TSG-RAN Working Group 3 meeting #7 **_TSGR3#7(99)C05_**
Sophia Antipolis, France
21st – 24th September 1999

Agenda Item: 6.5

Source: Siemens, Italtel

Title: Study Item (ARC/3) "Overall Delay Budget within the Access Stratum"

Document for: Status report

---

1 Introduction

This contribution reports the status of the e-mail discussion on Study Item ARC/3 "Overall Delay

2 Report of the reflector activity

During the last month Tdoc. TSGR3#6(99)A18 by Vodafone has been commented.
Even though this contribution was not discussed, some valuable information contained did not receive negative comments on the reflector and has been included in the present report.
The discussion about Tdocs. TSGR3#6(99)949, TSGR3#6(99)955 and TSGR3#6(99)999 was not concluded at last meeting, for this reason the figures derived in these documents have not been included.

The following changes have been brought to the Delay Component Description and to the Delay Template since the last version (see Appendixes):

- Reference GSM 03.05 added
- Macro-diversity Combining Delay does not need being evaluated
- Interleaving and Turbo-Coding component evaluated for speech services
- Added description about Re-transmission Delay evaluation
- New Radio Interface Propagation sub-chapter added
- New Processing Delay sub-chapter added
- AAL Packetisation, Mux and De-Packetisation Delay sub-chapter enhanced
- New informative sub-chapter about UTRAN external Delay Components added
- New Processing Delay component (U2) added in the template
- Template updated accordingly to the changes introduced

3 Next Action Items

The following components must be still analysed to finalise the report:

- Packetisation, de-packetisation and end-system play-out delay for NRT services (U1)
- Interleaving for NRT services (U3)
- MAC scheduling delay (U4) for NRT services
- Re-transmission delay (U5) for NRT services
- AAL packetisation, multiplexing and de-packetisation delay (TN1)

It is suggested to leave the Study Item open to allow the evaluation of the still missing figures.

Appendix A

1 Scope, Abbreviations, etc.

2 References

[1]: Selection procedures for the choice of radio transmission technologies of the UMTS (UMTS 30.03 version 3.2.0) - TR 101 112 V3.2.0 (1998-04)

[2]: Requirements for the UMTS Terrestrial Radio Access system (UTRA) (UMTS 21.01 version 3.0.1)- TR 101 111 V3.0.1 (1997-10)

[3]: Quality of Service and Network Performance (UMTS 22.25 version 3.1.0) - TR 22.25 V3.1.0 (1998-03)

[4]: ITU-T Recommendation G.174: Transmission Performance Objectives for Terrestrial Digital Wireless Systems using Portable Terminals to access the PSTN (6/94)

[5] ITU-T Recommendation G.114: Transmission Systems and Media General Characteristics of International Telephone Connections and International Telephone Circuits - One-Way Transmission Time (02/96)

[6]: Technical characteristics, capabilities and limitations of mobile satellite systems applicable to the UMTS (UMTS 30.20 version 3.1.0) - TR 30.20 V3.1.0 (1998-01)

[7]: ITU-T Recommendation I.356: Overall network aspects and functions – Performance objectives (10/96)

[8]: ATM Forum af-vtoa-0113.000, *ATM Trunking using AAL2 for Narrowband Services*, (2/99)

[9]: ITU-T Recommendation I.363.2, *B-ISDN ATM Adaptation Layer Specification Type 2 AAL*, (10/97)

[10] Liu Chunlei, Munir Sohail and Jain Raj, *Packing density of Voice Trunking using AAL2*, paper submitted to Globelcom'99

[11] ATM Forum 98-0630 - Packet Delay Variation in Voice Trunking using AAL2, COMSAT Laboratories, (10/1998)

[12] ATM Forum 98-0830 - Packing Density of Voice Trunking using AAL2, The Ohio State University & NOKIA Research Center (Burlington), (12/1998)

[13]: Technical performance objectives (GSM 03.05 version 6.0.0 Release 1997) - TR 101 631 V6.0.0 (1999-04)

3 External requirements on UTRAN

Reference [1] gives a minimum set of services to be supported in UMTS. The list characterises services by the following set of parameters:

- Range of supported data rates;
- BER requirements;
- One way delay requirements;
- Activity factor.

