TSG-RAN Working Group 3 meeting #2
Stockholm 15th – 19th March 1999

TSGW3#2(99)168

Agenda Item: 7.1, UTRAN Architecture (S3.01)

Source: Siemens, Italtel

Title: **UTRAN Delay Estimation**

Document for: Information

Abstract

This proposal seeks to determine transmission delays across the UTRAN. A study has been undertaken to examine the causes and effects of delay. From this, some UTRAN protocol layers have been identified as adding considerable delay, while others have been demonstrated to be delay sensitive. Once the delay components have been identified a proposed maximum delay for a branch is concluded.

Acknowledgements

Some of the results presented in this paper are a result of work carried out under the ACTS project EXPERT AC094.

Introduction

UMTS Network Delay

The UTRAN can be broken down in to several delay components, the processing delay in the nodes, delay in the transport network, and the delay over the air interface. Within this paper these delays are examined and initial estimates about the delay performance are given.

The radio frames of a radio access bearer service experience different transmission times on diversity branches between the UE and the SRNC. Especially in case of inter-RNC soft handovers both the absolute data delay of different diversity branches in the same connection can be significant.

Figure xx.1 shows a scenario of an inter-RNC soft handover that is seen as one possible scenario with maximum absolute delays and delay difference between data of different diversity branches.

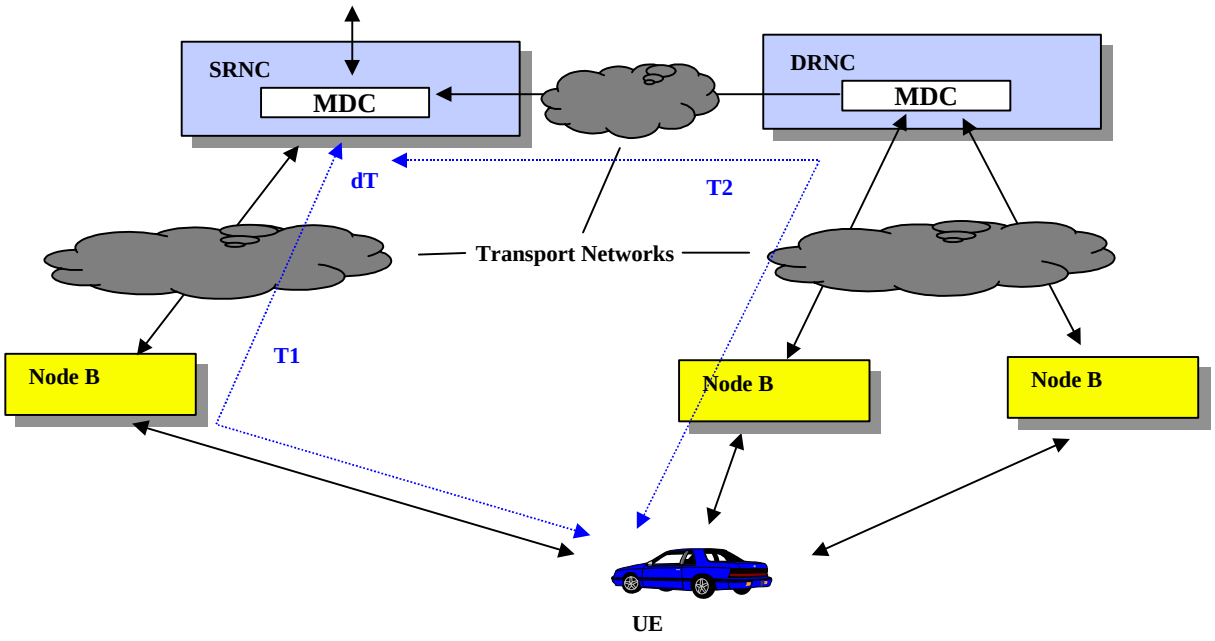

Figure xx.1: Transmission delays on distinct diversity branches in case of inter-RNC soft handover

UMTS Protocol Stacks

Branches T1 and T2 in figure xx.1, follow different routes and so will have different performance. Branch T2 has an additional RNC and transport network to pass, and should therefore have a greater delay. The protocol stack of branch T1 is shown in figure xx.2.

