Technical Specification Group Services and System Aspects **TSGS#23(04)0080**
Meeting #23, Phoenix, Arizona (USA)
15-18 March 2004

Source: **TSG-SA WG4**

Title: CR 26.937 001 rev 2 on "Rate Adaptation simulation results for PSS" (Release 6)

Document for: **Approval**

Agenda Item: **7.4.3**

The following CR, agreed at the TSG-SA WG4 meeting #30, is presented to TSG SA #23 for approval.

Spec	CR	Rev	Phase	Subject	Cat	Vers	WG	Meeting	S4 doc
26.937	001	2	Rel-6	Rate Adaptation simulation results for PSS	D	5.0.0	S4	TSG-SA WG4#30	S4-040159

3GPP TSG-SA4 Meeting #30
Malaga, Spain, February 23 – 27 2003

Tdoc S4-040159

CR-Form-v5

CHANGE REQUEST

⌘ **26.937** **CR** **001** ⌘ **rev** **2** ⌘ Current version: **5.0.0** ⌘

*For **HELP** on using this form, see bottom of this page or look at the pop-up text over the ⌘ symbols.*

Proposed change affects: ⌘ (U)SIM ☐ ME/UE ☐ Radio Access Network ☐ Core Network ☐

Title:	⌘	Rate adaptation simulation results for PSS
Source:	⌘	TSG SA WG4
Work item code:	⌘	PSSrel6
Date:	⌘	16 March 2004
Category:	⌘	**D**
Release:	⌘	REL-6

Use one of the following categories:
F *(correction)*
A *(corresponds to a correction in an earlier release)*
B *(addition of feature),*
C *(functional modification of feature)*
D *(editorial modification)*
Detailed explanations of the above categories can be found in 3GPP TR 21.900.

Use one of the following releases:
2 (GSM Phase 2)
R96 (Release 1996)
R97 (Release 1997)
R98 (Release 1998)
R99 (Release 1999)
REL-4 (Release 4)
REL-5 (Release 5)

Reason for change:	⌘	Simulation results on rate adaptation for PSS are included
Summary of change:	⌘	A new section on rate adaptation for streaming in Rel. 6 is included
Consequences if not approved:	⌘	Implementations lack from reference simulation results.

Clauses affected:	⌘	6.3		
Other specs affected:	⌘	☐ Other core specifications ☐ Test specifications ☐ O&M Specifications	⌘	
Other comments:	⌘			

How to create CRs using this form:
Comprehensive information and tips about how to create CRs can be found at: http://www.3gpp.org/3G_Specs/CRs.htm. Below is a brief summary:

1) Fill out the above form. The symbols above marked ⌘ contain pop-up help information about the field that they are closest to.

2) Obtain the latest version for the release of the specification to which the change is proposed. Use the MS Word "revision marks" feature (also known as "track changes") when making the changes. All 3GPP specifications can be downloaded from the 3GPP server under ftp://ftp.3gpp.org/specs/ For the latest version, look for the directory name with the latest date e.g. 2001-03 contains the specifications resulting from the March 2001 TSG meetings.

3) With "track changes" disabled, paste the entire CR form (use CTRL-A to select it) into the specification just in front of the clause containing the first piece of changed text. Delete those parts of the specification which are not relevant to the change request.

6.2.5.4 Example scenario relying on 3GPP QoS guarantees

A streaming session setup scenario comprising the following steps is an example of how the different buffering and rate control related parameters can be interpreted and applied in a rate adaptive service environment.

1. Offline encoding of a set of bitstreams at different bitrates. The bitrate range should be around the highest bitrate allowed by the codec level in use in PSS, but should also include lower and higher bitrate streams. Each of which bitstreams together with its transmission schedule is conformant to the hypothetical pre-decoder buffering model with the default parameters (or close to it).

2. Client sends to the server in the capability exchange process a pre-decoder buffer size parameter which is close to its maximum pre-decoder buffer size.

3. Using the given bitstream set (i.e. I-frame placement and stream bitrate) and assuming a given worst case transmission rate adaptation sequence (assuming a pre-defined transmission curve-reception curve control strategy and worst case reception rate variation), server estimates whether it can guarantee without significant quality loss a maximum sampling curve-transmission curve difference smaller than or equal to the client signalled parameters. It can also decide to not commit to the client signalled parameters, but require higher values than that. This algorithm also outputs a safe recommended initial pre-decoder buffering period to be applied for the bitstream set.