Table 1 reports the list of services, as provided in [1].

Table 1: List of test data rates for evaluation purposes

	bit rates (kbps)				**BER**	**One way delay**	**Channel activity**
Test environments	**Indoor Office**	**Out- to Indoor and Pedestrian**	**Vehicular 120 km/h**	**Vehicular 500 km/h**			
Representative low delay data bearer for speech*1	8				≤ 10^{-3}	20 ms	50%
LDD Data (circuit-switched, low delay)*1	144-384-2048	64 - 144 - 384	32 - 144 - 384	32 -- 144	≤ 10^{-6}	50 ms	100%

LCD Data (circuit-switched, long delay constrained)*1						300 ms	

*1 One-way delay (excluding propagation delay, delay due to speech framing and processing delay of voice channel coding) for all the test services.
NOTE: For LDD services, a BER threshold of 10^{-4} will be considered for the initial comparison phase of the different concepts in order to reduce simulation times. The BER threshold of 10^{-6} will be considered in the optimisation phase.
UDD Data (packet) - Connection-less information types not reported.

According to references [3] and [4], the total one-way end-to-end delay for voice services should be kept within 40 ms.

In the following figure it is reported from [2] the reference model for the computation of the transmission delay.
It should be considered that:

- The error protection contains any FEC, CRC, interleaving coding and macro-diversity processing.
- UTRAN inter-node delays are not taken into account.
- Implementation overheads, such as processing time, are not included in the definition.
- Speech encoding is not included in the radio transmission chain since it is assumed there will be bearer definitions applicable for speech transmission, as well as for video compression etc.

The one-way delay figures are only applicable for defining the radio technology bearers and not for defining the complete access delay for the radio access network. This means that the total delay will be larger. Thus the figures Tt and Ti must be lower than the requirement for total delay in the access network.