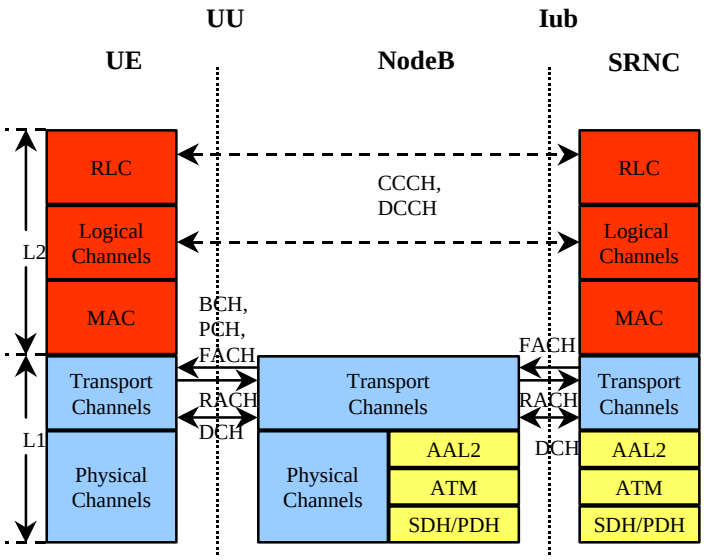

Figure xx.2: Protocol stack of route T1.

In figure xx.2 above, branch T1 represents the user-data protocol stacks that interconnect the mobile with the terrestrial network. The diagram shows directly connected ATM links. However a third party network can be used between UTRAN nodes that have an indeterminate amount of switches. In figure

xx.3, branch T2 is presented in terms of the user-data protocol stack. Branch T2 has an additional RNC in the transmission path and may have an unknown amount of switches over the Iub and Iur interface.

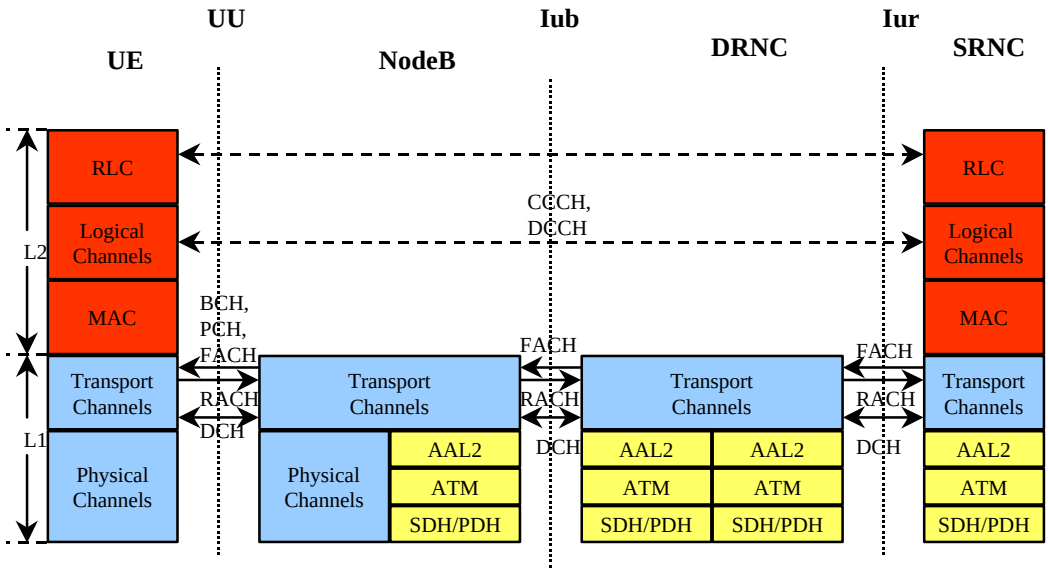

Figure xx.3: Protocol stack of route T2.

The author has shown the directly connected case where the UTRAN components interface to themselves directly. However, a “transport cloud” may exist between the UTRAN entities. It is up to the UTRAN operator to specify the performance characteristics required in the ATM traffic contract.

