Information • Knowledge • Systems *Management* 4 (2004) 55–76
IOS Press

QoS Requirements of Network Applications on the Internet

Yan Chen, Toni Farley and Nong Ye*

Department of Industrial Engineering, Arizona State University, Tempe, AZ, USA

ABSTRACT: In this paper, we present quality of service (QoS) metrics for various network applications based on human factors and technology attributes. The first term, human factors, addresses human perception of different kinds of media, such as conventional text, audio and video. The second term, technology attributes, represents the different technological aspects of these network applications, such as time-dependence and symmetry. Both of these terms are key factors that lead to variations of requirements for QoS. Establishing these requirements is paramount to providing QoS on computer networks and the Internet. With the metrics presented in this paper we can provide the criteria necessary for such QoS assurance.

INTRODUCTION

Information services on the Internet come in varying forms, such as web browsing, e-mail, and multimedia on demand. The main motivation behind the design of next-generation computer and communications networks is providing universal and easy access to these various types of information services on a single multi-service Internet. This means that all forms of communications (video, voice, data and signaling), along with all types of services (from plain text web pages to multimedia applications), are bonded in a single-service platform through Internet technology.

Media are transmitted on a network by bit streams. Each bit stream is represented by a flow of bits of 1s and 0s through the network. Although all media are transmitted in the same digital form, each application generating a bit stream has an associated set of service characteristics. Regardless of the type of media (picture or voice) there are different QoS requirements perceived from the user's standpoint that determine whether the user finds this type of service acceptable. Examples of user perception include phone-to-phone delay and playing music online.

In general a phone-to-phone delay, from a user's perspective, should be no more than 150 milliseconds (Silveira et al, 1996) to allow for appropriate and easy understanding. Outside of this limit, the user will become annoyed and find the service unacceptable. To support this kind of a call over the Internet, the voice media need to be transmitted with a constant and reliable bit transmission rate, low jitter, and a low error and loss rate.

User perception is also important when a user plays music online. If the variation of delay across the Internet is too large the song sounds desultory, and the user stops listening. When a network cannot provide the corresponding support necessary to make a service acceptable, the user feels that this kind of service is senseless.

The Internet today is made up of a simple convergence of joining application media bit streams. This convergence does not take into account the service characteristics associated with each type of

* Corresponding author: Professor Nong Ye, Department of Industrial Engineering, Arizona State University, Box 875906, Tempe, AZ 85287-5906, USA. Tel.: +1 480 965 7812; Fax: +1 480 965 8692; E-mail: nongye@asu.edu.

application. This means that the Internet does not currently provide any quality assurance.

Concise specification of service requirements is vital to realizing quality assurance on computer networks. Since different applications have different service features, each application should specify its explicit requirements to a computer network in order to achieve the desired QoS. If there are no requirements given, the network will take for granted that any level of service is acceptable, and therefore can provide any level of network support. The types of network support provided, such as bandwidth and priority in a router's queue, may lead to delay and jitter that can then render the service unacceptable. Consequently, if we want to satisfy QoS of various applications in next-generation networks, the first step is to identify the QoS metrics of each network application.

In this paper, we first analyze the influence of two key factors, human factors and technology attributes, on QoS requirements. Next, we present network application classification based on the application's different service characteristics, resulting mainly from these two key factors. Finally, based on existing literature, we systematically propose numerical measures, which can quantitatively represent the QoS requirements of various applications with different service characteristics in Multi Service Networks, along with the rationale behind these choices.

QOS MEASURES

In general, QoS has three attributes to measure the output performance of a process: timeliness, precision and accuracy (Ye, 2002). Timeliness measures the time taken to produce the output of the process. Precision measures the amount or quantity of the produced output. Accuracy measures the correctness of the produced output, usually relating to the content of the output. Specific measures of the three QoS attributes depend on the process of interest. Existing work on QoS of computer networks has used the following QoS measures.

Response Time Expected by Users: The users' expected response time is the time elapsed between sending a request and the reception of the first response by the user.

Delay: The network transmit delay is the time elapsed between the emission of the first bit of a data block by the transmitting end-system, and its reception by the receiving end-system.

Jitter: In transmission technology, jitter refers to the variation of delay generated by the transmission equipment.

Data Rate: Data rate refers to the raw data rate of encoded multimedia data before transmission, that is, the rate in which data are encoded.

