TSG-SA Working Group 1 (services) meeting #3 TSG S1#3(99) 362
Hampton Court, Surrey, UK 10th-12th May 1999 Agenda Item 6.3.3

3GPP TSG-SA-S1 QoS adhoc Tdoc TSGS1 QoS (99)012
Copenhagen 7th April 1999

Source: Nortel Networks

QoS Performance requirements for UMTS

The purpose of this paper is to discuss QoS from a user requirement perspective with the aim of producing a CR to 22.05 & 22.25

1 User requirements

A typical user is not concerned with how a particular service is provided. However, the user is interested in comparing one service with another in terms of universal, user-oriented performance parameters which apply to any end-to-end service. From a user's perspective, performance should be expressed by parameters which:

- Focus on user-perceivable effects, rather than their causes within the network
- Are independent of the networks internal design
- Take into account all aspects of the service from the user's point of view which can be objectively measured at the service access point
- Can be assured to a user by the service providers(s)

1.1.1 Classification of requirements

At the user level, communications applications can be classified into three broad classes depending on the time-sensitivity of the information:

i) Real-time streaming

A real-time streaming application is one that delivers time-based information in real-time, where time-based information is user data that has an intrinsic time component. Video, audio and animation are examples of time-based information, in that they consist of a continuous sequence of data blocks that must be presented to the user in the right sequence at pre-determined instants.

ii) Real-time block transfer

A real-time block transfer application delivers one or more blocks of data, each of which must be delivered within a deadline. Each block of data can be time-based or non time-based. There is no intrinsic timing relationship between consecutive data blocks, which is a key difference relative to a real-time streaming application. However, since each block delivery is time-sensitive, it is considered a real-time application.

iii) Non real-time

A non real-time application is one that does not carry time-sensitive information. Non real-time applications are similar to real-time block transfer applications in that they both deliver blocks of data (which can be time-based or non time-based information), the key difference is the urgency of delivery.

1.2 Key parameters impacting the user

Delay

Delay is an important parameter from a user perspective, manifesting itself in a number of ways, including the time taken to establish a particular service from the initial user request and the time to receive specific information once the service is established.

Delay variation

Delay variation is generally included as a performance parameter since it is very important at the transport layer in packetised data systems due to the inherent variability in arrival times of individual packets. However, as will be discussed later, services that are highly intolerant of delay variation will usually take steps to remove (or at least significantly reduce) the delay variation by means of buffering, effectively eliminating delay as perceived at the user level (although at the expense of adding additional fixed delay).

Information loss

Information loss is an obvious performance parameter to the user, since it has a very direct affect on the quality of the information finally presented to the user, whether it be voice, image, video or data.

2 Performance requirements for real-time streaming applications

2.1 Audio

2.1.1 Conversational voice

Requirements for conversational voice are heavily influenced by one-way delay. In fact, there are two distinct effects of delay. The first is the creation of echo in conjunction with two-wire to 4-wire conversions or even acoustic coupling in a terminal. This begins to cause increasing degradation to voice quality for delays of the order of tens of milliseconds, and echo control measures must be taken at this point (provision of echo cancellers etc). Considering that delays in wireless systems are typically well in excess of these values, such echo control measures are mandatory. The second effect occurs when the delay increases to a point where it begins to impact conversational dynamics, ie the delay in the other party responding becomes noticeable. This occurs for delays of the order of several hundred milliseconds.

For connections with adequate echo control, audio transfer delay requirements depends on the level of interactivity of the user task. To preclude difficulties related to the dynamics of voice communications, ITU-T Recommendation G.114 [2] recommends the following general limits for one-way transmission time (assuming echo control already taken care of):

0 to 150 ms	preferred range
150 to 400 ms	acceptable range (but with increasing degradation)
above 400 ms	unacceptable range

However, the human ear is highly intolerant of short-term delay variation (jitter) and a limit as low as 1 msec is reported in [3]. As a practical matter, for all voice services, delay variation due to variability in incoming packet arrival times must be removed with a jitter buffer.

Requirements for information loss are influenced by the fact that the human ear is tolerant to a certain amount of distortion of a speech signal. In digital transmission systems a prime source of distortion is due to the use of low bit-rate speech compression codecs and their performance under channel error conditions. Detailed requirements are dependent on the specific coder and channel error statistics, but subjective tests [4] have shown that acceptable performance is typically obtained with frame erasure rates (FER) up to about 3 %.