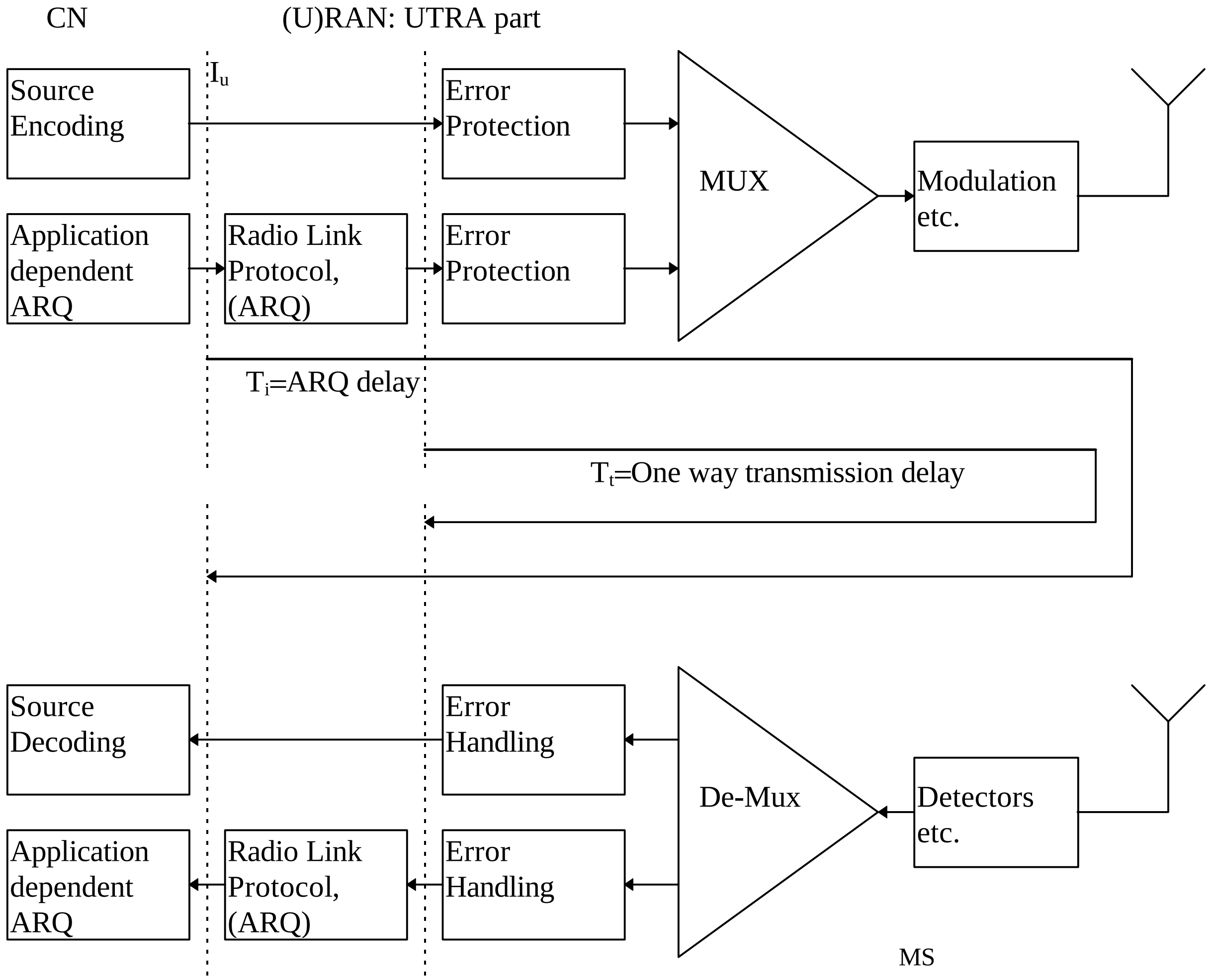


Figure 1: Reference Model for Transmission Delay

4 UTRAN Delay Components: Definitions

In this chapter the transmission delay components across the UTRAN are identified and described.
A short description of each component is given, along with affected services and impacting parameters.
The paper is mainly based on 3GPP RAN WG3 Tdoc. TSGW3#2(99)169 – UTRAN Delay Estimation

4.1 Symbols used

↑ = Direct proportionality
↓ = Inverse proportionality
RT = Real Time
NRT = Non Real Time
CBR = Constant Bit Rate

4.2 UTRAN Network Components

4.2.1 Packetisation, De-packetisation and End-System Play-Out Delay

RT **NRT** The originating terminal adds a packetisation delay. Factor influencing this delay is the ↓ instantaneous source data rate and the allowable transmission rate, i.e. PCR and SCR.
CBR When a real time CBR data stream terminates at an application end-point, play-out buffering is required to remove the CDV caused by the statistical sharing effects of the packet network. Once this variation is removed and de-packetisation applied, the resulting traffic stream from the protocol stack can be fed to higher layers as a constant stream of data. This delay is dependent on the bit-rate of the connection, the ↑ play-out buffer depth (dimensioned on the maximum CDV allowed) and the packet size.

[Evaluation part ffs.]

4.2.2 Macro-diversity Combining Delay

RT **NRT** The Macro Diversity Combination function may require additional switching and processing in the RNC. Even though the delay introduced is heavily implementation-dependent, it has to be considered as a component of the overall delay evaluation.
The MDC function combines signals together at the same moment in time. Therefore, the main delay component in this function is dependent on the difference path delays of each branch involved in a single connection.

~~*[Evaluation part ffs.]*~~

4.2.3 Interleaving and Turbo Coding

RT **NRT** 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 transmission delay to the data stream over the air interface ↑ proportional to the interleaving factor.
NRT Turbo coding has it own internal interleaving mechanism, for data services this is an additional delay depending on the ↑ block dimension and on the ↓ service data rate.

Due to speech coder handling of 20ms speech blocks, an interleaving of 2 has been chosen resulting in a component delay of 20 ms.

[Evaluation part ffs.]