Transport Network

ATM Transport Capabilities

The transport network interconnecting the entities of a UMTS network is ATM, which is based on hard, fast switching techniques. ATM is the leading network technology that has the ability of differentiate between traffic categories. The categorisation of traffic into ATM transfer capabilities (ATCs) is done through buffer management strategies. These methods include time and space priority mechanisms, and feedback control, see [BRIE98].

Three parameters: Cell Transfer Delay (CTD); Cell Delay Variation (CDV); and Cell Loss Ratio (CLR) measure ATM network performance. With the introduction of buffer management techniques the overall utilisation of an ATM network is increased, while maintaining network performance guarantees for real-time traffic. However, this causes some traffic categories to have unpredictable performance. Thus, speed performance parameters are only specified to real-time traffic categories and the accuracy parameter to all services except best effort (UBR). Table xx.1 summarise performance measures to ATCs, more detail can be found in [ATM056]

	CBR	rt-VBR	nrt-VBR	ABR	UBR	ABT
CTD	Specified	Specified	Unspecified	Unspecified	Unspecified	N/A
CDV	Specified	Specified	Unspecified	Unspecified	Unspecified	N/A
CLR	Specified	Specified	Specified	Specified	Unspecified	N/A

Table xx.1: The ATCs with their respective performance parameters

Transport Network Performance Degradation

When a cell stream is introduced to an ATM network, small performance degradation will certainly occur. This performance degradation is dependent on the type of media; switching mechanism and associated buffer management strategy; and the load and characteristics of other users.

Table xx.2 highlights the causes of performance degradation in ATM.

	CTD	CDV	CLR	CER	SECBR	CMR
Propagation Delay	✓					
Bit Error Statistics			✓	✓	✓	✓
Switch Architecture	✓	✓	✓			
Buffer Capacity	✓	✓	✓		✓	
Traffic Load	✓	✓	✓			✓
Number of Nodes in Tandem	✓	✓	✓	✓	✓	✓
Resource Allocation	✓	✓	✓			
Failures			✓	✓	✓	

Table xx.2: Degradation of Network Performance.

Discussion

Causes of Delay in the UTRAN

The UTRAN protocol stack and associated functions has been analysed to determine causes of delay. Every layer in the protocol stack will have some processing delay associated with transmission. However, some layers will effect the performance more than others. While some of the functions in the UTRAN architecture add delay, others are delay sensitive. Thus, by identifying the causes and effects of delay, the architecture can be enhanced to provide the best overall performance in terms of usability and speed. The layers with the greatest effects have been identified and are highlighted in the following section.

Main causes of delay are:

- Packetisation/de-packetisation of user data streams in the end-system
- Removal of delay variation with “play-out” buffering.
- Turbo coding and interleaving in the physical layer
- AAL2 PDU routing and AAL2/5 packetisation delay
- MAC Scheduling delay

- Transport network delay

Delay sensitive functions are:

- Power Control
- RLC Acknowledgement control loop

The locations of these delays are highlighted in figure xx.4. For clarity only one side of the peer protocol has been highlighted.