2.1.2 Voice messaging

Requirements for information loss are essentially the same as for conversational voice (ie dependent on the speech coder), but a key difference here is that there is more tolerance for delay since there is no direct conversation involved. The main issue, therefore becomes one of how much delay can be tolerated between the user issuing a command to replay a voice message and the actual start of the audio. There is no precise data on this, but based on other studies [5,6] related to the acceptability of stimulus-response delay for telecommunications services, a delay of the order of a few seconds seems reasonable for this application.

2.1.3 Streaming audio

Streaming audio is expected to provide better quality than conventional telephony, and requirements for information loss in terms of packet loss will be correspondingly tighter.

However, as with voice messaging, there is no conversational element involved and delay requirements can be relaxed, even more so than for voice-messaging.

2.2 Video

2.2.1 Videophone

Videophone as used here implies a full-duplex system, carrying both video and audio and intended for use in a conversational environment. As such, in principle the same delay requirements as for conversational voice will apply, i.e. no echo and minimal effect on conversational dynamics, with the added requirement that the audio and video must be synchronized within certain limits to provide “lip-synch” [3]. In fact, due to the long delays incurred in even the latest video codecs, it will be difficult to meet these requirements.

Once again, the human eye is tolerant to some loss of information, so that some degree of packet loss is acceptable depending on the specific video coder and amount of error protection used. It is expected that the latest MPEG-4 video codecs will provide acceptable video quality with frame erasure rates up to about 1%.

2.2.2 One-way video

The main distinguishing feature of one-way video is that there is no conversational element involved, meaning that the delay requirement will not be so stringent, and can follow that of streaming audio.

2.2.3 Data

2.2.4 Two-way control telemetry

As will be seen later, most data services are either non-real-time, or not streamed. Two-way control telemetry is included here as an example of a data service which does require a real-time streaming performance. Clearly, two-way control implies very tight limits on allowable delay and a value of 250 msec is proposed, but a key differentiator from the voice and video services in this category is the zero tolerance for information loss (obvious if you are controlling an important industrial process, for example).

Requirements for real-time streaming services are summarised in Table 1 below.

Table 1: End-user Performance Expectations - Real-time Streaming Services

Medium	**Application**	**Degree of symmetry**	**Customer demand**	**Amount of traffic**	**Data rate**	Key performance parameters and target values			
						One-way Delay	**Delay variation**	**Information loss**	**Other**
Audio	Conversational voice	Two-way	High	High	4-13 kb/s	<150 msec preferred <400 msec limit	< 1 msec	< 3% FER	
Audio	Voice messaging	Primarily one-way	High	Medium	4-13 kb/s	< 1 sec for playback < 2 sec for record	< 1 msec	< 3% FER	
Audio	High quality streaming audio	Primarily one-way	Low	Low	32-128 kb/s	< 10 sec	< 1 msec	< 1% FER	
Video	Videophone	Two-way	Low	Low	32-384 kb/s	< 150 msec preferred <400 msec limit		< 1% FER	Lip-synch : < 100 msec
Video	One-way	One-way	Low	Low	32-384 kb/s	< 10 sec		< 1% FER	
Data	Telemetry - two-way control	Two-way	Low	Low	<28.8 kb/s	< 250 msec	N.A	Zero	

3 Performance requirements for real-time block transfer applications

3.1 Data

Although there may be some exceptions, as a general rule it is assumed that from a user point of view, a prime requirement for any data transfer application is to guarantee essentially zero loss of information. At the same time, delay variation is not applicable. The different applications therefore tend to distinguish themselves on the basis of the delay which can be tolerated by the end-user from the time the source content is requested until it is presented to the user.

3.2 Web-browsing

This is perhaps the most obvious 3G data application. For the purposes of this document, in this category we will refer to retrieving and viewing the HTML component of a Web page, other components eg images, audio/video clips are dealt with under their separate categories. From the user point of view, the main performance factor is how fast a page appears after it has been requested. A value of 2-4 seconds per page is proposed.

3.3 Bulk data

This category includes file transfers, and is clearly influenced by the size of the file. As long as there is an indication that the file transfer is proceeding, it is reasonable to assume some what longer tolerance to delay than for a single Web-page.

3.4 High-priority transaction services (E-commerce)

The main performance requirement here is to provide a sense of immediacy to the user that the transaction is proceeding smoothly. A value of 2-4 seconds is suggested to be acceptable to most users.

