3GPP TR 26.937 V0.1.0 (2001-12)

Technical Report

3rd Generation Partnership Project;
Technical Specification Group Services and System Aspects;
Transparent end-to-end packet switched streaming service (PSS);
RTP usage model
(Release 5)

Keywords

UMTS, IP, packet mode, protocol

3GPP

Postal address

3GPP support office address

650 Route des Lucioles - Sophia Antipolis
Valbonne - FRANCE
Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Internet

http://www.3gpp.org

Contents

Foreword

This Technical Report has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

x the first digit:

1 presented to TSG for information;

2 presented to TSG for approval;

3 or greater indicates TSG approved document under change control.

y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.

z the third digit is incremented when editorial only changes have been incorporated in the document.

1 Scope

The objective of this document is to define an "RTP usage model" for the 3GPP Packet-switched Streaming Service (PSS). In doing so, the document considers how a 3G network could be optimally configured for transporting the RTP traffic, and how the streaming mechanism itself should be designed and optimised given an understanding of the underlying transport characteristics.

The scope of this document includes consideration of (non-exhaustive):

- Trade-off between radio usage efficiency and streaming QoS
- Feedback of network conditions and adaptation of stream and/or the transmission of the stream
- Optimal packetisation of the media stream in line with the segmentation within the transport mechanism
- Error robustness mechanisms (such as retransmission)
- Client buffering to ease the QoS requirements on the network and enable more flexibility in how the network transport resources are applied
- Optimal selection of media and bearer based on prior knowledge in session establishment

2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] 3GPP TR 41.001: "GSM Release specifications".

[2] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

[3] 3GPP TS 26.234 (V5.0.0 onwards): "Transparent end-to-end packet switched streaming service (PSS); Protocols and codecs".

[4] 3GPP TS 23.107: "QoS Concept and Architecture".

[5] IETF RFC 1889: "RTP: A Transport Protocol for Real-Time Applications", Schulzrinne H. et al., January 1996.

[6] 3GPP TS 22.233: "Transparent end-to-end packet-switched streaming service. Service aspects (Stage 1)" (Release 5)

3 Definitions and abbreviations

3.1 Definitions

For the purposes of the present document, the terms and definitions given in 3G TR 21.905 [3] and the following apply:

network: in the context of the RTP usage model network refers to the UMTS bearer service between the entry-point of the UMTS network (i.e. GGSN) and the UE.

3.2 Abbreviations

For the purposes of the present document, the abbreviations given in 3GPP TR 21.905 [3] and the following apply:

RTCP	RTP Control Protocol
RTP	Real-time Transport Protocol

4 Background and motivation

4.1 Mobile RTP

The IETF defined Real-time Protocol - RTP/RTCP [5] for real-time multimedia transport over IP is the selected protocol for audio, video and speech media in the 3GPP PSS [3]. The RTP/RTCP protocol specification is a general description of the functionality that RTP can provide, but the specification itself does not define how an application should use RTP/RTCP most efficiently in a given network environment. Application designers should consider commonly the characteristics of the real-time source and network environment in which the application operates, in order to achieve optimal delivered media quality in a given application scenario.

UMTS networks have specific characteristics, so that a PSS application using RTP/RTCP should be designed specifically for UMTS networks in order to achieve optimal performance.

4.2 Vertical awareness

Vertical awareness means, that the assumptions that network makes about the application and application makes about the network should be consistent. Better application quality can be achieved if the algorithms and enhancements used in the application and the network are tuned to each other (application, network co-design).

No vertical awareness can result in

- degraded application quality.

- inefficient UMTS bearer utilisation, unnecessary high cost of implementation of a bearer.

4.3 Horizontal awareness

Horizontal awareness means, that the behaviour and optimal operating point of the server is known by the client and vice versa. Enhanced application quality can be achieved through collaboration of the server and client.

No horizontal awareness can result in

- degraded application quality when server (client) does not behave as anticipated by the client (server).

5 Overview

5.1 Application and network modelling

A system model is presented where both network and application are modelled in fairly abstract terms, considering all aspects that have effect on the QoS performance of the system. The models don't reveal internal workings of the network or the application, but provide enough detail about their behaviour and characteristics to be able to tune operations of one to the other in the effort of system performance optimisation (application quality and network utilisation).