4. Server sends an SDP using the average bitrate stream bitrate and the pre-decoder buffer parameters (i.e. max difference between the sampling and the transmission curve) that it attempts to guarantee.

5. Client requests a streaming RAB with QoS parameters similar to those in Annex J of TS26.234 [3].

6. Client analyses the granted QoS parameters by the network and decides how much jitter buffering there needs to be. In case of strict QoS scheduling on the network, the maximum expected time difference between transmission curve and reception curve is in fact the granted “transfer delay” QoS parameter.

7. Client decides whether it can accept the server signalled parameters (i.e. whether the sum of the server signalled pre-decoder buffer size and buffer size required for jitter buffering exceeds some hard limit of the client pre-decoder buffer size). It can decide not to continue with the session setup if it can not provide the required pre-decoder buffer, and can release the streaming bearer.

8. Client sets up a total pre-decoder buffer size as the sum of server signalled pre-decoder buffer size (i.e. maximum sampling curve-transmission curve difference) and estimated maximum transmission curve-reception curve difference.

9. Client sends a SETUP request and waits for the OK from the server.

10. The client sends a PLAY request, the server responds OK and starts streaming.

11. Client pre-rolls into the pre-decoder buffer for a time which is the sum of initial pre-decoder buffering period and the maximum transfer delay.

12. The server will operate the sampling curve-transmission curve control with the parameters that it signalled.

13. The server will be responsible to explore the max transfer delay limit of the network, and operate its transmission curve-reception curve control to avoid packet drops by the network due to enforcing of the max transfer delay.

6.3 Signalling for rate adaptation in Release 6

6.3.1 Implementation of the signalling for rate adaptation

This section gives implementation guidelines of the signalling for rate adaptation. The goals for a rate adaptation implementation should be:

- Optimising the throughput through the network, i.e. avoid buffer underflow or overflow of the network buffers.
- Pauseless playback at the PSS client, i.e. avoid buffer underflow or overflow of the client buffers.
- Optimising the “quality”, i.e. transmit with the best content rate possible.

- Limiting the delays at the receiver and in particular avoid as much as possible the need for rebuffering.

In order to describe the implementation, we make the distinction between transmission rate and content rate. At any point of time, the transmission rate is "how much" is sent on the network. On the other hand, the content rate is "what" is sent on the network. Transmission rate control and content rate control are what regulate the behaviour of the server. When RTCP reports are received, the sender may adapt its current transmission and content rates based on the feedback.

The transmission rate control is based on the statistics available in the RTCP Receiver Report (RR). Through the statistics, the sender is able to estimate the network throughput and react accordingly.

There are many tools one may use to perform content rate adaptation. They generally fall into two categories: bitstream switching and bitstream thinning (the two of the them can be combined.) There can be many variants such as the number of bitstreams used, the types of frames used for switching, whether packets can be skipped, etc…

For validation purposes, the implementation described here uses a simple bitstream-switching scheme. The sender may take the decision to switch between bitstreams at I frames.

The server keeps track of the receiver buffer status through the OBSN reports. The sender uses this information to avoid underflow and overflow of the receiver buffer. The OBSN reports allows the sender to know how much playout time the receiver currently has (underflow condition) and how many bytes are in the buffer (overflow condition).

The server keeps in memory the following information about the packets it sends: sequence number, timestamp and size. The sender can delete this information after a packet has been played by the receiver (i.e. when it is not in the receiver buffer anymore).

In order to avoid buffer overflow, the server can estimate through the OBSN how many bytes are currently waiting for playout at the receiver. By comparing this value to the total buffer size signalled in the RTSP at the start of the session, it can derive if the receiver is close to overflow and should thus decrease its transmission rate. As explained above, the sender chooses its transmission rate in order to maximise the network throughput and to guarantee a pauseless playback to the PSS client. However, if the sender gets closer to the receiver overflow point, it will send at a lower rate than the optimum rate supported by the network in order to avoid the overflow.

In order to avoid underflow, the sender monitors the current receiver buffer delay. This can be estimated through the OBSN APP packet since the APP packet contains information about the next packet to be played out (OBSN field) and the delay until this packet will be played out.