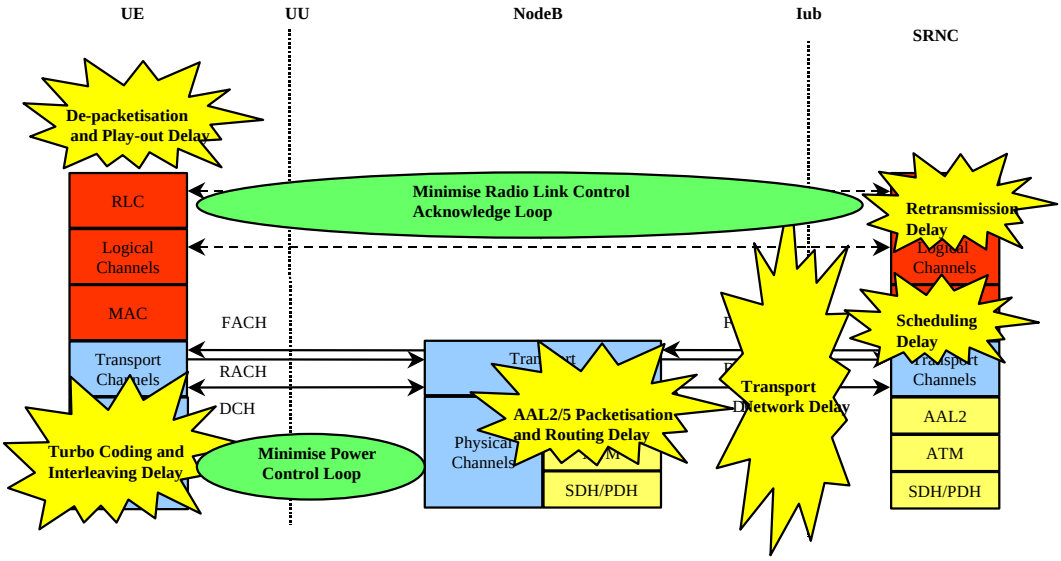

Figure xx.4: Location of delay causing and delay sensitive functions

Estimation of Delay Component

UTRAN Network

De-packetisation and End-System Play-Out Delay

When a real time CBR data stream terminates at an application end-point, play-out buffering is required to remove the variation in delay caused by the statistical sharing effects of the packet network. Once this variation is removed, the resulting traffic stream from the protocol stack can be fed to higher layers as a constant stream of data.

The CDV in ATM networks is removed in terminating end-point by play-out buffers. In addition, like packet networks, both the originating and terminating end-points will introduce packetisation and de-packetisation delay. This delay is dependent on the bit-rate of the connection and the packet size. In figure xx.5 the graph show the cell delay caused by ATM end terminals.

Using figure xx.5, it can be seen that the originating terminal will add a packetisation delay (this being based on ATM AAL1 type networks). In the terminating endpoint will add a de-packetisation and play-out buffer delay.

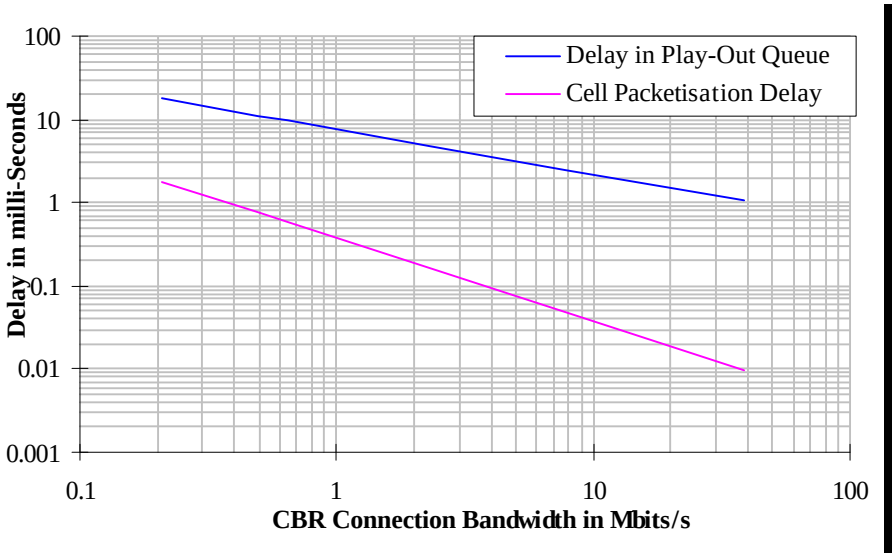

Figure xx.5: Delay within a CBR End-Terminal.

Interleaving and Turbo Coding

Interleaving is a physical layer function that segments transport blocks over several radio frames. These blocks can be interleaved over 1, 2, 4, and 8 transport blocks. Thus, the interleaving will add a large delay to the data stream. Thus delays due to interleaving will be 20, 40, 80 or 160 ms depending on the number of transport blocks used.

Turbo coding has it own internal interleaving mechanism, for real-time services this is thought to be an additional delay of 10 ms.

