capacity required in the link (C) to the aggregated mean bit rate (R_b) as follows:

$$C = R_b \cdot (1+\beta) = \left(\sum_i R_b^i \right) \cdot (1+\beta)$$

where R_b^i represents the mean bit rate of user connection i traversing over the link. Thus, the required capacity depends on the aggregated mean bit rate, on the traffic pattern characteristics (i.e. statistical properties) of the individual sources and on the considered QoS constraints. Assumming that the main QoS constraint imposed by the radio access network is the delay, we need to set an upper bound on the delay of the link under study. Also, since delay is a random variable, we will assume that the delay bound is met if it is met for 99.9% of the packets.

With respect to the dependency of the proposed approach with the traffic pattern characteristics of the sources, two different traffic models showing quite different dynamics are analyzed: voice traffic and web browsing. Then, for each type of traffic, a detailed characterization of the complete Iub protocol stack is addressed so that the mechanisms used there are reflected into the traffic patterns observed at the transport network. The analysis of the two types of services is done separately, without mixing services. Hence, the obtained results provide the link capacity needed to support a given amount of traffic of a given type of service. Otherwise, assessing capacity requirements in mixed services scenarios it is out of the scope of this work since the problem under those situations is highly dependent on the model used to share or differentiate resources among services. In any case, results obtained here could be applied in the quantification of the resources needed for a given type of traffic in a best-effort network, as well as in a given MPLS (Multi-Protocol Label Switching) path or in a given PHB (Per Hop Behaviour) in Diffserv. Of course, the benefits arising from the statistical multiplexing of sharing different MPLS paths or PHB in the same transport resources are not captured in the model but it still can be used as an upper bound.

In order to establish a reference model we consider a generic topology as the one depicted in Fig. 1. A simple IP-RAN model can be derived from it if we concentrate on a single link, marked (a) in Fig. 1, which connects several Nodes-B to one or more RNCs. Taking into account that all the traffic originated at a known set of Nodes-B is routed through the link under study, the mean aggregated traffic traversing the link can be estimated applying the assumptions mentioned before. Fig. 2 depicts the single path network model for the IP-RAN.

III. Scenario Definition

The studied scenarios, the protocol stacks and the corresponding delay restrictions are explained in this section. Our focus is in the Iub interface defined between NodeB and RNC [8]. The Iub user plane includes various frame protocols (FP), options for the support of random access channels (RACH/FACH), dedicated channels (DCH) and shared channels (HS-DSCH). These latter channels are the ones used in HSDPA.

A. Scenario A: Dedicated Channels over the Iub

The objective in this scenario is to analyze the impact on transport network resources when DCH channels are used in the air interface. DCH channels were introduced in R99 and currently they are commonly used in radio access bearers (RABs) for voice as well as for non real-time data services. Fig. 3 shows the user plane protocol stack for this scenario. The Radio Link Control (RLC) handles segmentation and retransmission of user data between the User Equipment (UE) and the RNC. The MAC layer handles the mapping between the logical channels and the transport channels as well as the selection of the data rates being used. At the output of the MAC layer bursts of Transport Blocks (TB) are generated every Transmission Time Interval (TTI) of the corresponding transport channel. Then, for each DCH channel, the DCH Framing Protocol (DCH-FP) layer assembles the bursts transmitted in one TTI into one FP frame which is subsequently delivered to the IP transport network layer.