The models try to capture all possible network and application characteristics that are implied by the different implementation options. Applications and network implementations are classified according to their characteristics. The most suitable network characteristics, implied by certain bearer implementation options, are found for an application class with given characteristics.

The model interface is defined and clarified for unambiguous sharing of the model parameters between the network and the application.

5.2 Recommended PSS implementation

5.2.1 Mobile aware RTP application

An RTP application firstly has to implement algorithms (transmitter and receiver respectively) complementing RTP (e.g. rate control) and utilising functionality provided by RTP (e.g. jitter buffer for restoring timing of the media stream). These algorithms have to be tuned to the UMTS network for an application to run optimally. Certain algorithms are recommended to be used in a PSS application.

5.2.2 RTP aware mobile network

Certain UMTS streaming bearer implementation options are recommended to be used for the RTP flows of the PSS application.

5.2.3 Application-Network QoS interface

Values for the QoS parameters of the QoS profile (as in [4]) are recommended that correspond to the recommended application and network implementation options.

5.2.4 Design requirements

The design of the recommended PSS implementation has been driven by the following requirements:

- Optimal application quality
- Optimal network utilisation
- Fair to other services
- Operator controllable
- Leave room for vendor differentiation
- UMTS specific, but backwards compatible to legacy systems

6 Modelling

6.1 Use case

The streaming use case considered in the model assumes a streaming server located in the mobile operator's network or connected to the mobile network through a UMTS QoS aware interface. The streaming client is located in the mobile User Equipment.

6.2 Network modelling

6.2.1 Objectives of network modelling

The output of the network modelling work is a description of the dynamic characteristics of the network that is perceived by the application as the quality of service of the streaming bearer. This description provides the basic assumptions that the application can make about the network behaviour.

To find these dynamic characteristics of the network and to identify the relevant model parameters, the implications of different bearer implementation decisions are analysed in the following clauses.

TBD: The network model is to be supported and verified through network simulations, including in particular testing of the different mentioned RAN implementation options.

6.2.2 UTRAN streaming bearer implementation options

The most critical quality of service limitations in the UMTS network are at the RAN. The details and dynamics of the physical layer is not discussed, only layer-2 and higher implementation options. The listed options for streaming bearer implementation are not meant to be exhaustive, but only meant to show that alternatives for the implementation exist. The network model is constructed based on these mentioned alternatives. In an implementation other not mentioned options and algorithms might be used. The streaming service should actually work independently from the bearer implementation details, as stated in the PSS service requirements [6].

6.2.2.1 Link layer traffic handling modes

6.2.2.1.1 UTRAN RLC modes

There are three different traffic handling modes in UTRAN radio link layer (i.e. RLC) for transporting user-plane data: Transparent Mode, Unacknowledged Mode and Acknowledged Mode.

The transparent mode passes RLC SDUs without additional header information through. No SDU concatenation or padding is possible, and no integrity checking is provided. The transparent mode is primarily targeted to be used with circuit switched bearers. In a packet switched bearer, transparent mode is useful if the RLC SDU size is adapted to the RLC PDU size. In a general video (and some audio) stream, size of packets will vary and it can not always be an integer multiple of the size of an RLC-PDU. Therefore the transparent mode is not recommended to be used with the streaming traffic class.

The unacknowledged mode introduces a more flexible RLC SDU mapping to RLC PDUs, and thereby makes it suitable for general packet based traffic.

Transparent and unacknowledged mode L2 bearers normally carry delay sensitive traffic, as there is no delay introduced for error detection and correction.

The acknowledged mode provides error correction by applying re-transmission for erroneously received RLC blocks. As the acknowledged mode provides in-order delivery of SDUs, enabling the retransmission scheme results in added delay for SDUs whose RLC blocks are being re-transmitted. This appears as SDU delay jitter at the receiver.

The retransmission is not guaranteed to provide full reliability. Any yet unacknowledged RLC block may be discarded from a sender retransmission buffer (i.e. the retransmission attempts for that block stopped) if one of the following occurs: timer expiration, maximum number of retransmission attempts reached or sender retransmission buffer overflow.

This means, that RLC acknowledged mode can be flexibly configured to trade off the required reliability and maximum delay allowed in the RLC layer.