The basic idea for receiver underflow prevention is simple. If the buffer level in time decreases, the sender switches down to a lower content rate. Decreasing the content rate allows the sender to send packets earlier and increase the receiver buffer level again. Throughput variations because of varying network conditions (in particular handover) and network load can be significant. To this end, the sender aims at maintaining at least the target buffer level (in time).

6.3.2 Test results over EGPRS

6.3.2.1 Parameters used for the simulation

Client: initial buffering of 8 seconds

Content: NASA video clip, duration 3 minutes (180 seconds). Pre-encoded at 3 different bitrates 20kbps, 35kbps and 50 kbps. The packet size is 300 bytes (excluding RTP/UDP/IP headers).

Server switching mechanism: upswitch or downswitch only on I-frames.

Rate adaptation parameters:

- Buffer size: 115000 bytes
- Target buffer level: 12 seconds
- RTCP interval: 1 second

Network: EGPRS (emulator)

- Two timeslots (MCS-7 coding scheme) are used with RLC ACK mode.. The network load is divided into real-time traffic and non real-time traffic. In addition to the RTP application (real-time), there are two other mobiles with ON/OFF TCP traffic (non real-time) emulating the load that would occur because of Web Browsing.
- The theoretical maximum channel bitrate at the radio layer when using two MCS-7 timeslots is 89.6 kbps, but because of protocol header overhead and RLC layer retransmission, the real available throughput is less.

Therefore, the RTP traffic will experience variations in bitrate because of both:

- Network conditions (and handovers)
- Varying network load

6.3.2.2 Results

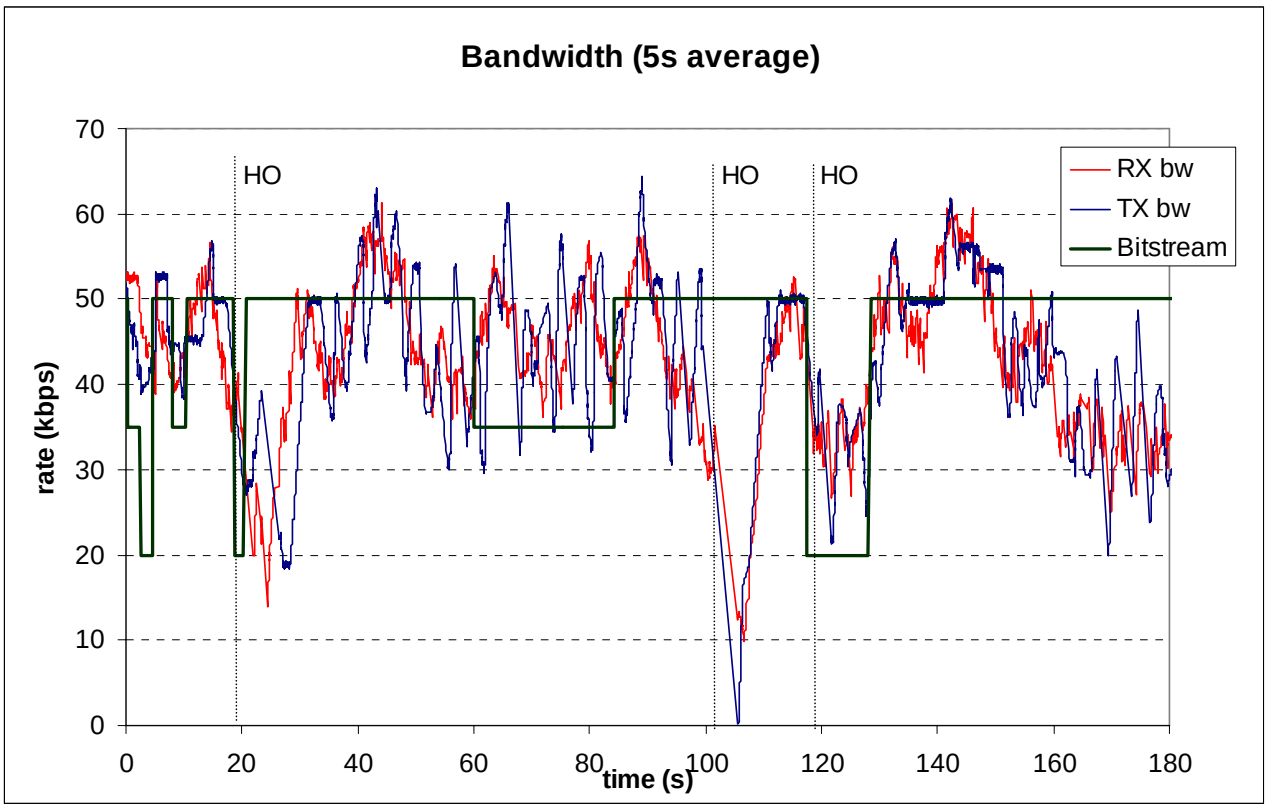
The following plot shows the bitrate received by the receiver over time (averaged over 5-second intervals) and the adapted transmission bitrate. It can be seen that the transmission rate (blue curve) is adapted to the reception rate (red curve) through estimation of the network throughput.