4.2.4 MAC Scheduling Delay

RT For real-time services, a set of resource units will be allocated on a deterministic basis. This implies that a delay no bigger that one transport block is foreseen.

NRT Non-real-time services using shared channels require statistical scheduling, the delay introduced may become important: even though delay guarantees will not be applicable, the delay introduced has an impact on acknowledgement delay and on the resulting QoS. The component depends on ⇒ the load factor of the used resource and possibly the round trip delay between the UE and the RNC.

[Evaluation part ffs.]

4.2.5 Re-transmission Delay

NRT 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 amount of retransmissions needed for a single transport block is a multiplication factor for delay, i.e. if it takes two re-transmissions to transfer a transport block successfully, then twice the physical layer delay would be added plus the delay needed to send the NAK information back to the MAC: ⇒ the maximum number of allowed re-transmissions defines the weight introduced by this component.
In other words the delay introduced by the component is $N_{retransmissions}$ * UE-SRNC Round trip delay.

Re-transmission Delay = Max. allowed re-transmission times * (Interleaving and Turbo Coding + Packetisation, De-packetisation and End-System Play-Out Delay + MAC Scheduling Delay + Macro-diversity Combining Delay + AAL Packetisation, Multiplexing and De-packetisation Delay + Media Delay + Switch delay) * 2.

This delay component can be reduced if prioritisation mechanisms are used for data re-transmission.

4.2.6 Radio Interface Propagation Delay

Since in one millisecond radio waves cover a distance of around 300 km, the delay introduced by the component can be neglected even considering a macro cell environment.

4.2.7 Processing Delays

Node internal processing delay is due mainly to SW processing and to information transfer inside nodes: for this reason this component is heavily implementation dependent and only a rough evaluation can be given.

a) AAL2 switching inside a node is a function requiring higher processing effort than simple ATM switching: a reasonable guess for this component can be 600 µs.
   AAL2 switching is introduced by S- and DRNCs except for PS services.
   In this case AAL2 switching does not take place on the Iu side and the total delay introduced in the SRNC because of this component can be halved.

b) On the uplink soft- (RNC) and softer (Node B) handover processing introduce further delay, but since the task does not require the execution of complex algorithms, a short processing time of 400 µs can be guessed.

c) In the downlink the Node B processing consists essentially of CRC calculation, convolutional/turbo-encoding, interleaving, softer handover splitting, spreading and modulation: the total delay introduced could be totally estimated in 1 ms.

d) In the uplink the processing in Node B is quite different in TDD and FDD, a reference with GSM is also quite difficult due to the different technology used.
Some of the operations required are quite time consuming even when processing is performed by powerful DSPs.
According to [13] the time required for processing UL frames in GSM is 8.8ms: assuming that CDMA required functions (demodulation, despreading, deinterleaving, convolutional/turbo-decoding, error checking and measurements) are more complex (e.g. turbo-decoding and TDD joint detection) and that, on the other hand, DSP performance has increased, it could be derived a guess of about 9 ms for voice services and, due to turbo-decoding, additional 15 µs/(bit/s) for data services.

e) Internal margins should be considered in every node in the range of 1 ms.

The total delay introduced in every node is summarised in the following:

RNC UL	a +b + e (- a/2 if S-RNC, packet switched)	0.3 + 0.4 + 1 (- 0.3)
RNC DL	a + e (- a/2 if S-RNC, packet switched)	0.3 + 1 (- 0.3)
Node B UL	c + d + e	1 + 9 (+ 0.15*(bit/s) if turbo-coding) + 1
Node B DL	b + e	0.4 + 1

It can be assumed a delay of:
2 ms for UL RNC,
1.3 ms for DL RNC,
1.4 ms for DL Node B,
11 ms for UL Node B, no turbocoding,
11 + 0.15*throughput for UL Node B with turbocoding (throughput given in bit/s).