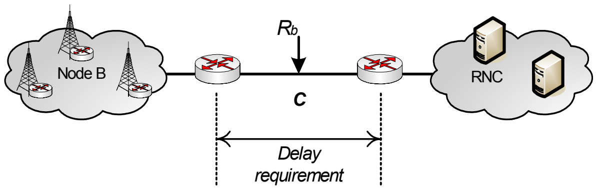

Figure 2. IP-RAN reference model

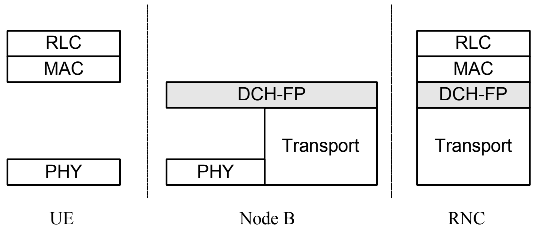

Figure 3. Iub protocol stack for DCH channels

As described in [9], delay in the UTRAN depends on many factors and components such as the processing at each network node, transport network, and radio interface. However, there is no 3GPP specification defining specific delay requirements for the Iub interface. The tolerable delay bounds in the transport network layer for DCH channels are dependent on (1) the delay requirements of the user traffic itself but also on (2) the requirements derived from supporting radio control functions such as outer-loop power control and soft handover. These requirements result in particularly tight delay budgets to be satisfied. For instance, in [13] the acceptable delay value considered for voice services is around 5 ms for 99.9% of transmissions and around 50 ms for data services. In the case of voice, the service itself is the limiting factor, while in the case of data services the radio functions are the limiting factor. These delays are taken as a reference for this scenario as indicated in Table I. Along with the previous values, and in order to assess the sensitivity of the obtained results with the delay constraints being considered, a softer delay restriction (20 ms) is also considered for voice as well as tighter delay restriction (5 ms) for data.

B. Scenario B: High Speed Channels over the Iub

In the scenario *B* we now take into account the use of high-speed channels (HSDPA). The launch of HSDPA radio channels leads to higher data rates in Iub interface, as well as different traffic patterns characteristics in the transport network due to the fact that radio packet scheduling is moved to the Node B. HSDPA introduces new elements in the protocol architecture that have a direct impact on transport network requirements on the Iub interface. Fig. 4 depicts the user plane protocol stack for scenario B.

Unlike R99 where MAC layer was completely located at the RNC, a fast packet scheduling functionality is now introduced at the Node B (MAC-hs for HSDPA). The RNC retains only part of the MAC (MAC-d) mainly to handle logic channel multiplexing. It is worth noting that the RLC layer stays mainly unchanged except for some optimizations for real-time services such as VoIP.

The use of buffering in the Node B permits a peak rate for the connection as high as the terminal and Node B capabilities allow, while keeping the maximum bit rate over Iub in line with the QoS parameters received from the packet core. In fact, having the transmission buffer at the Node B also requires flow control mechanisms to be applied, so that Node B buffer does not overload if radio conditions in the downlink make data to be retained at the Node B. Also in the downlink direction, the Node B buffer shouldn't get empty as long as there are still user data pending for transmission at the RNC. The FP protocol specified to carry HSDPA data in the Iub interface is called the high-speed downlink shared channel FP (HS-DSCH FP). Under this scenario, the delay requirements for HS-DSCH FP frames are mainly due to the service itself since neither outer-loop power control nor soft handover are supported on these channels. According to this, softer delay restrictions can be considered for voice (e.g. 50 ms) and data traffic (e.g. 150 ms). However, attending to potential delay values given in [11] for next release of UTRAN, denoted as Long Term Evolution (LTE), values ranging between 1 ms and 15 ms are accounted for packet transmissions in the transport part of the radio access network. Thus, in accordance with previous arguments, delay upper bounds of 5ms/50ms for voice and of 5ms/150ms for data have been considered in this scenario (see Table I).

IV. SIMULATION DESCRIPTION

A. *Traffic Models*

The voice model consists of a series of ON and OFF periods with a service rate of 12.2 kbps, which corresponds to one of the bit rates achieved by the Adaptive Multi-Rate (AMR) codec specified by 3GPP. ON and OFF states are exponentially distributed with a mean duration of 3 sec. We assume that all users' sessions are kept active during the simulation elapsed time. The background noise description packets sent by the codec during the silence periods are not considered in our analysis.

In the case of data traffic we consider a web browsing model [10]. A web session is modelled as a sequence of packets corresponding to the download of pages. The number of pages in a session is a geometrically distributed random variable with a mean of 5 pages. A truncated Pareto distribution is used to model the packet size of web traffic, resulting in a mean packet size of 366 bytes. The number of packets per downloaded page is modelled by a geometrically distributed random variable with a mean of 25 packets. Packet-calls are separated by an interval (reading time) which is a geometrically distributed random variable with a mean of 10 seconds.

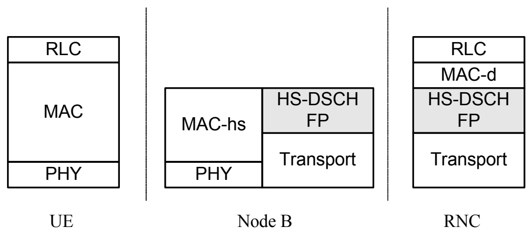

Figure 4. Iub protocol stack for HS channels

TABLE I. TRANSPORT DELAY REQUIREMENTS

Scenario	Transport scheme	Delay upper bound	
		Voice	***Web***
A	Iub with DCH	5-20 ms	5-50 ms
B	Iub with HS-DSCH	5-50 ms	5-150 ms

B. *Iub Interface Modelling*

In accordance to guidelines provided in [7], we consider a modular structure for the Iub interface modelling. Such structure is separated into the following modules: link, IP transport, Radio Protocols/FP, and traffic sources. Fig. 5 is a diagram of the IP-based Iub reference models for both scenarios. Notice that the link and IP transport modules are common for both scenarios and main differences are related to the Radio Protocols/FP functions and traffic model assumptions.