There were three handovers during the run:

- At time 19.8 s that lasted for 2.2s
- At time 101.6s that lasted for 4.0 s
- At time 116.8 that lasted for 1.8s

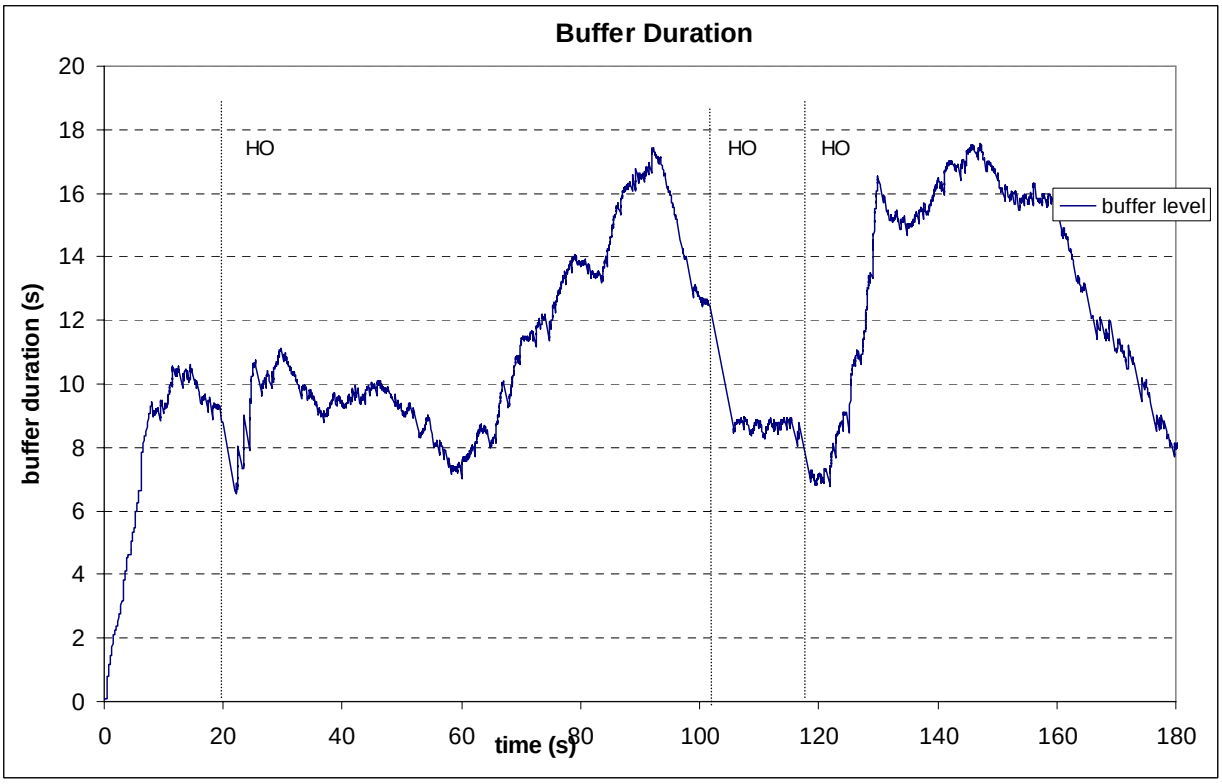
Start of the handover periods are marked with vertical lines on the plot.