Required Bandwidth: The required bandwidth is defined by the required data transfer rate, measured in bits per second, of each specific application in telecommunication. This metric includes raw data and overhead.

Loss Rate: The bit loss rate is the number of bits lost between two points in telecommunications after transmission.

Error Rate: The bit error rate is the frequency of erroneous bits between two points in telecommunication after transmission.

TECHNOLOGY ATTRIBUTES OF APPLICATIONS

Different applications can be characterized by certain technology attributes. Applications in one class of technological attribute may not share the same QoS requirements as those in another class. The technology attributes used to classify applications in this paper are defined in this section.

Time Dependence

Applications can be classified by their time dependency requirements. We classify all applications into two time dependant categories: real time (RT) applications and non real time (NRT) applications.

Real time is defined by the Oxford Dictionary of Computing as: "A system in which the time at which the output is produced is significant. This is usually because the input corresponds to some movement in the physical world, and the output has to relate to that same movement. The lag (delay) from the input time to output time must be sufficiently small for acceptable timeliness".

RT applications can further be divided into soft real time and hard real time. The main difference between soft and hard real time is that hard real time applications will fail should QoS requirements not be met. In this paper we group all soft and hard real time applications as RT.

In RT applications, the network needs to deliver time-based information without changing its built-in time properties. For adequate user satisfaction, we need to maintain more stringent delay and jitter requirements for RT applications. The delay requirements must be strict in order to maintain system timing. The jitter requirements are essential to transmitting data at a constant, reliable rate.

NRT applications are any applications that do not have stringent timing requirements. This type of application does not fail if timeliness metrics are not met, nor does it require timing accuracies to be considered acceptable. NRT applications, which do not have time-based sensitivity requirements, are mostly concerned with delay.

Symmetry

An application's symmetry property allows us to classify all applications into two categories: symmetric applications and asymmetric applications.

In symmetric applications, requests and responses are comparable in terms of resource consumption. Videophony is an example of a symmetric audio and video application. This application requires equal resource consumption on both host machines. In this case, the request consumes an equivalent amount of resources as the response.

In asymmetric applications, requests are considerably less resource consuming than responses. An example of an asymmetric application is Video-on-Demand (VoD). This application is highly asymmetric since it consumes much more resources in the VoD Server (response) than in client machine (request).

The grades of symmetry in multimedia services result in different grades of interactivity, which in turn lead to different response time and delay requirements. For example, Videophony requires an almost instantaneous feedback to satisfy the human perception of conversation, whereas VoD does not. This immediate feedback requires minimal delay and jitter on the network.

HUMAN PERCEPTION OF QOS

In this section we present the human perception of the QoS measures presented previously. These are the perceptions from a user's standpoint, which differ from the perception of these QoS aspects from a technology standpoint.

Human Perception of Delay

In conventional text and data networking, delay requirements are the least stringent. The response time in these types of applications can increase from 2 to 5 seconds before becoming unacceptable (Szuprowicz, 1995). Even at 5 seconds, the response time may still be considered tolerable.

End-to-end audio latency refers to the latency experienced across a link. In other words, the time between when the sound is sent on one end of a link, and when it is received on the other. The relationship between End-to-end Audio Latency and the Human Ear is shown in Table 1. (Szuprowicz, 1995).

In interactive applications of real time sound transmission, as well as in virtual reality, the overall one-way delay needs to be short in order to give the user an impression of real time responses. A maximum value on the order of 0.1 to 0.5 seconds is required to accomplish this goal. (Fluckiger, 1995).

Based on subject tests, the International Telecommunication Unit (ITU) G.114 specification recommends less than a 150 millisecond one-way end-to-end delay for high-quality real time traffic in telecommunication. (ITU G.114, 1996) The G.114 time limits are shown in Table 2.

Table 1
End-to-end Audio Delay and the Human Ear

Audio Delay (ms)	Effect of Delay on Human Voice Perception
> 600	Speech is unintelligible and incoherent
600	Speech is barely coherent
250	Speech is annoying but comprehensible
100	Imperceptible different between audio and real speech
50	Humans cannot distinguish between audio and real speech

Table 2
G.114 Limits for One-way Transmission Time

One-way transmission time	Effect of Delay on Human Voice Perception
0 to 150 ms	Acceptable for most users
150 to 400 ms	Acceptable, but had impact
400 ms and above	Unacceptable

In this paper, we adopt the G.114 limit of delay, 150 milliseconds, for most real time traffic that does not require a high level of interaction, such as teleconferencing, and streaming audio and video. On the contrary, interactive applications like Voiceover IP (VoIP) and Videophony require the value of delay given in Table 1 of 100 milliseconds due to the desire for instantaneous feedback.