3.5 Still image

This category includes a variety of encoding formats, some of which may be tolerant to information loss since they will be viewed by a human eye. However, given that even single bit errors can cause large disturbances in other still image formats, it is argued that this category should in general have zero information loss. However, delay requirements for still image transfer are not stringent, given that the image tends to be built up as it is being received, which provides an indication that data transfer is proceeding.

3.6 Interactive games

Requirements for interactive games are obviously very dependent on the specific game, but it is clear that demanding applications ("twitch games") will require very short delays, and a value of 250 msecs is proposed, consistent with demanding interactive applications.

3.7 Telemetry (monitoring)

Monitoring covers a wide range of applications, but in this category it is taken to apply to relatively low priority activities, eg status updating, rather than control.

3.8 Telnet

Telnet is included here with a requirement for a short delay in order to provide essentially instantaneous character echo-back.

3.9 E-mail (server access)

E-mail is generally thought to be a store and forward service which in principle can tolerate delays of several minutes or even hours. However, it is important to differentiate between communications between the user and the local email server and server to server transfer. When the user communicates with the local mail server, there is an expectation that the mail will be transferred quite rapidly, although not necessarily instantaneously. Consistent with the research findings on delay tolerance for Web-browsing, a requirement of 2-4 seconds is proposed.

Requirements for real-time block transfer services are summarised in Table 2 below.

Table 2: End-user Performance Expectations - Real-time Block Transfer Services

Medium	**Application**	**Degree of symmetry**	**Customer Demand**	**Amount of traffic**	**Data rate**	**Amount of data**	Key performance parameters and target values		
							One-way delay	**Delay variation**	**Informatio n loss**
Data	Web-browsing - HTML	Primarily one-way	High	High		10 KB	< 4 sec /page	N.A	Zero
Data	Bulk data transfer/retrieval	Primarily one-way	High	Medium		10 KB- 10 MB	< 10 sec	N.A	Zero
Data	Transaction services – high priority eg e-commerce, ATM	Two-way	High	Low		< 10 KB	< 4 sec	N.A	Zero
Data	Still image	One-way	Medium	Medium		< 1 MB	< 10 sec	N.A	Zero
Data	Interactive games	Two-way	Medium	Low		< 1 KB	< 250 msec	N.A	Zero
Data	Telemetry - monitoring	One-way	Low	Low	<28.8 kb/s		< 10 sec	N.A	Zero
Data	Telnet	Two-way (asymmetric)	Low	Low		< 1 KB	< 250 msec	N.A	Zero
Data	E-mail (server access)	Primarily One-way	High	High		< 10KB	< 4 sec	N.A	Zero

4 Performance requirements for non real-time applications

In principle, the only requirement for applications in this category is that information should be delivered to the user essentially error free. However, there is still a delay constraint, since data is effectively useless if it is received too late for any practical purpose.

4.1 Fax

Fax is included in this category since it is not normally intended to be an accompaniment to real-time communication. Nevertheless, there is an expectation in most business scenarios that a fax will be received within about 30 seconds. The information loss requirement is based on established wireline requirements, but further work is needed to extend this to take into account the bursty error nature of a wireless channel.

4.2 Low priority transaction services

An example in this category is Short Message Service (SMS). 30 seconds is proposed as an acceptable delivery delay value in [7].

4.3 Email (server to server)

This category is included for completeness, since as mentioned earlier, the prime interest in email is in the access time. There is a wide spread in user expectation, with a median value of several hours.