In the modelling of the link, we consider a single queue onto which all concurrent connections are multiplexed, and where the service time is a linear function of the IP packet size. The link model assumes no losses, i.e. the buffers are large enough to accommodate the potential overload.

The IP transport module includes the following components: segmentation, multiplexing queues and packetizer. The segmentation module guarantees that large FP frames are fragmented in order to fit into the maximum container payload. The multiplexing queue retains FP frames from various streams (i.e. user connections), so that the packetizer can arrange several of them into the same IP packet. This process introduces an additional delay to the streams (e.g. FP frames wait in the multiplexing buffer until either there is enough data to build a complete transport packet or a maximum packetizer delay is exceeded). Moreover the following overheads are considered: (1) Overhead/Stream, added to each FP PDU so that several of them can be packet into the same container; (2) Overhead/Container, added to the set of FP frames multiplexed in a single IP packet; and (3) the UDP/IP overhead of the packet to be delivered to the transport.

Modelling of the Radio Protocols/FP block is addressed by means of two main aspects: overheads and queuing. On one hand, overheads values are derived from headers added in the PDCP, RLC, MAC and FP layers, which depend on the service type and on the scenario. In particular, in scenario A, voice traffic is assumed to be supported under the transparent RLC mode whereas web traffic uses RLC acknowledge mode. PDCP usage for header compression is considered for data services. In scenario B, voice traffic is seen as Voice over IP (VoIP) so that PDCP is also used in this case for header compression. We assume that after compression the resulting header size is approximately four bytes [12]. On the other hand, RLC/MAC queuing is introduced to account for the effect of having a maximum bit rate for the DCH channel that, in case of scenario A, should be enforced by the MAC scheduler at the RNC. Notice that in scenario B, we do not model this effect since the assumption here is that data arriving at the RNC can be directly forwarded to the NodeB scheduler (i.e. no RLC/MAC buffer waiting time is considered in the RNC for scenario B).

Finally, we have also introduced in the model the effect of a rate limiter for web traffic. The purpose of a rate limiter is to avoid large traffic bursts reaching the RLC/MAC buffers at the RNC. In a real system, this rate limiter could be located at the gateway of the packet-switched core network.

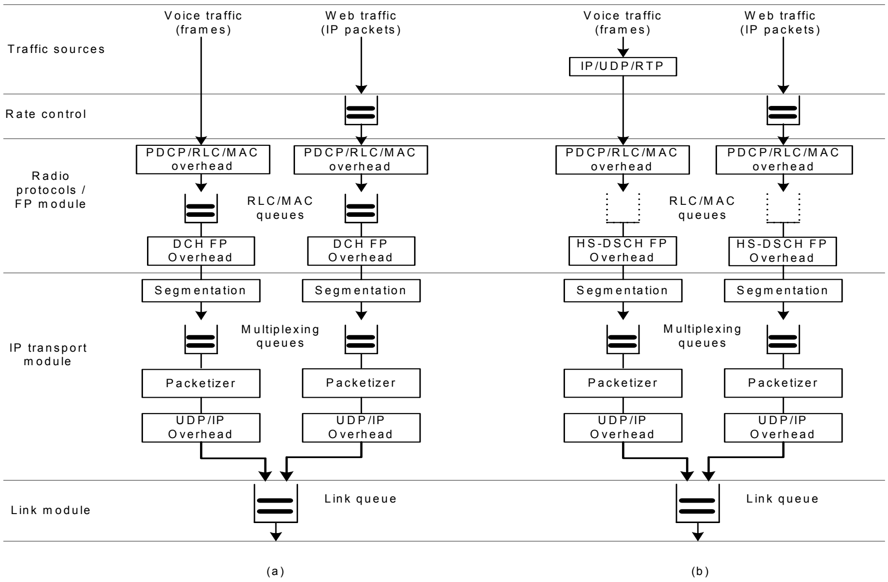

Figure 5. Iub reference models with IP-based Transport: scenario A (a), and scenario B (b)

In our model, although this rate control is depicted in Fig. 5 as an extra queue, this component is easily simulated by adjusting the inter-arrival packet time of web model, which limits web traffic to a maximum data rate. In both scenarios, a rate limitation to 512 kbps has been considered.