In video applications, it is necessary to preserve the timing relationships between audio and video streams, as well as the timing relationships within individual video streams. First, we consider the way in which human beings perceive sounds and images.

Humans are much more sensitive to alteration of audio than of visual signals (Fluckiger, 1995). Lip-synchronization is defined as the difference between the visual reception of a video, and the auditory reception of its sound. For adequate user perception, lip-synchronization should not exceed 100ms (Fluckiger, 1995)

In many applications, an audio and a video stream are transmitted concurrently. Compressor/ Decompressor (codec) technology processes audio and video data. In some codec technology, audio and video data are put into separate packets. In such cases, the audio stream must have priority over the video stream. Since the delay of a real time stream should be less than 150 milliseconds, the delay of a video stream should not exceed 250 milliseconds to preserve the 100 milliseconds limit on lip-synchronization. In other codec technologies, audio and video data are put into the same packets. Under these circumstances, we have to satisfy the delay requirement of the audio stream, which has a maximum limit of 150 milliseconds. In this paper, we consider the case in which audio and video data are in the same packets.

Human Perception of Jitter

Jitter is an essential performance parameter of a network intended to support real time sound and image media. Of all information types, real time sound is the most sensitive to network jitter.

The solution to overcoming jitter in real time sound transmissions is for the receiving system to wait a sufficient length of time, called a delay offset, before playing received sounds. Incoming blocks are stored in a temporary memory location, called a buffer. After an adequate delay, these sound blocks can be played smoothly. This process is called delay equalization or delay compensation.

One of the shortcomings of the delay equalization technique is that of the additional delay, which is introduced at the sink end. Therefore, the value of jitter has a close relationship with the delay requirement of a specific application.

With typical personal computers or workstations as end-systems, the variation of the network transit delay should, in general, not exceed 100ms for CD-quality compressed sound and 400ms for telephone-quality speech. For multimedia applications with strong bounds on transit delay, like virtual reality, this jitter should not exceed 20–30 ms (Fluckiger, 1995).

Video applications have different human perception requirements for the minimization of jitter. For example, the variation of network transit delay should not exceed 50ms for HDTV quality, 100ms for broadcast quality, and 400ms for videoconferencing (Fluckiger, 1995).

Human Perception of Error

Three considerations must be taken into account when defining the acceptable human requirements for error. The first is current retransmission techniques. In conventional text and data transmissions, retransmission techniques are employed to correct errors in transmitted data. In the case of real time audio and video, retransmission techniques are not appropriate. Secondly, we need to realize that humans usually tolerate a high level of transmission error. This tolerance is generally derived from human experiences in comparable situations. Finally, in the case of compressed bit streams, erroneous or lost packets may invalidate greater numbers of subsequent packets. As a result, lower values of network error and loss rates must be maintained when audio and video data are in a higher compression rate scheme.

Furthermore, limits on error rates vary by application. The manner in which data are presented to humans (data type) dictates limits on error rates. In the case where data are presented to human users without recording for further processing, the residual bit error rate of a telephone-quality audio stream should be lower than 10^{-2}. The residual bit error rate of a CD-quality audio stream should be lower than 10^{-3} in the case of an uncompressed format, and lower than 10^{-4} in the case of a compressed format (Fluckiger, 1995).

The end-to-end network bit error rate, before possible error recovery between end-systems, should not exceed 10^{-6} for HDTV quality, 10^{-5} for broadcast TV quality, and 10^{-4} for videoconference quality (Fluckiger, 1995). These figures are for compressed video data streams.

CLASSIFICATION OF APPLICATIONS

Based on the different technology attributes and human perception characteristics discussed in this paper, we propose the taxonomy of traffics on network in this section. The technology attributes previously defined in this paper serve as the basis for classifying multimedia applications into four categories based on time dependence and symmetry.

Non Real Time & Asymmetric

Non real time, asymmetric applications generate traffic that is considered “best effort” traffic. This means that the traffic is transmitted using a best effort protocol. The best effort protocol does not provide any specific QoS reliability and simply processes traffic on a first-come, first-served basis. The most common NRT, asymmetric network applications are as follows:

Web Browsing — Hyper Text Transport Protocol (HTTP)
Enhanced Web Browsing
Email
File Transfer Protocol (FTP)
Telnet.

