TSG-RAN Working Group 3 meeting #6 ***TSGR3#6(99)A18***
Sophia Antipolis, France, 24th - 27th August 1999

Agenda Item:

Source: Vodafone Limited

Title: UMTS Delay Budget

Document for: Discussion

---

1. UMTS Delay Budget

This document is based on the work done in 3GPP RAN3 (in particular documents 700, 800 and 805) and on GSM 03.05 v4.1.0. GSM 03.05 gives a delay template for GSM (basic) full rate speech. This document is intended to add more detail and greater clarity to the current delay template. This template focuses on speech as this is likely to be the most delay critical application. The template can be replicated for other data services/data rates – although some modifications are likely to be needed.

GSM 03.05 was written quite some time ago. Since then, maximum DSP speeds have got a lot faster and speech coders have got more complicated. It seems likely that network infrastructure will use faster DSPs while mobiles will use DSP advances to find cheaper chipsets that run at the old speeds. These trends have been used to adapt the processing delays given in 03.05.

While the figures in the table are not deliberately incorrect, it is expected that many of them need adjustment. The primary intention of this document is break the delay down into its component parts, so that each component can be considered in more detail.

A data rate of 13 kbps is assumed. This allows for some control bits to be added to the 12.2 kbps AMR codec. The peak data rate (rather than average) seems to be the rate which is relevant for the delay calculations. Note that the use of silence detection will probably mean that the average data rate required from the radio interface is around 8 kbit/s.

Major issues
1) the model for the speech (and other) ATM traffic on the Iub interface is not yet clear. Speech packet arrivals are almost certainly not independent. It is possible that they are dependent on the radio interface frame boundaries and thus all arrive at the same time. Alternatively the radio interface protocol is developed to ensure that speech packets have a uniform distribution. The output of the Synchronisation discussions may need to be influenced in order to reduce the speech delay.

2) A time alignment protocol between BTS and TRAU needs to be developed. Is this possible when the TRAU is in fact communicating with several BTSs if the mobile is in soft handover?

3) the ATM delays for queues, jitter and for switching need checking (and probably changing).

1.1 Downlink Speech Delay Budget

Service (kbit/s)	**13 (RT)**	
Delay Component	**Delay (ms)**	**Description and basis for chosen value**
T-sample	25	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).

T-transc	4	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 across the Iu interface. 03.05 quotes T-transc for GSM full rate as 8 ms. Advances in DSP power are partially offset by increases in codec complexity.
T-MSC/MGW Margin	0.5	Allows for internal interface processing/handling delays, etc.
TN1-Iu ATM packetisation and queuing delay	0.1	The Iu interface is likely to be reasonably broadband. For example, much greater than 8 Mbit/s. The speech packets are roughly 260 bits (=13 kbit/s * 20ms). However a full ATM cell is needed to move this AAL 2 packet. An ATM cell has 48 bytes of payload and 5 bytes of header. A queue of 2 ATM cells gives 2*53*8 bits / 8 Mbit/s = 106 us. On this interface, this delay seems to be small.
TN2-Iu Media delay, eg 5 us/km.	1	If the RNC and MSC are co-located this delay is virtually zero. With RNCs remote from the MSC, the distance could be, say, 200 km. This gives about 1 ms of delay.
SRNC processing delay	1	The SRNC probably needs to do some work on top of the ATM switching; however, hopefully the delay is low. AAL2 switching cannot be quicker than ATM switching. I.356 suggests less than 300 us for real time services. Guess at 600us for AAL2.
TN1-Iur ATM packetisation and queuing delay	0.1	The quantity of traffic that needs to use the Iur interface is not clear. However, all RNCs need to be able to communicate with all other RNCs. Hence all the Iur interfaces (that are not contained within one switch building) will probably form an ATM network. This ATM network can share the Iu interface's physical transport. Hence this interface can be fairly broadband, eg > 8Mbit/s. The packetisation/queuing delay is then similar to that for TN1 on the Iu interface.
TN2-Iur Media delay, eg 5 us/km.	3	In order to get trunking benefits, an "Iur network" seems to be needed. This will tend to require the "worst case" Iur interface to go from the SRNC, through an MSC, through, say, 2 ATM switches to the MSC and then to the DRNC. Overall the distance could be, say, 600 km. This gives 3 ms of delay.
TN3-Iur Switch delays in an ATM, Iur network	1.2	Four ATM switches (see TN2-Iur, above) and I.356's 300 us per node gives 4*0.3 = 1.2 ms of delay in an Iur network. This assumes that these ATM switches are performing native ATM switching and not AAL.2 switching.
DRNC processing delay	0.6	The DRNC has less work to do than the SRNC. This delay could be just the real time I.356 delay, increased to allow for AAL.2 switching delay; guess at 600 us.
TN1-Iub ATM packetisation and queuing delay	0.4	The Iub may be a single 2 Mbit/s point to point (RNC to node B) link. There are many alternatives. However, a ring that serves multiple node Bs will need to have much higher bandwidth than that of a single 2 Mbit/s. The worst case seems to be that of the single 2 Mbit/s. 260 bits at 2 Mbit/s takes 0.13 ms. However need to fill a whole ATM cell… If we **assume** that speech frames arrive randomly at the DRNC (or SRNC) for transmission down the Iub, then a queue of 2 ATM cells gives a delay of 2*53*8 bits / 2Mbit/s = 0.4 ms
TN2-Iub Media delay, eg 5 us/km or some microwave delay.	1	200 km gives 1 ms of delay. Although such distance might be rare, other transmission hardware associated with the "last mile" are likely to have an impact.