Tables II and III list, respectively, the overheads considered for scenario A and scenario B.

V. SIMULATION RESULTS

In this section we present some simulation results obtained with OPNET Modeler under the described framework. A set of simulations has been performed in each scenario, where each simulation is run with a specific mean service traffic load and obtains packet delay percentiles during a 5-minute period (excluding a warming time needed to stabilise the mean load).

The results are given in terms of the extra capacity required in order to support the mean traffic load and to fulfil the transport network delay requirements. Accuracy in the obtained values is estimated around 2 %. Results are presented in four graphs, which correspond to the two services (voice and data) per scenario. We show the extra capacity required in terms of the over-provisioning factor, β, in percentage for a given mean rate value in the traffic aggregate. The legend of the figures relates the colours of the bar graphs to the considered delay bounds in ms.

Figs. 6 and 7 show the results for dedicated channels (scenario A). It can be seen that while β values around 40-60% may suffice to support voice traffic, values as high as 165% are required for more bursty data services like web browsing. DCH channels rates of 256 kbps have been considered for web traffic. As expected, the degree of over provisioning decreases for higher traffic aggregates. However, unlike web traffic, voice traffic shows less drastic changes between different traffic loads.

Figs. 8 and 9 show the results when high speed channels are used (scenario B). We notice a significant increase in the over-provisioning factor, with respect to scenario A, for both voice and data services. In the case of voice, the increase is mainly due to the larger overhead incurred by the support of VoIP. Notice that the mean bit rate values indicated in the abscissas axis of the graphs only account for voice frames. On the other hand, in the case of data, quite different situations can be envisaged. Focusing on Fig 7 and Fig 9, we can observe that, when considering the same delay constraint, the overhead in scenario B is higher. This is due to the high variability of traffic injected by the RNC into the transport because, unlike scenario A, there is no smoothing effect due to the RLC/MAC queuing.

TABLE II. TRAFFIC OVERHEADS IN SCENARIO A

Module	**Component**	**Overheads**	
		Voice	***Web***
Radio Protocols/ FP	PDCP/RLC /MAC	0 bytes	2 bytes
	DCH-FP overhead	8 bytes	5 bytes
IP Transport	Stream / Overhead	3 bytes	3 bytes
	Container / Overhead	8 bytes	8 bytes
	UDP/IP overhead	28 bytes	28 bytes

TABLE III. TRAFFIC OVERHEADS IN SCENARIO B

Module	**Component**	**Overheads**	
		Voice	***Web***
Traffic Source	IP/UDP/RTP[a]	4 bytes	N/A
Radio Protocols/ FP	PDCP/RLC /MAC	2 bytes	2 bytes
	HS-DSCH FP overhead	10 bytes	10 bytes
IP Transport	Stream / Overhead	3 bytes	3 bytes
	Container / Overhead	8 bytes	8 bytes
	UDP/IP overhead	28 bytes	28 bytes

a. VoIP is considered when offering voice service in HSDPA

On the other hand, if scenario B is operated with higher delay bounds, the required capacity can be less than in scenario A. Although it could seem an unfair comparison, notice that the delay bound in scenario A is not due to the service itself but to specific radio functions within the FP-DCH protocol. Contrarily, this limitation does not exist in scenario B so that softer delay requirements can be applied whenever the final service is not deteriorated.

VI. CONCLUSIONS

This paper explores the option of over-provisioning radio access network in order to satisfy traffic load conditions and delay requirements in the transport network layer. Factors like the rapidly increasing volume of traffic in current 3G networks, as well as new enhancements to the system (e.g. HSDPA) directly influence the required transport capacity in the UTRAN. The results show the degree of over-provisioning that would be needed to support the Iub, with its stringent delay requirements, over an IP-based backhaul network.

The over-provisioning factor is mainly influenced by the inherent nature of the dynamics of the traffic injected and by traffic overheads. In this sense, web traffic results are mainly impacted by the former, thus requiring large values of extra capacity in order to cope with high traffic fluctuations. On the other hand, voice traffic is more stable (i.e. low dynamism) and therefore the most part of β exclusively depend on protocol overheads of the Iub interface.

ACKNOWLEDGMENT

This work is part of the EU IST-AROMA project [14], partially funded by the European Commission under the IST framework. The work is also partially funded by the Spanish government in the CICYT project TIC2004-00518.