Non Real Time & Symmetric

Most common NRT network applications are asymmetric. This makes sense because NRT traffic usually consists of client machines requesting data and services from a host machine. However, there is one commonly used NRT application that is symmetric in nature and thus requires equivalent resource consumption at each host endpoint of the communication link:

Internet Relay Chat.

Real Time & Asymmetric

Real time applications inherently have more stringent QoS requirements due to the nature of real time transmissions. Common real time, asymmetric applications are as follows:

Audio Broadcasting
Video Broadcasting
Interactive Audio on Demand
Interactive Video on Demand
Telemetry.

Real Time & Symmetric

Common real time, symmetric applications include those which are conversational in nature:

Teleconferencing (including Audio, Audiographics, and Video Conferencing)
Videophony
VoIP.

QOS METRICS FOR EACH CLASS OF APPLICATION

In this section we systematically propose the desired QoS metrics for each application listed in the previous section. We base these metrics on knowledge about human perception of QoS and the technology features of various traffics put forth in this paper. Capacities of networking equipment are described in Appendices I and II, which can be used to examine the capability of the current networking technology in supporting QoS metrics.

QoS Metrics for Web Browsing

The QoS metrics for conventional web-browsing using HTTP are outlined in Table 3.

Table 3
QoS Metrics for Web Browsing

Traffic Class	Technology Attributes	QoS Metrics						
		Timeliness			Preciseness			Accuracy
		Response time Expected by Users	Delay (ms)	Jitter (ms)	Data Rate (bps)	Required Bandwidth (bps)	Loss Rate	Error Rate
Web Browsing	Non Real Time and Asymmetric	2–5 Seconds	< 400	N/A	< 30.5 K	< 30.5 K	Zero	Zero
Network Capacity								
Link: Refer to Appendix I				Router: Refer to Appendix II				

Requirements for conventional web browsing are mainly influenced by response time, which is limited to no more than 5 seconds. Delay less than 400 ms is expected for all best effort network traffic although, under some circumstances, delays greater than 400 ms may be considered acceptable. Jitter is not applicable in HTTP because it has little impact on traditional static text and picture web browsing. Data rate and required bandwidth for typical applications of web browsing are less than 30.5 kbps. The expected loss rate and error rate are zero since HTTP is a reliable transfer protocol in which erroneous packets are sent again using a retransmission policy.

QoS Metrics for Enhanced Web Browsing

Enhanced web browsing refers to high-priority transaction services, such as those related to e-commerce. The primary performance requirement for enhanced HTTP is to provide a sense of immediacy to the user. This sense of immediacy acts as a security blanket by assuring the user that a transaction is preceding smoothly. The QoS metrics for enhanced web browsing are outlined in Table 4.

Table 4
QoS Metrics for Enhanced Web Browsing

Traffic Class	Technology Attributes	QoS Metrics						
		Timeliness			Preciseness			Accuracy
		Response time Expected by Users	Delay (ms)	Jitter (ms)	Data Rate (bps)	Required Bandwidth (bps)	Loss Rate	Error Rate
Enhanced Web Browsing	Non Real Time and Asymmetric	2–4 Seconds	< 400	N/A	< 24 K	< 24 K	Zero	Zero
Network Capacity								
Link: Refer to Appendix I				Router: Refer to Appendix II				

For enhanced web browsing, a value of 2–4 seconds for response time is suggested as acceptable to most users. The best effort traffic delay specification of less than 400 ms is sufficient to satisfy this 2–4 second response time requirement. Again, jitter is not applicable to enhanced HTTP. Compared with traditional web browsing, enhanced web browsing requires less bandwidth. This is due to the fact that web pages used in a transaction service are typically simpler and more practical than web pages that are customarily browsed. However, this bandwidth requirement may be increased by adding security measures, such as encryption and SSL. The expected loss rate and error rate are zero.

QoS Metrics for Email

Email is a store and forward type of service which, in principle, can tolerate delays of several minutes or even hours from a technology perspective. However, it is important to differentiate between communications between the user and the local email server and server to server transfers. When a user communicates with a local mail server, there is an expectation that the mail will be transferred quite rapidly, although not necessarily instantaneously. The QoS metrics for email are given in Table 5.