MAC Scheduling Delay

At the moment this is difficult to estimate. Although, for real-time services a single code or resource unit will be allocated on a deterministic basis. This implies that a delay no bigger that one transport block is foreseen. With the introduction of non-real-time services this scheduling delay may become important. However, it is desirable to have a small delay, but as non-real-time services suggest, delay guarantees will not be applicable.

Re-transmission Delay

It is thought that the retransmission of data streams will not take place over real time bearers. When retransmission is used in non-real time services, guaranteed delivery over the radio interface is performed by the RLC. The physical layer delay can be quite considerable and the amount of

retransmissions needed for a single transport block is a multiplication factor for delay, i.e. if it take two re-transmission to transfer a transport block successfully then twice the physical layer delay would be added. In addition, if data transmission has to halt for transport block acknowledgement, the throughput can be considerably reduced.

Transport Network

AAL1/2/5 Packetisation and De-packetisation Delay

A simulation was made of ½ rate coded voice connections to examine the packetisation delay across a 2Mbits/s ATM link. The study examined different types of AAL and different connection configurations; i.e. VC or VP connected traffic streams. The source and destination were directly connected and the packet transfer delay in one direction was determined. Figure xx.6 shows the average transfer delay over a single link.

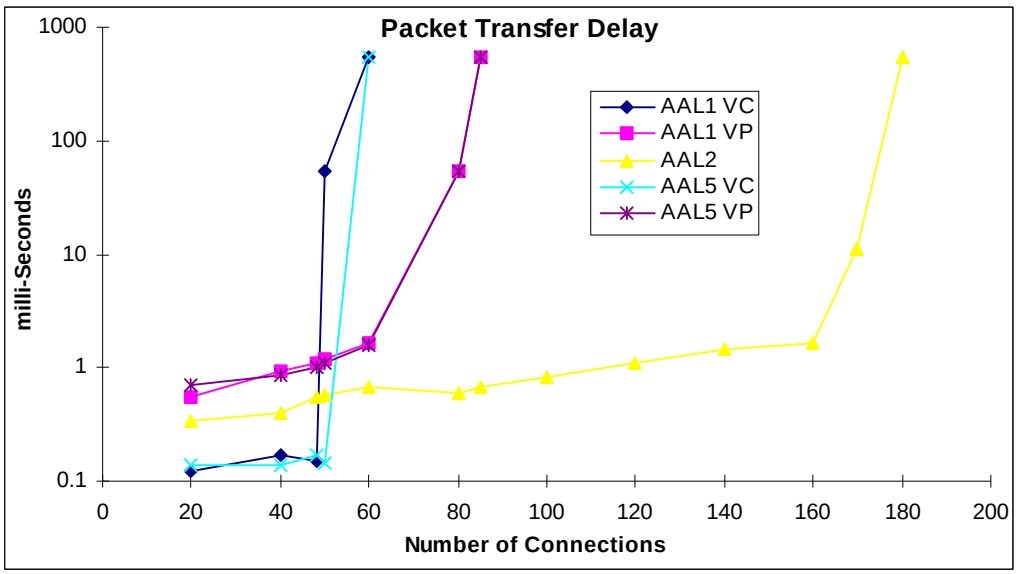

Figure xx.6: Packetisation delay of a 2Mbits/s link using AAL 1/2/5.

From figure xx.6 the packetisation and de-packetisation on to a 2 Mbits/s links is in the order of 1 ms.

Media Delay

The propagation delay over cabled networks can assumed to be fixed. The commonly assumed delay measurement is 5µs per kilometre.

Switch Delay

The ITU-T has defined a delay for real-time services through an ATM switch, [I.356]. This delay is 300 µs. The ITU-T is rather vague over the definition of this delay and the author of this paper assumes that 300 µs is a maximum delay for a switch with 155.52 Mbit/s ports.