TN3-Iub Switching delays in the 'Iub network'	0.6	One of the advantages of ATM for the Iub is the possibility to do transmission concentration "downstream" of the RNC. Hence 2 ATM switches delays could be expected. Expect this to be native ATM switching, not AAL 2 switching.
BTS buffering: lack of time alignment	<10 or 0	In GSM, the 08.60/08.61 protocols provide the means to move the transcoder's frame boundaries into alignment with the radio interface speech frame boundaries. Will UMTS have a similar protocol? If not then this will produce a delay of up to 10 ms. I think that UMTS should define some time alignment protocol: in which case this delay ought to tend to 0 ms. GSM TFO disables the time alignment protocol: hence the use of TFO does not dramatically reduce the delay in a GSM mobile to GSM mobile call. This will also be a problem with UMTS TFO unless the mobiles do end to end time alignment of both their transcoder frames AND their RADIO INTERFACE FRAMES.
BTS buffering: ATM jitter	12	In GSM, there is a fixed, constant delay between BTS and TRAU. The introduction of a packet network in between BTS and TRAU will add some jitter to this fixed delay. Hence the time alignment protocol has to aim to get the speech packets delivered Y ms before they need to start being transmitted over the radio interface in order to ensure that [99.99%] of packets are available in time. GSM 05.05 implies that Frame Erasure Rates of around 1% are tolerable. However what FER is required for 'toll quality'? I guess that we should not plan to have more than 0.01% of speech packets discarded because they arrive 'too late' over the Iub. Some allowances have already been made for jitter in the TN1 and TN3, and RNC processing delay values above. However, extra allowances will be needed to achieve 99.99%. How to try and produce a safe guess? We need the mean and variance of the distributions. If we assume mean=variance (why? - I seem to remember that this applies for a Poisson distribution) and then allow [4]*variance for jitter (why 4? - some one needs to look at some maths tables and work out how many variances gives 99.99% confidence) then we add [4] times the mean delay: Mean delay = TN1-Iu + SRNCatm + TN1-Iur + TN3-Iur + DRNCatm +TN1-Iub + TN3-Iub = 0.1 + 0.3 + 0.1 + 1.2 +0.3 +0.4 + 0.6 = 3.0 ms. 4* mean = 12 ms
Node B to cells Softer handover splitting	1	When a mobile is in soft(er) handover between 2 or more cells on one site (node B), the node B needs to do some work to copy the received packet to the relevant cells. Guess that this will take less than 1 ms.
BTS internal delays "margin"	1	This is a margin for internal signalling interface delays,etc
T-encode CRC calculation, convolutional encoding and interleaving	1	These processes ought to be relatively straightforward for a base station.

Time taken to send speech packet over the radio interface (U3)	20	Given that CDMA is not a TDMA system, a Iub packet containing 20 ms of speech will take 20 ms to transmit over the CDMA radio channel. As a result interleaving over the whole 20 ms could be used with 'no' delay penalty. The greater the interleaving depth/delay, the better the averaging on the radio interface. GSM uses 40 ms (which 'optimises' to a delay of 37.5 ms). Owing to the speech coder's handling of 20 ms speech blocks, the use of 10 ms interleaving depth would seem to be wasteful and would not lower the speech delay–
Radio interface propagation delay (U6)	0	30 km at the speed of light is 0.01 ms. For this document, this delay is negligible.
T-rxproc Channel decoding and (if they exist) equalisation, interference cancellation and joint detection.	4	In GSM 03.05 this is the time needed to perform equalisation and channel decoding and is quoted as 8.8 ms. GSM mobiles are probably designed to be able to (just) equalise one timeslot per frame. Hence 4.6 of this delay is for the equaliser. Assuming that UMTS (FDD) does not use an equaliser or joint detection then the equaliser delay can be discounted. In the GSM case, this leaves 4.2ms for the de interleaving and convolutional decoding. Ongoing DSP speed increases probably mean that this delay can be could be cut by a factor 4 to about 1 ms; however the tendency in mobiles to just use cheaper DSPs will act against this: hence my guess is 4 ms.
UE margin	1	Time is needed to move the 260 bits of the speech packet from the "rake receiver module" to the "convolution decoder module" and then on to the "speech decoder" module. If any of these modules are built on separate chips then more significant delay can build up.
T-proc Delay needed for speech decoder to produce **first** 'PCM' sample.	2	This is the delay from the "260 speech bits" being loaded into the speech decoder to the "first PCM" sample being sent to the earpiece. GSM 03.05 lists this as 1.5 ms. Increased DSP speed will probably be offset by more complex speech coders (although an N times increase in speech coder complexity may only lead to an N/2 increase in decoder complexity) and the use of lower quality DSPs in the mobile. Guess that this figure could be 2 ms.