4.3 Transport Network

In this sub-chapter the delay components are described, which are introduced by the transport network interconnecting UTRAN nodes.
To help defining the performance of the transport network, the following delay and bandwidth performance parameters shall be used:

- ***P**acket **T**ransfer **D**elay* (**PTD**) defines the elapsed duration between two measurement points. Mean packet transfer delay is the arithmetic average of a specified number of packet transfer delays.

- ***P**acket **D**elay **V**ariation* (**PDV**) is introduced in [8]. Across the AAL2 CPS, a 2-point measurement defines PDV: the 2-point PDV for a packet between two measurement points (MP) is the difference between the absolute packet transfer delay of this packet between the two MPs and a defined reference packet transfer delay between those MPs.

- ***Packing density*** is defined in [11] and [12] as the ratio of the average user byte number (excluded ATM and CPS headers) in a cell onto the ATM cell length.

In the present evaluation End-to-end PTD and PDV shall be considered: in other words a measurement point shall correspond to one CPS Service Access Point (SAP) at which a CPS-SDU is submitted to the CPS. The second measurement point is localised at the peer SAP delivering CPS-SDU, as shown in Figure 2.

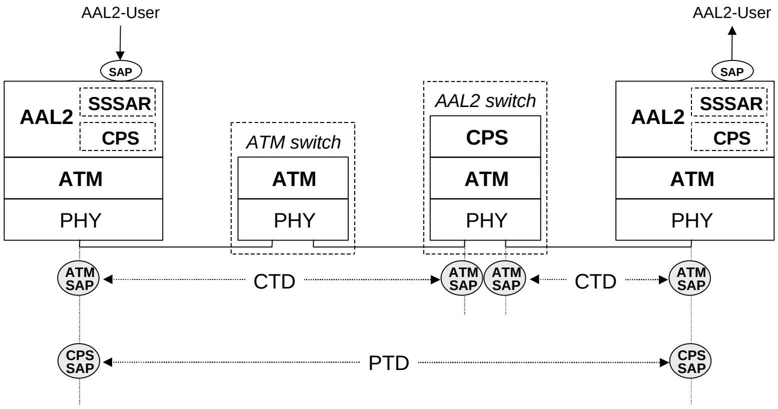


Figure 2 - The differences between CTD and PTD

The packing density results from characteristics of AAL2 user traffic: larger the length of submitted user data is, more significant packing density is. However, there is a theoretical limit of packing density equal to $45/48 \cdot 47/53 \approx 83.14\%$.
Figure 3 depicts the evolution of average packing density according to average length of AAL2 user data (CU_Timer = ∞)
Figure 3

Figure 3 - Evolution of packing density according to average length of user data

According to [10], the CU_Timer_ also affects packing density: if the cell is not completely packed within the time period determined by the CU_Timer_ value, the timer expires and the partially packed cell will be sent. Consequently, the number of VCs on which AAL2 traffic is distributed can also impact packing density.

4.3.1 AAL Packetisation, Multiplexing and De-packetisation Delay

RT **NRT** This component, considered on a point-to-point link, is due to the ATM SAR sub-layer action and to the multiplexing of cells and sub-cells (for AAL2) on the ATM link. link performed in the CPS (Common Part Sub-layer). The PTD and PDV are especially impacted by the *packet queuing delay* in the CPS transmitter buffer. This queuing delay depends on:

- the negotiated *QoS of the ATM connection* (especially ↓ the Peak Cell Rate (PCR));
- the *number of active multiplexed AAL2 connections* ↑ due to an increase of ATM connection load.
- Considering a low ATM connection load, the *CU_Timer_* value affects ↑ the PTD and the PDV in the extent that probability of CU_Timer_ expiration is no null. Defined in [9], CU_Timer_ is optionally used to ensure that a CPS packet does not wait for a too long time before transmission. Besides, [9] has not specified any value for CU_Timer.
- Due to the CU_Timer_ effects, the *number of selected VCs* ↑ to carry AAL2 user data also affects the AAL2 performance. By distributing AAL2 traffic over several VCs, each ATM connection load decreases

(and consequently, probability of Timer_CU expiration increases). Consequently, the PTD and PDV tend to increase.