6.2.2.1.2 Implications of RLC mode decision

The PSS specification [3] defines a default value of 1 second pre-decoder buffering time. In practice, the pre-decoder buffering delay at the receiver can be in the order of multiple seconds (e.g. 2-4 seconds), which largely relaxes the delay jitter requirements in the network.

This implies that PSS applications are not overly sensitive to network delay jitter. In addition to that, streaming applications, particularly video, are much more sensitive to packet loss than delay jitter. It gives a worse viewing experience to see some video picture data missing, than having some video picture displayed late.

Therefore, despite the high delay jitter introduced by using RLC acknowledged mode, it is possible to use RLC retransmission for correcting damaged RLC blocks instead of reflecting directly the RLC loss up to the application.

Typically the radio link is adapted in UTRAN by transmission power (in GERAN by selection of coding schemes). Instead of relying on high transmission power (or protective coding scheme) in order to achieve a given SDU error ratio as requested by a given QoS profile, RLC re-transmissions can be used. It makes the implementation of the streaming bearer in the network cheaper at the expense of possibly introducing higher delay jitter.

6.2.2.2 Transport channel mapping

6.2.2.2.1 Dedicated or shared channel

Several schemes may be considered for channel allocation for streaming traffic class connection (downlink): dedicated channel (only streaming packets are sent through a reserved pipe), shared channel with other non-real time application packets (from the same user or not) or shared channel with other real time packet flows.

One of the latter two cases (i.e. when radio resources are shared among different flows) could be chosen by the RRM for the sake of better network resource utilisation, fairness, statistical multiplexing gain or some other reasons.

When mapping a streaming traffic class RAB to a radio bearer in UTRAN, the following applicable bearer services (transport channels) can be identified:

- DCH (Dedicated Channel) is an up- and downlink channel and is the main transport channel for packet data. DCH is dedicated to one flow and can be used for fairly constant bitrate packet traffic.

- DSCH (Downlink Shared Channel) is a common channel that can be shared among multiple users and multiple flows. DSCH downlink channel is particularly efficient for bursty Non Real Time packet traffic. It is good for asymmetric services, where downlink is the main transmission direction.

It should be noted that the support of DSCH is optional to terminals, therefore there must always be an alternative way to use only DCH, even though the DSCH would be the preferred option.

6.2.2.2.2 Implications of channel mapping decision

If a streaming source generates less traffic than its allocated bearer was set-up for, or generates a variable rate traffic, other services could use the unused resources. In this case a shared channel (DSCH) could be used. It is, however, difficult to guarantee QoS to each individual flow competing for the same shared resource. On the other hand, the network wants to make sure, that if a dedicated fixed-rate channel is allocated (DCH) the resource is utilised efficiently by the streaming application. These are the factors driving the choice of transport channel to be used for streaming.

It can be assumed that the effective radio throughput on average will be the same throughout the session independently of the transport channel chosen. Thus the application can assume, that it can transmit at this average radio throughput rate, and the variation of the available radio rate will be hidden behind a large enough scheduler buffer. Similarly, this buffering can also smooth out any temporal variation of the transmission rate around the average rate. Application rate adaptation is necessary when, for any reason this assumption proves not to be valid (e.g. due to different time window sizes used at the network and the application over what the rate is averaged).

The flow mapping decision puts different requirements on the rate adaptation algorithm required. Depending on the expected channel rate variation, a streaming application should be prepared to apply different rate measurement and rate adaptation schemes. Depending on the rate variation model, for example, rate measurements might be interpreted differently. A model of available rate variation in the network, can be built based on the understanding how a streaming bearer with different maximum and guaranteed bitrate QoS parameters is implemented in the network (e.g. mapped to what transport channel).

When a dedicated channel (DCH) with a given bitrate is allocated for the downlink flow, no available rate variation on the air interface is expected. However, if RLC re-transmission is used the rate variation due to retransmission can not always be neglected. The radio channel allocation is usually such, that the expected L2 throughput after re-transmission should reach the guaranteed bit rate.

When streaming is implemented over a shared channel (DSCH), the available bitrate for a single flow varies over time according to some pattern, which depends on many factors e.g. the scheduler algorithm used in the RAN, the load in the cell or some other rate allocation policies. The RRM however aims to maintain on average the guaranteed bitrate.