REFERENCES

[1] A. Bolle, H. Herbertsson, "Backhaul must make room for HSDPA", Wireless Europe, January 2006.

[2] The Business Case for GSM and UMTS Backhaul Optimization, www.cisco.com/go/mobile, 2005.

[3] H. Holma, A. Toskala, "HSDPA/HSUPA for UMTS", Ed. John Wiley & Sons, 2006.

[4] A. Samhat, T. Chahed, "On the QoS-capable transport in the IP-based UTRAN", 2nd International Symposium on Wireless Communication Systems, pp. 318-321, Sept. 2005.

[5] Y. Guo, Z. Antoniu, S. Dixit, "IP Transport in 3G Radio Access Networks: an MPLS-based approach". Proceedings of WCNC'02 Conference, March 2002.

[6] C. Fraleigh, F. Tobagi, C. Diot, "Provisioning IP Backbone Networks to Support Latency Sensitive Traffic", IEEE INFOCOM 2003, San Francisco, March 2003.

[7] 3GPP, TR 25.933 v5.4.0, "IP Transport in the UTRAN (Release 5)".

[8] 3GPP TS 25.430 v7.0.0, "UTRAN Iub interface: general aspects and principles (Release 7)"

[9] 3GPP, TR 25.853 v3.0.0, "Delay Budget within the Access Stratum"

[10] 3GPP TR 101.112 V3.2.0, "Universal Mobile Telecommunications System (UMTS); Selection procedures for the choice of radio transmission technologies of the UMTS", April 1998.

[11] 3GPP TR 25.912v7.0.0, "Feasibility study for evolved Universal Terrestrial Radio Access (UTRA) and Universal Terrestrial Radio Access Network (UTRAN) (Release 7)".

[12] S. Casner, V. Jacobson, "Compressing IP/UDP/RTP Headers for Low-Speed Serial Links", RFC 2508, February 1999.

[13] MWIF "IP in the RAN as Transport Option in 3rd Generation Mobile Systems", MTR-W-5 Technical Report. Rel. v2.0.0. June 2001.

[14] IST-AROMA project web page http://www.aroma-ist.upc.edu/.

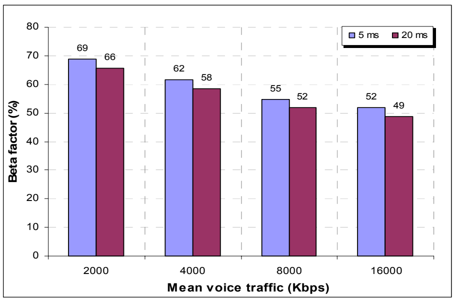

Figure 6. Over-provisioning factor for scenario A with voice traffic

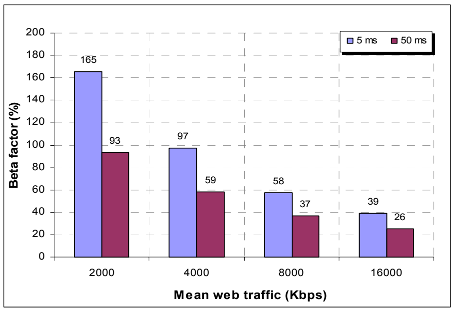

Figure 7. Over-provisioning factor for scenario A with web traffic

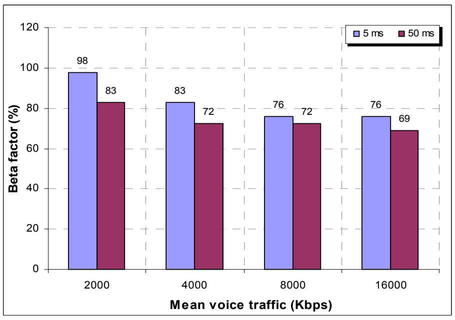

Figure 8. Over-provisioning factor for scenario B with voice traffic

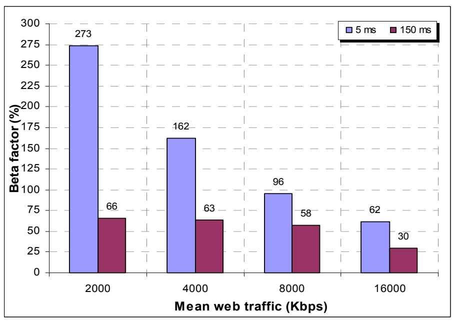

Figure 9. Over-provisioning factor for scenario B with web traffic