As a result of these handovers, the average bitrates were very low at these times.
The plot also shows the content bitrate, i.e. the bitstream (20kbps, 35kbps or 50kbps) selected by the server at a given time instant.



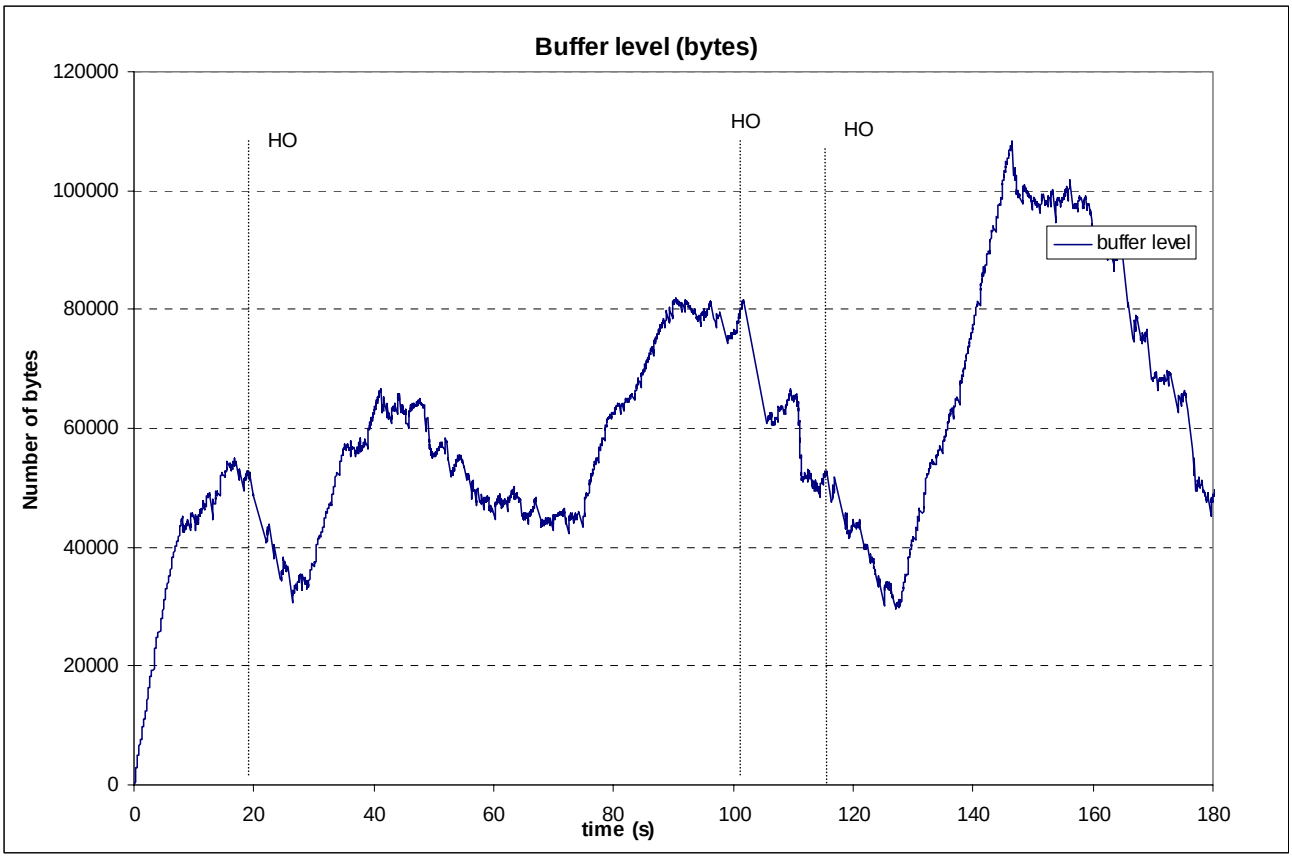
The average content rate during the session is 40 kbps.

The buffer duration is shown in the figure below.



The target buffer level is 12s and is the minimum protection against throughput variation that the sender aims at providing. When the network conditions are good and the sender maximises the throughput available from the network, the buffer duration will be higher than the target level.

The buffer level in bytes is shown in the figure below.