Several switches were tested and the results for the minimum and maximum delay is given in table xx.3

Switch	Min Delay	Max Delay	Estimated Queue Length
AT&T, RUM	57.25 µs	140 µs	30 cells
Philips, LaTEX (Pure ATM)	47.03 µs	188.80 µs	52 cells
Philips, LaTEX (SDH)	151.31 µs	314.89 µs	61 cells
Alcatel, ALEX	127.46 µs	498.24 µs	140 cells
Ascom, AAU	104.28 µs	769.51 µs	254 cells
Fore, ASX-200	68.84 µs	766.78 µs	293 cells

Table xx.3: Switch delay measurements

The minimum delay demonstrates a switch with a low load and hence the delay is due to cell processing only. The switch under heavy load with a high CLR causes maximum delay. Hence, the delay is caused by cell processing plus the queueing delay. The results highlighted in grey are from ATM switches with configurable buffer lengths. These switches exceed the ITU-T recommended values when their buffers are used at the maximum limit, however, by reducing the buffer size reduces the switch delay

Network Delay

To gain an overall appreciation of ATM network performance this section examines the performance of a network of real ATM switches. A long-distance ATM connection was studied in [DEL06], [DEL10] [DEL15]. The connection used international ATM links across Europe to determine a real performance value for ATM. In [DEL10] a further intercontinental connection was constructed and measured.

The international ATM connection was to be divided into portions as described in [I.356]. Figure xx.7 show the international ATM connection apportioned. From this the ITU-T recommended CDV and CTD can be calculated.

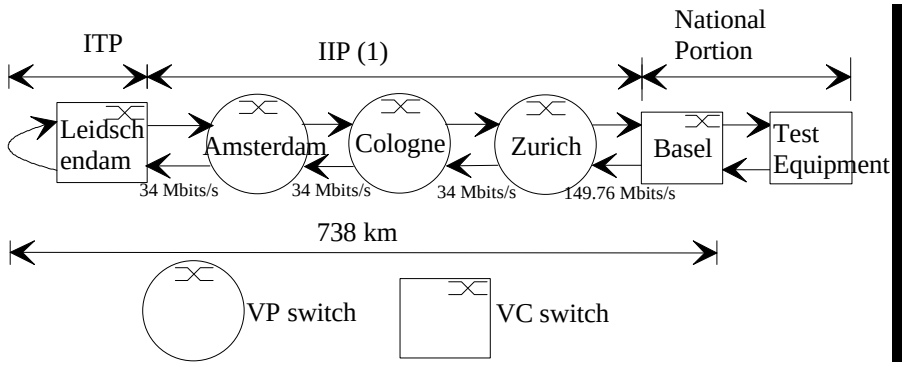

Figure xx.7: Basel-Leidschendam-Basel Portioned

The estimated delay, according to the ITU-T and the measured delay are presented in figure xx.8. The ITU-T specify average delays on connections, as in figure xx.7, to be 21.3 ms. The average delay measured over this connection was 17.6 ms.