4.4 Usenet

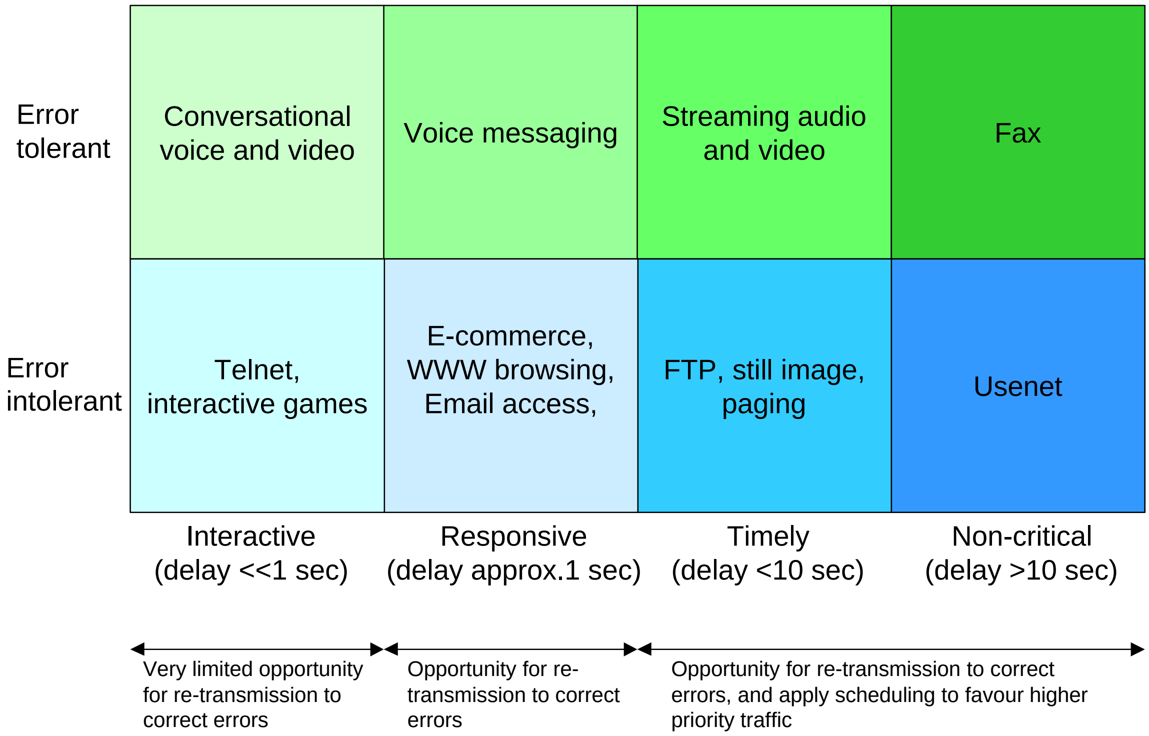
Usenet is a world-wide distributed discussion system. It consists of a set of "newsgroups" with names that are classified hierarchically by subject. "Articles" or "messages" are "posted" to these newsgroups by people on computers with the appropriate software -- these articles are then broadcast to other interconnected computer systems via a wide variety of networks. This is a very low priority service, with corresponding delay requirements.

Requirements for non real-time services are summarised in Table 3 below.

Table 3: End-user Performance Expectations - Non Real-time Services

Medium	Application	Degree of symmetry	Customer demand	Amount of traffic	Data rate	Amount of data	Key performance parameters and target values		
							One-way delay	**Delay variation**	**Information loss**
Data	Fax	Primarily one-way	High	High	9.6 kb/s		< 30 sec /page	N.A	<10-6 BER
Data	Email (server to server transfer)	One-way	High	High		<10 KB	Can be several hours	N.A	Zero
Data	Usenet	Primarily one-way	Medium	High		Can be 10's MB/day	Can be several hours	N.A	Zero
Data	Transaction services – lower priority eg SMS	Two-way	Medium	Low		< 10 KB	< 30sec	N.A	Zero

5 Summary of performance requirements into Quality of Service categories



	Interactive (delay <<1 sec)	Responsive (delay approx.1 sec)	Timely (delay <10 sec)	Non-critical (delay >10 sec)
Error tolerant	Conversational voice and video	Voice messaging	Streaming audio and video	Fax
Error intolerant	Telnet, interactive games	E-commerce, WWW browsing, Email access,	FTP, still image, paging	Usenet

Very limited opportunity for re-transmission to correct errors

Opportunity for re-transmission to correct errors

Opportunity for re-transmission to correct errors, and apply scheduling to favour higher priority traffic

6 References

1. “Residential broadband Internet services and application requirements” T.Kwok, IEEE Communications Magazine, June 1997

2. ITU-T Recommendation G.114 “One-way transmission time”, 1996

3. “Human perception of jitter and media synchronisation”, R. Steinmetz, IEEE JSAC Vol 14, No. 1, Jan 1996

4. “Subjective evaluation of speech codecs for CDMA”, L.Thorpe, P. Coverdale, TR96-001, June 1996

5. “Response time and display rate in human performance with computers “ B. Shneiderman, Computing Surveys, Vol. 16, No.3 Sept. 1984

6. “Using customer perception in setting objectives for Intelligent Network services”, D.M.MacDonald, S. Archambault , ITC-14, 1994

7. “Report on UMTS/IMT2000 spectrum requirements”, UMTS Forum, Oct 1997

8. ITU-T Draft Recommendation G.109 “Definition of categories of voice quality”, Dec 1998