Moreover, PTD and PDV are impacted by ATM cell queueing at the network switches:

- *The* data-rate of the physical link and the physical protocol *(IMA upon PDH, SDH, …),*
- *Included in the QoS, the CTD and the CDV of the corresponding ATM connection also affect AAL2 performance.*
- *The ATM link load obviously impacts AAL2 performance.*

In this component it is also considered the delay introduced by fractional ATM, i.e. by the partitioning of a physical resource into different interfaces. As an example, it could be considered the case of the share of a PCM E1 interface into one Abis interface and one I_{ub} interface, in order to link over a single physical interface a site supporting both GSM and UMTS services.
The delay introduced by this component is inversely proportional to the grade of fractionalisation, in case of an E1 link the component is about:

Delay = **Error!** µs

The delay introduced by this component on different interfaces should be reasonably different, since Iu and I_{ur} will be presumably broad band interfaces, while I_{ub} data rate will be smaller, even in case of ring topology. The characteristic of the traffic on this interface will be different on up- and downlink, as well, since the traffic profile in the uplink would assume a distribution conditioned directly by the radio interface, while on the downlink the traffic would be originated by an ATM network.

Simulations are in any case required to have more precise figures.
[Evaluation part ffs.]

4.3.2 Media Delay

RT **NRT** The propagation delay over cabled networks can assumed to be fixed and proportional to the connection length.
The same can be assumed for microwave and satellite connections, but to the medium delay a further component must be added, which considers the technology used for the link, e.g. point-to-multipoint, point-to-point, radio ATM.
For satellite links, the delay can be time-dependent, in accordance with the orbit eccentricity .

The following delay can be assumed, according to [5]:

Coax cable: 4 µs/km;
Optical fibre: 5µs/km.

In case µwave links are used, the following indicative values can be considered (ffs):

PDH microwave link: 1.5 ms
SDH microwave link: 1 ms
Point-to-multipoint microwave link: 5 ms

According to [5] and [6] the delay introduced by a satellite link can range between 60ms (max. value for LEOs) and 310 ms (max. value for HEOs).
It is therefore suggested to allow a single satellite hop along a link over the UTRAN between a UE and the Core Network.

4.3.3 Switch Delay

RT **NRT** This is the component due to switching nodes (Cross-Connects and Switches) along UTRAN terrestrial interfaces, only. Its value is proportional to the number of intervening nodes and has a heavy dependence on the traffic load of each node. [7] defines 300 µs as the maximum delay for real-time services through ATM switches. Even if the definition of this component is not clear in the quoted reference, this assumption is accepted as a worst case.

4.4 UTRAN external Delay Components

To help deriving end-to-end and round-trip delay figures, delay components located outside the UTRAN are reported in the following.
The content of this chapter must be considered purely as a guideline, since outside the scope of the present document.

4.4.1 Transcoding for Speech Services

To encode the speech between time X and time X+20ms, the speech coder needs to gather PCM speech samples from time X to X+25ms. The speech coder also uses information from speech samples gathered before time X, but these do not contribute to delay.
Once all the speech samples have been gathered, processing needs to be performed. This is the delay from time "X+25" to when the whole encoded speech 'packet' is available for transmission.
The collection phase introduces 25 ms delay, while the processing phase is carried out in an implementation and coder dependent time.
The delay quoted in [13] for a full rate coder (8 ms) can be considered as a reference for the UE located function, while more powerful processing could reduce the delay introduced by the TRAU coder to the half (4 ms).

The decoding delay must be calculated between the time when the whole speech frame is in the decoder and when the first sample is made available at the decoder output.
[13] provides a value for this component referred to the full rate decoder, as well. As in the coder case the same value (1.5 ms) can be kept for the UE decoder, while the TRAU located decoder should introduce the half delay (0.8 ms)
As far as concerns the component introduced in TRAU, it must be noted that the worse case occurs when two end UEs engaged in a call are using different codecs and different TRAUs: this prevent the support of TFO and the optimisation of a back-to-back decoding-coding function.