Total downlink: 80.5 (+<10) ms

1.2 Uplink Speech Delay Budget

Service (kbit/s)	**13 (RT)**	
Delay Component	**Delay (ms)**	**Description and basis for chosen value**
Echo control	2	Do (or should) mobiles have to implement some form of electronic echo cancellation within the handset? This seems likely (if the echo requirements are actually to be achieved), and is probably feasible to perform without incurring huge delay. GSM assumes that mobiles limit the echo by "good acoustic design" of the handset. Given the current shape and forms of handsets, there seems to be potential for cheaper/better handsets via the use of electronic techniques.
T-sample	25	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).

T-transc	8	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 to be passed to the convolution encoders (etc) in the mobile. 03.05 quotes T-transc for GSM full rate as 8 ms. Advances in DSP power are partially offset by increases in codec complexity and the desire to use the least expensive DSP in the mobile. Guess that this figure remains at 8 ms. (Absolute maximum is 20 ms.)
Alignment of speech and radio frames in the mobile	1	A decent UE design will ensure that the speech frames are always delivered at the 'right time' so that no extra delay is incurred. Some delay is however inevitable, eg due to a cheap, slow internal bus. 260 bits across a 512 kbit/s bus is 1ms. If the voice codec is implemented in, say, a laptop then extra delay can be expected, eg <10ms.
T-encode CRC calculation, convolutional encoding and interleaving	1.5	These processes ought to be relatively straightforward but a cheap DSP in the mobile will do them slower than in the base station.
Time taken to send speech packet over the radio interface (U3)	20	Given that CDMA is not a TDMA system, an speech packet containing 20 ms of speech will take 20 ms to transmit over the CDMA radio channel. As a result, interleaving over the whole 20 ms could be used with 'no' delay penalty.
Radio interface propagation delay (U6)	0	30 km at the speed of light is 0.01 ms. For this document, this delay is negligible.
T-rxproc Channel decoding and (if they exist), equalisation, interference cancellation and joint detection.	4	In GSM 03.05 this is the time needed to perform equalisation and channel decoding and is quoted as 8.8 ms. The same delay is quoted for mobile and BTS. GSM mobiles are probably designed to be able to (just) equalise one timeslot per frame. Hence 4.6 of this delay is for the equaliser. Assuming that UMTS (FDD) does not use an equaliser or joint detection then this delay can be discounted. In the GSM case, this leaves 4.2ms for the de interleaving and convolutional decoding. Ongoing DSP speed increases in the base station probably mean that this delay can be cut by a factor 4. CDMA specific features may add complexity: allow a factor four increase. Guess that the end result is about 4 ms.(* There is an assumption here that the BTS's receiver processing task "per mobile" is about 4 times more complex than the mobile's receiver task. *)
BTS internal delays "margin"	1	This is a margin for internal signalling interface delays,etc
Node B processing Softer handover	1	The processing needed for softer handover within the node B has to take some time. However, guess that this might be less than 1ms.