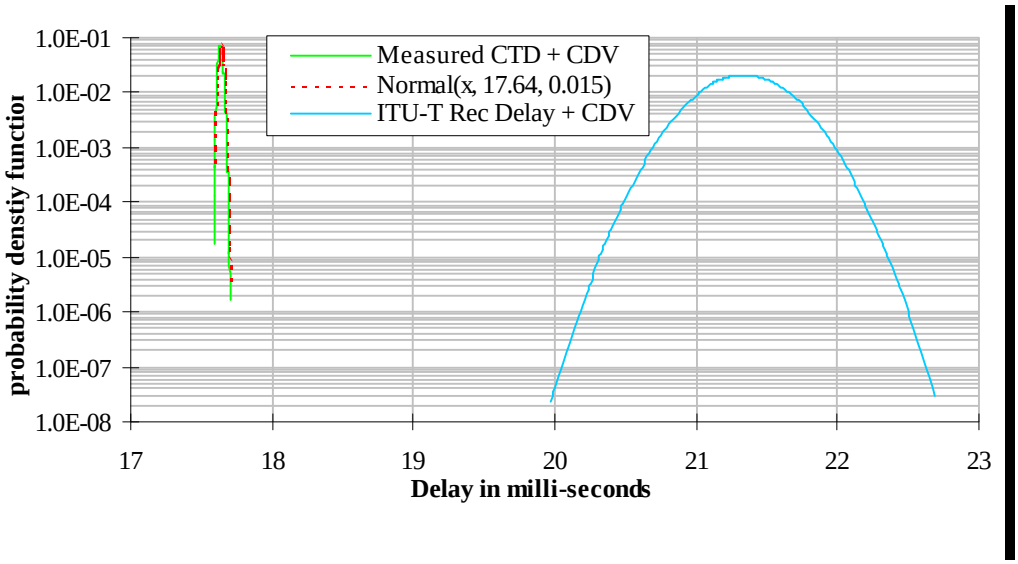

Figure xx.8: Cell Transfer Delay Basel-Leidschendam-Basel

Distance and complexity cause the degradation of performance according to the ITU-T. Therefore, as more juridical boundaries are crossed a greater distance and increased number of switches are included in the estimation

Delay Sensitive Components

Power Control

It is felt that in the FDD system fast power control is obligatory. Power Control Indications will be included in the radio frames at the Node B. As the PC bits are added after interleaving the round-trip delay of the power control signal will be small.

RLC Loop Delay

The RLC loop delay is thought to be a delay sensitive function of the network. The smaller the loop delay the quicker the acknowledgement can be confirmed the receipt of data packets. With larger delays, any loss of data may make the re-transmitted packet invalid because of latency criteria. Using smaller delay loops, more control can be applied and more efficient use of radio resources can be made. This will lead to an increased throughput of traffic.

Proposal

Using the above measurements the delay across the UTRAN can be estimated. Examining branch T1 in figure XX.1 the expected delays caused by each component is as follows:

Delay Location branch T1	Delay in milliseconds
Turbo Coding	10
Interleaving	40
Uu Delay	.05
Iub AAL2 Packetisations delay (2 Mbits/s Link)	1
Iub Transport Network Delay (assume 5 loaded switches over 50km)	2.5
Macro Diversity Delay	1
Removal of Packet Delay Variation	20
Packetisation de/packetisation of transport block	2
Total	Approx. 77 milliseconds

Table xx.4 Estimated delay on branch T1

Thus, on branch T1 and expected delay between data entering layer 3 in the RNC to leaving layer 3 as a constant data stream in the UE is 77 milliseconds.

Delay Location branch T2	Delay in milliseconds
Turbo Coding	10
Interleaving	40
Uu Delay	.05
Iub AAL2 Packetisations delay (2 Mbits/s Link)	1
Iub Transport Network Delay (assume 5 loaded switches over 50km)	2.5
Iur AAL2 Packetisations delay (2 Mbits/s Link)	1
Iur Transport Network Delay (assume 5 loaded switches over 50km)	2.5
Macro Diversity Combining	1

Removal of Packet Delay Variation	20
Packetisation de/packetisation of transport block	2
Total	Approx. 80 milliseconds

Table xx.5 Estimated delay on branch T2

With the increase in branch length an additional 3 milliseconds is added to the path. This result highlights the greatest delay in the UTRAN network to be the physical layer functions over the air interface. The delays reported in this paper examine the static delays across the network. The signalling delay of the branch addition, reconfiguration or deletion was not examined, when taking signalling delays into account an increase in delay is likely.

References

[ATM056] ATM Forum Recommendation, Traffic Management Specification Version 4.0. AT-TM-0056.000, April 1996.

[BRIE98] U.Briem, E.Wallmeier, C.Beck, F.Matthiesen. Traffic Management for an ATM Switch with Per-VC Queueing: Concept and Implementation, IEEE Communications Magazine January 1998

[DEL06] S.B.Winstanley (Editor), Deliverable 06: Specification of Inetgrated Traffic Control Architecture. WP4.1/WP4.2 AC094, September 1996

[DEL10] S.Peeters, K. Spaey (Editors) Deliverable 10: First Results from Trials of Optimised Traffic Control Features. WP4.1/WP4.2 AC094, March 1997

[DEL15] S.B.Winstanley (Editor) Definition of Optimum Traffic Control Parameters & Results of Trials, WP4.1 AC094 December 1997.

[I.356] ITU-T Recommendation I.356, B-ISDN ATM Layer Cell Transfer Performance, May 1996, Geneva.