4.4.2 Core Network

[To be added as an informative part, since CN architecture is out of scope of RAN3, the following text must still be completed and improved.
For speech services: in addition to the delay introduced by the transcoding function, a TRAU introduces a component due to the AAL2 termination/switching (?? ms) plus some further interface and processing delay (0.5 ms),in GSM the delay introduced by a MSC according to [13] is 300 µs, the media delay on an E interface (MSC-MSC) of 200 km is 800 µs).
The location of the echo canceller in the CN would introduce a further component: deriving the figure from [13] and considering the improvement in the processing capability of DSPs a value of 0.5 ms could be guessed.
In case the echo canceller is located in the UE, a delay of 1 ms could be assumed.

For data services transcoding and AAL2 related delays are absent, but more processing must be considered for the GTP-GTP double tunnelling mechanism. The media delay on one Gn interface of 200 km is 800µs, to be considered twice being the GGSN the anchor point.]

4.4.3 User Equipment

In addition to the transcoding component already described, some further delay is introduced in the UE for example for the A/D conversion, for the processing and to move speech packets from module to module, in case modules are implemented in different HW components. Even though this delay is strongly implementation dependent, a guess of 3 ms could be acceptable for both uplink and downlink.

5 UTRAN Delay Estimation

6 UTRAN Overall Delay Budget (conclusion to be included in Ch. 13.1 of S3.01)

Appendix B

1 Delay Budget Template

1.1 Delay Components

1.1.1 UTRAN Nodes

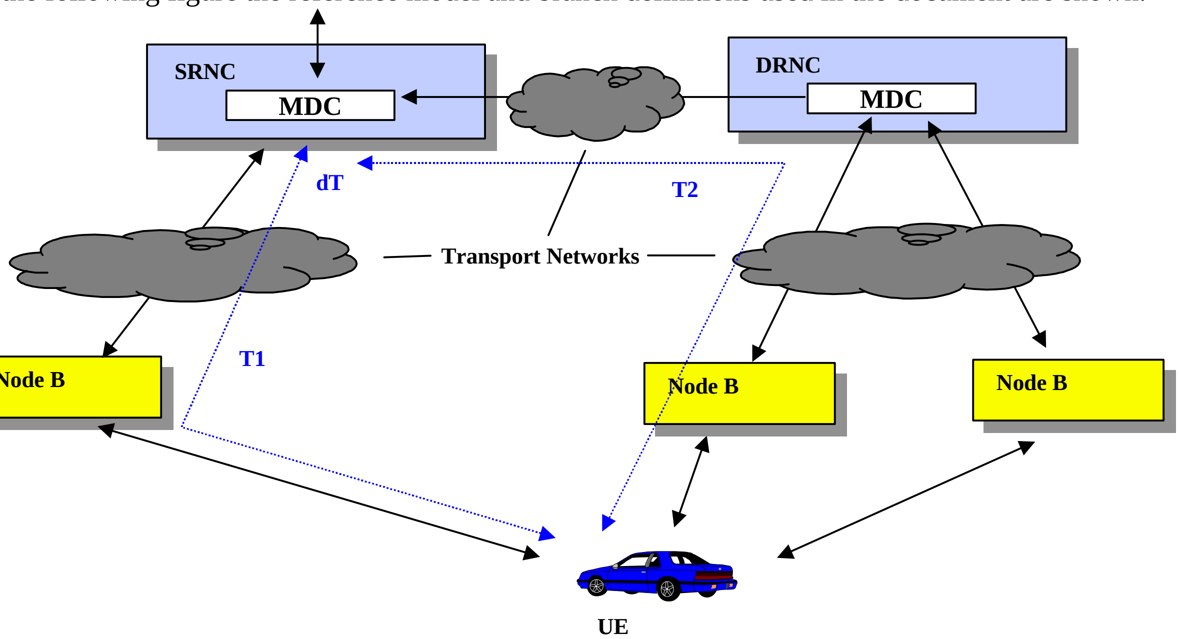
U1): Packetisation, De-packetisation and End-System Play-Out Delay
U2) Processing Delay
U3): Interleaving and Turbo Coding
U4): MAC Scheduling Delay
U5): Re-transmission Delay
U6): Uu delay

1.1.2 Transport Network

TN1): AAL Packetisation, Multiplexing and De-packetisation Delay
TN2): Media Delay
TN3): Switch Delay

1.2 UTRAN Reference Configuration

In the following figure the reference model and branch definitions used in the document are shown.