6.2.3 GERAN streaming bearer implementation options

TBD: Add for GERAN similar analysis as in the previous clause for UTRAN.

6.2.4 Core network characteristics

TBD.

6.2.5 Network model definition and parameters

TBD: Create formulas or description that model network behaviour:

- What type of throughput variation the network has?
- What kind of delay jitter the network introduces?
- How the reliability variation (i.e. SDU loss ratio) can be modelled?
- Etc.

TBD: Identify parameters and map them to QoS parameters of [4].

6.3 Application modelling

6.3.1 Objectives of application modelling

Whereas applications such as conferencing and non real-time text applications have distinctive traits, different streaming applications may have different traffic characteristics and bearer QoS requirements. Conferencing applications aim at transferring data with end-to-end minimum delay and non real-time text applications aim at transferring data without end-to-end loss. Conversely, different streaming applications may have very different delay versus loss trade-off.

The output of the streaming application modelling work is a description of the dynamic characteristics of the application, which is seen by the network as changing traffic characteristics. This description provides the basic assumptions that the network can make about the application behaviour.

6.3.2 Streaming application traffic characteristics

Packet sizes, packet rate, constant vs. variable bitrate

TBD: Elaborate on these characteristics of the application traffic.

6.3.3 Application implementation options

6.3.3.1 Delay and loss tolerance, required bearer QoS

The stream is normally time aligned at the receiver (using delay jitter compensating buffers), therefore the acceptable delay jitter over the transmission media is higher than the one tolerated in conversational applications. Ensuring low mean transfer delay of the packets of the stream is not as crucial as for conversational applications, as there is no interactive feedback required from the client to the server.

Due to the more relaxed delay requirements packet loss robustness can be increased by using different techniques (e.g. retransmission, FEC, packet interleaving).

6.3.3.2 Adaptation capability

Streaming application can be classified according to their capability to adapt their behaviour to changing network characteristics.

6.3.3.2.1 Server

Simple fixed bitrate bitstream transmission (low adaptation capability): Only basic encoding tools are used, such as error-resilient tools, but no scalability tools. The server can only send a pre-encoded bitstream at the designated target bitrate. Thus the server does not react and rely on any feedback from the receiver.

Rate adaptive transmission of a pre-encoded bitstream that includes scalability tool(s) (medium adaptation capability): Using a bitstream encoded with scalability tools, the server can adjust the transmission rate according to the feedback from the receiver and/or directly from the UMTS network. The server relies on the feedback to adjust its transmission bitrate.

Adaptive rate control (high adaptation capability): Equipped with "scalability transcoding " module, the server can adjust its transmission bitrate with fine granularity. It can also change other application traffic characteristics, such as the packet size, according to the characteristic of the network. It relies on feedback.

6.3.3.2.2 Client

Basic PSS client. Equipped with only a pre-decoder buffer and a decoder (player), this client can simply play the received video stream from the server. Only passive enhancement function is implemented (such as error concealment), but no reactive or proactive collaboration with the server.

Active PSS client. The receiver expects to see direct server action or change in the traffic characteristics as a response to an implicit request for server action through RTCP reports. The client in this case is actively monitoring and is aware of the QoS conditions.

6.3.4 Application model definition and parameters

TBD: Create formulas or description that model the application behaviour:

- What characteristics the streaming traffic has? (e.g. variable rate, packet sizes)
- How adaptive the streaming application (streaming server) is to varying network conditions and to the resulting quality of service variation?
- Etc.

TBD: Identify parameters and map them to QoS parameters of [4].

6.4 Model parameter interface

6.4.1 UMTS QoS Profile

Vertical awareness is facilitated by communicating the parameters of the application model to the network and vice versa the network model parameters to the application. The parameters and the interface for this communication has to be clarified and specified. The interface should be explicit and unambiguous in the context of the model in which it is interpreted.

The UMTS QoS concepts document [4] defines the "streaming traffic class" which term in the RTP usage model context refers to the model of the streaming application and the network implementing a streaming bearer. The "QoS Profile" of the streaming traffic class refers to the model parameter interface. The QoS profile with the QoS parameters is used as the interface for negotiating the application and network model parameters. The following sections give unambiguous interpretation of the QoS parameters as the parameters of the network and application models defined in sub-clause 6.2 and 6.3.