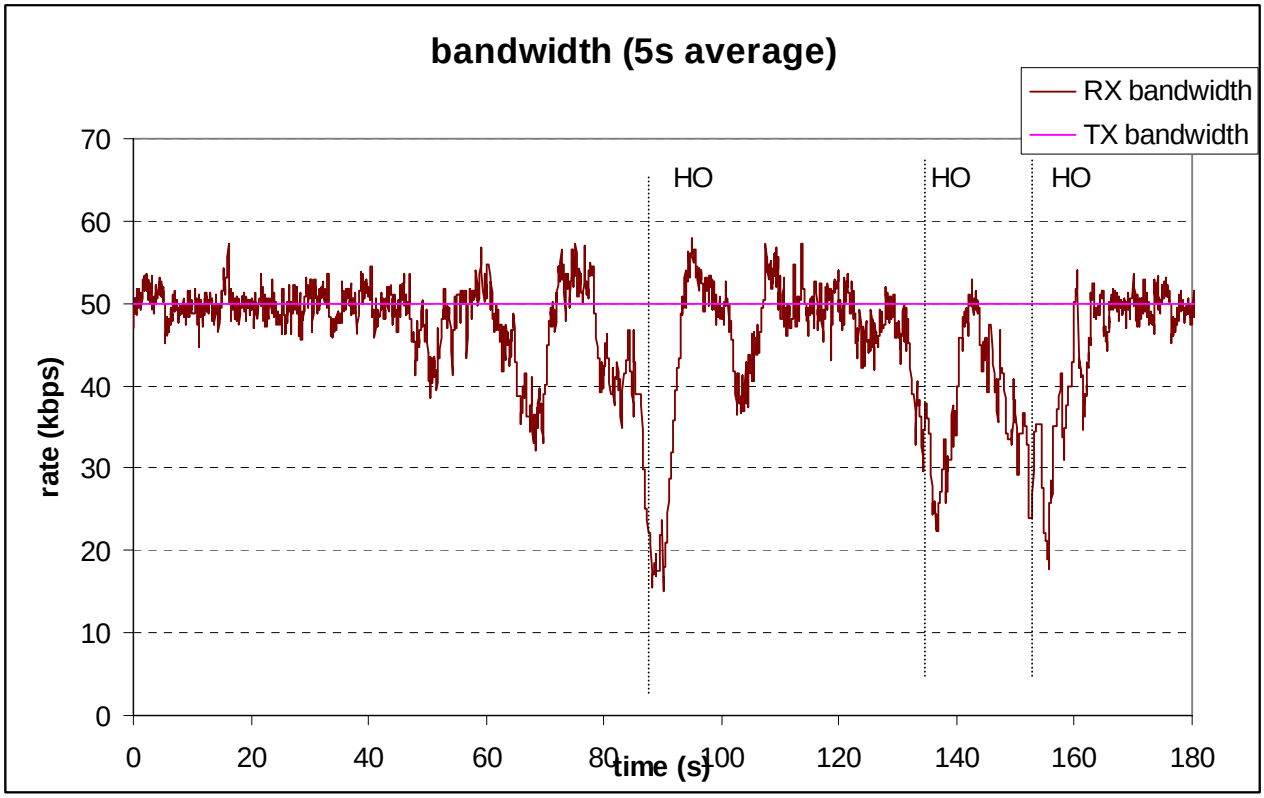


Despite high bandwidth variations, the sender is capable through the signalling for rate adaptation to control the receiver buffer level and thus provides a better end-user experience.

As a reference for comparison, it is given below some plots for a server that would not implement the rate adaptation and that would send at a constant bitrate (50kbps).

Since the simulation environment is dynamic, the simulated throughput is different. However, it has similar characteristics as in the previous simulation.