Network Assumptions

For the evaluation of delay components introduced by the transport network the following assumptions for a typical worst case scenario have been made:

Iub interface: 6-hop PDH µwave link
6-hop SDH µwave link

Iur interface: 600 km STM-1, optical fiber
9 ATM switches/cross-connects

Iu interface: 200 km STM-1, optical fiber
4 ATM switches/cross-connects

For a best case scenario, branch T1 is assumed to consist of co-located RNC and Node B.

1.3 Delay Budget Template

Service (kbit/s)	**8 (RT)**	**32**	**64**	**144**	**384**	**2048**	**Source/Reference**
Delay Component	Delay (ms)						
T1 Branch							
U3	20	100	100	100	100	100	
U6	0.05						
U2 – Node B	11	15.8	20.6	32.6	68.6	318.2	UL value. 1.4 ms for DL
TN1 – I_{ub}	1	1	1	1	1	1	
TN2 – I_{ub}	14						TSGR3#3(99)313, Nokia
TN3 – I_{ub}	0						
U2 – SRNC	2	2	2	2	2	2	UL value. Subtract 0.7 ms for DL
U1	<14	1	1	1	1	1	
U4	0	10	10	10	10	10	
U5	0						
T1 Branch Delay	***62***						
T2 Branch							
U3	20	100	100	100	100	100	
U6	0.5						
U2 – Node B	11	15.8	20.6	32.6	68.6	318.2	UL value. 1.4 ms for DL
TN1 – I_{ub}	1	1	1	1	1	1	
TN2 – I_{ub}	14						TSGR3#3(99)313, Nokia
TN3 – I_{ub}	-						
U2 – DRNC	2	2	2	2	2	2	UL value. Subtract 0.7 ms for DL
U1 – DRNC	<14	2	2	2	2	2	
TN1 – I_{ur}	1	1	1	1	1	1	
TN2 – I_{ur}	3						
TN3 – I_{ur}	2.7						
U2 – SRNC	2	2	2	2	2	2	UL value. Subtract 0.7 ms for DL
U1 – SRNC	<6	2	2	2	2	2	
U4	0	10	10	10	10	10	
U5	0						
T2 Branch Delay	***77.2***						
I_u Interface							
U1 (packetisation only)	0	1	1	1	1	1	
TN1 – I_u	0.5	0.5	0.5	0.5	0.5	0.5	
TN2 – I_u	1						
TN3 – I_u	2.5						
Iu Delay	***4.5***						

Note 1) processing times are not considered, their evaluation requires further study; TN1 has still to be integrated with CPS scheduling component.

In the following table the delay estimation results are reported; delay definitions are reported after the table.

62 77.2

Service (kbit/s)	8 (RT)	32	64	144	384	2048
Delays (processing time to be added)	Delay (ms)					
α) Total delay T1 worst case	66.5					
β) Total delay T2 worst case	81.5					
γ) Total delay T1 best case	33					
δ) Max T2-T1 delay difference	47.5					
ε) SRNC delay	17					
θ) DRNC delay	17					
η) Node B delay	32					

Definitions (with reference to template):

α = T1 Branch Delay + Iu Delay

β = T2 Branch Delay + Iu Delay

γ = T1 Branch Delay + Iu Delay

The evaluation of γ) assumes that components U3, U6 and U2 are unchanged and components TN1 TN2 TN3 U1, U4 and U5 are neglectable.

δ = β – γ

The maximum delay difference between T1 and T2 branches has been compared, T1 being the best case and T2 being the worst case.

ε = U1 + U4 + U2 + TN1

θ = U1DRNC + U2 + TN1

η = U3 + U2 + TN1