TN1-Iub ATM packetisation and queuing delay	<10 or 1.3	The radio interface speech frames may be aligned so that all the mobiles using that cell finish their 10 ms frames at the same time. This means that there is a sudden 'burst' of packets which need to be sent up the narrow bandwidth Iub interface. This would cause up to 10 ms of delay. This delay might be easily avoided, provided the technique is standardised. For example, different mobiles could be told to align their frame boundaries with different slot boundaries (currently there are 15 slots per frame). This requires debate in the RAN 1 group and standardisation in RAN 1 and RAN 2. RAN 3 synchronisation ad hoc may also be involved. Provided that the mobiles are "staggered" then uplink speech packets would assume a uniform distribution and their queuing delay could be assumed to be less than 1 radio interface slot - if the Iub interface has enough bandwidth to handle all mobile users talking at the same time. However Iub dimensioning might assume that X% of users are silent at any one time. A delay of less than 2 slots seems reasonable, ie 2*10/15 ms = 1.3 ms.
TN2-Iub Media delay, eg 5 us/km or some microwave delay.	1	200 km gives 1 ms of delay. Although such distance might be rare, other transmission hardware associated with the "last mile" are likely to have an impact.
TN3-Iub Switching delays in the 'Iub network'	0.6	One of the advantages of ATM for the Iub is the possibility to do transmission concentration "downstream" of the RNC. Hence 2 native ATM switch delays could be expected.
DRNC processing delay	0.6	The DRNC needs to do some work; however, hopefully the delay is low. AAL2 switching cannot be quicker than ATM switching. I.356 suggests less than 300 us for real time services. Guess at 600 us for AAL 2 switching.
TN1-Iur ATM packetisation and queuing delay	0.1	The quantity of traffic that needs to use the Iur interface is not clear. However, all RNCs need to be able to communicate with all other RNCs. Hence all the Iur interfaces (that are not contained within one switch building) will probably form an ATM network. This ATM network can share the Iu interface's physical transport. Hence this interface can be fairly broadband, eg > 8Mbit/s. The packetisation/queuing delay is then similar to that for TN1 on the Iu interface.
TN2-Iur Media delay, eg 5 us/km.	3	In order to get trunking benefits, an "Iur network" seems to be needed. This will tend to require the "worst case" Iur interface to go from the SRNC, through an MSC, through, say, 2 ATM switches to the MSC and then to the DRNC. Overall the distance could be, say, 600 km. This gives 3 ms of delay.
TN3-Iur Switch delays in an ATM, Iur network	1.2	Four ATM switches (see TN2-Iur, above) and I.356's 300 us per node gives 4*0.3 = 1.2 ms of delay in an Iur network. This should be native ATM switching, not AAL.2 switching.

ATM jitter - buffering before/soft handover combining	10.4-X	In the uplink, the SRNC will need to buffer the frames arriving from the different paths until all of them have arrived and the combining can be performed. Some allowances have already been made for jitter in the TN1 and TN3, and RNC processing delay values above. However, extra allowances might be needed to achieve 99.99%. Could reuse the same value as for the downlink field: "BTS buffering- ATM jitter", except that if we have regular uplink packet arrivals at the BTS, the the Iub delay should be less variable (eg regard it as fixed for this part). Mean delay = TN1-Iu + SRNCatm + TN1-Iur + TN3-Iur + DRNCatm + TN3-Iub = 0.1 + 0.3 + 0.1 + 1.2 +0.3 + 0.6 = 2.6 ms. 4* mean = 10.4 ms This buffering delay can be traded off against the buffering delay in the transcoder needed to ensure a steady stream of output speech samples.
SRNC processing delay	1	As a minimum, ATM AAL.2 switching delay. Extra delay for soft handover combining is likely. Guess at a total of 1ms.
TN1-Iu ATM packetisation and queuing delay	0.1	The Iu interface is likely to be reasonably broadband. For example, much greater than 8 Mbit/s. The speech packets are roughly 260 bits (=13 kbit/s * 20ms). However a full ATM cell is needed to move this AAL 2 packet. An ATM cell has 48 bytes of payload and 5 bytes of header. A queue of 2 ATM cells gives 2*53*8 bits / 8 Mbit/s = 106 us. On this interface, this delay seems to be small.
TN2-Iu Media delay, eg 5 us/km.	1	If the RNC and MSC are co-located this delay is virtually zero. With RNCs remote from the MSC, the distance could be, say, 200 km. This gives about 1 ms of delay.
ATM jitter - buffering before/after speech decoding.	X	This delay is the other part of the buffering in the SRNC. In GSM, there is a fixed, constant delay between BTS and TRAU. The introduction of a packet network in between BTS and TRAU will add some jitter to this fixed delay. The TRAU needs to add some delay to ensure that it can continue to stream out 64 kbit/s PCM speech. .
T-MSC/MGW Margin	0.5	Allows for internal interface processing/handling delays, etc.
T-proc Delay needed for speech decoder to produce **first** 'PCM' sample.	0.8	This is the delay from the "260 speech bits" being loaded into the speech decoder to the "first PCM" sample being sent out 'towards the PSTN'. GSM 03.05 lists this as 1.5 ms. Increased DSP speed will be partially offset by more complex speech coders (although an N times increase in speech coder complexity may only lead to an N/2 increase in decoder complexity). Guess that this figure changes to 0.8 ms.

Total uplink: 85.1 (+<8.7)