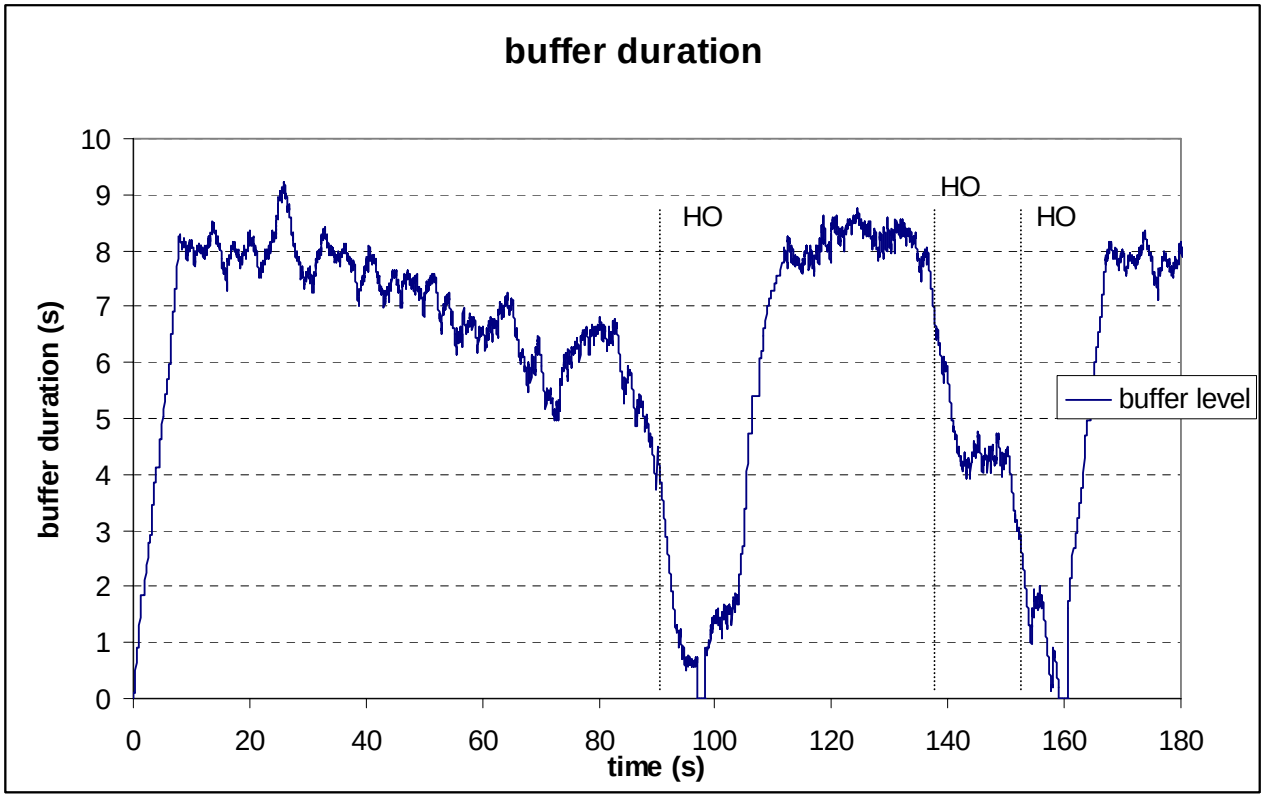
The bandwidth plot is shown below:



There were 3 handovers

- The first one at time 88.3s that lasted for 3.6s
- The second one at time 135s that lasted for 2.2s
- The third one at time 152 that lasted for 2.6 seconds

The buffer level plot is shown below. Because of the network load and handovers, the initial receiver buffer level decreases during the connection. When it underflows, the client needs to rebuffer which leads to interruption of playback.



6.43 UMTS QoS profile parameters

The UMTS QoS profile [4] is used as the interface for negotiating the application and network QoS parameters. In the following some PSS application specific interpretation of the QoS profile parameters is given. The shown PSS performance in the use cases should be achievable when the only knowledge available about the streaming bearer before starting the streaming session is the knowledge extracted through the following interpretation of the QoS parameters.

6.43.1 Guaranteed and maximum bitrate

The guaranteed bitrate can be understood as the throughput that the network tries to guarantee.

The maximum bitrate is used for policing in the core network (i.e. at the GGSN). Policing function enforces the traffic of the PDP contexts to be compliant with the negotiated resources. If downlink traffic for a single PDP context exceeds the agreed maximum bit rate, user IP packets are discarded to maintain traffic within allowed limits. IP packets could additionally be discarded at any bit rate between the guaranteed and the maximum, when enough resources are not available for the PDP context.

In case of a streaming application, it is possible to shape the excessive traffic and queue those packets exceeding the guaranteed bitrate since the application buffer relaxes the delay requirements. This queuing consists of scheduling packets from a connection up to the maximum throughput and the rest of the packets remain in the corresponding queue.