6.4.2 Interpretation of QoS parameters

TBD: Add interpretation of all parameters that can have ambiguous interpretation.

6.4.2.1 Maximum and guaranteed bitrate

6.4.2.1.1 Specification in [4]

Purpose of Maximum bitrate QoS parameter: Maximum bitrate can be used to make code reservations in UTRAN (or time-slot reservation in GERAN) in the downlink of the radio interface. Its purposes are 1) to limit the delivered bitrate to applications 2) to allow maximum wanted user bitrate to be defined for applications able to operate with different rates (e.g. non transparent circuit switched data).

Purpose of Guaranteed bitrate QoS parameter: Guaranteed bitrate may be used to facilitate admission control based on available resources, and for resource allocation within UMTS. The guaranteed bitrate can be understood as the throughput that the network tries to guarantee.

The UMTS bearer is not required to transfer traffic exceeding the Guaranteed bitrate. Quality requirements expressed by e.g. delay and reliability attributes only apply to incoming traffic up to the guaranteed bitrate.

6.4.2.1.2 Interpretation

TBD: The right interpretation is to be decided.

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 maximum 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.

The guaranteed and maximum bitrate parameters describe an application flow such that:

- average bitrate is described by the guaranteed bitrate parameter.
- allowed burstiness (i.e. instantaneous variations in bitrate around the average bitrate) is described by the maximum bitrate parameter.

TBD: Over which time period this is measured?

The bitrate parameters take into account the full RTP, UDP and IP headers. Thus, header compression is transparent to the set-up of the end-to-end UMTS bearer QoS.

6.4.2.2 SDU error ratio

TBD: The right interpretation is to be decided:

a) This is the target average SDU error ratio that the network attempts to keep all the time. What is the maximum SDU error ratio then?

b) This is the maximum SDU error ratio, and the actual SDU error ratio will be 95% of the time smaller than or equal to this value. What is the target average SDU error ratio then?

6.4.2.3 Residual bit error ratio

TBD: The right interpretation is to be decided:

a) This is the target residual bit error ratio that the network attempts to keep all the time. What is the maximum residual bit error ratio then?

b) This is the maximum residual bit error ratio, and the actual SDU error ratio will be 95% of the time smaller than or equal to this value. What is the target average residual bit error ratio then?

6.4.2.4 Maximum SDU size

To guarantee a given SDU error ratio, the larger the SDU size, the smaller RLC BLER the radio interface has to provide, which means that the reliability requirements for the radio link are more stringent. Maximum SDU size should be commonly considered with the required SDU error ratio. From the network viewpoint, smaller SDUs allow easier compliance to reliability requirements by relaxing the radio link adaptation. The application should always be conservative when specifying a maximum SDU size.

TBD: Are larger SDUs than the maximum SDU size discarded?

6.5 Monitoring the model state

In addition to the one-time initial set-up of the model parameters it is necessary to continuously (periodically) monitor the model state of the network and/or application in order to adapt to the dynamics of the model behaviour.

The network model state can be monitored by the application through for example RTCP reports. For example, the fraction of packets lost field in an RTCP receiver report tells about the reliability of the network in the last reporting period. This periodic feedback information from the network model could be utilised by the application to adapt its behaviour. Similarly, the application model state could be conveyed to the network by some means for possible network behaviour adaptation.

TBD: This clause is for further study.

6.6 Matching application and network characteristics

TBD.

7 Recommended PSS implementation

7.1 Network

Use RLC re-transmissions (even multiple retransmission attempts) while limiting the RLC delay. The maximum allowed RLC delay can be derived from the negotiated transfer delay QoS parameter.

TBD: Add more recommendations. Clarify bearer (PDP-context) mapping of the different media flows.

7.2 Application

TBD.

7.3 QoS Profile

TBD: Add recommendation for all the parameters of all the different media streamed over RTP.

Annex A: Change history

Change history							
Date	TSG #	TSG Doc.	CR	Rev	Subject/Comment	Old	New

Change history for TSG-SA4 PSM SWG internal working draft

Change history							
Date	TSG #	TSG Doc.	CR	Rev	Subject/Comment	Old	New
2001-11-28				0.0.1	Initial draft internal to SA4		S4-010613
2001-12-08				0.1.0	Revised scope and text identified as working draft	S4-010613	S4-010679