Table 5
QoS Metrics for Email

Traffic Class	Technology Attributes	QoS Metrics						
		Timeliness			Preciseness			Accuracy
		Response time Expected by Users	Delay (ms)	Jitter (ms)	Data Rate (bps)	Required Bandwidth (bps)	Loss Rate	Error Rate
Email	Non Real Time and Asymmetric	2–5 Seconds	Low	N/A	< 10 K	< 10 K	Zero	Zero
Network Capacity								
Link: Refer to Appendix I				Router: Refer to Appendix II				

Consistent with our research findings on response time for web browsing, a requirement of 2–5 seconds for email response time is proposed. The requirement of expected network delay is low, and jitter is not applicable. The typical bandwidth requirement for email is less than 10 kbps. The expected loss rate and error rate are zero.

QoS Metrics for FTP

Typically, FTP service requires a relatively high bandwidth, which is clearly influenced by the size of the file. Meanwhile, the size of files will also impact the time expected to finish the service. We also need to differentiate between the time elapsed between command transmissions, and file transmissions between the user and the FTP server. User expectation is that a file should begin transmission soon after the command sent out. QoS metrics for FTP are given in Table 6.

Again, a requirement for response time of 2–5 seconds is proposed based on our web browsing research findings. As for delay, as long as there is some indication that the file transfer is proceeding, it is reasonable to assume a somewhat greater tolerance for delay than the expected delay for a single web page. Jitter is not applicable, and the expected loss rate and error rate are zero.

Table 6
QoS Metrics for FTP

Traffic Class	Technology Attributes	QoS Metrics						
		Timeliness			Preciseness			Accuracy
		Response time Expected by Users	Delay (ms)	Jitter (ms)	Data Rate (bps)	Required Bandwidth (bps)	Loss Rate	Error Rate
FTP	Non Real Time and Asymmetric	2–5 Seconds	Med	N/A	High	High	Zero	Zero
Network Capacity								
Link: Refer to Appendix I				Router: Refer to Appendix II				

QoS Metrics for Telnet

Requirements for quality of service of telnet applications are focused on a short delay in order to provide essentially instantaneous echo-back of characters. The QoS metrics for telnet are given in Table 7.

Table 7
QoS Metrics for Telnet

Traffic Class	Technology Attributes	QoS Metrics						
		Timeliness			Preciseness			Accuracy
		Response time Expected by Users	Delay (ms)	Jitter (ms)	Data Rate (bps)	Required Bandwidth (bps)	Loss Rate	Error Rate
Telnet	Non Real Time and Asymmetric	< 2 seconds	< 250	N/A	< 1 K	< 1 K	Zero	Zero
Network Capacity								
Link: Refer to Appendix I				Router: Refer to Appendix II				

For telnet applications, a response time of less than 2 seconds is suggested in order to satisfy the human perception of users and a delay less than 250 millisecond is recommended. Since most of the data transmitted via telnet are plain characters, the requirement for bandwidth is quiet small, typically less than 1 kbps. Jitter is not applicable, and the expected loss and error rates are zero.

QoS Metrics for Internet Relay Chat

Internet Relay Chat is a multi-user chat system, where users convene on “channels” (a virtual place, usually with a topic of conversation) to talk in groups, or privately. Requirements for quality of service in Internet Relay Chat are focused on short response times in order to provide essentially instantaneous character feedback. The QoS metrics for Internet Relay Chat are given in Table 8.

Table 8
QoS Metrics for Internet Relay Chat

Traffic Class	Technology Attributes	QoS Metrics						
		Timeliness			Preciseness			Accuracy
		Response time Expected by Users	Delay (ms)	Jitter (ms)	Data Rate (bps)	Required Bandwidth (bps)	Loss Rate	Error Rate
Internet Relay Chat	Non Real Time and Symmetric	1 Second	< 200	N/A	< 1 K	< 1 K	Zero	Zero
Network Capacity								
Link: Refer to Appendix I				Router: Refer to Appendix II				

A response time of less than 1 second is suggested to satisfy users' expectations. A delay less than 200 milliseconds is recommended to guarantee a short response time. In general, Internet Relay Chat consists of plain character transmission, so the requirement for bandwidth is quiet small, typically less than 1 kbps. Jitter is not applicable in Internet Relay Chat. With the existence of a retransmission policy for erroneous packets, the expected loss rate and error rate are zero.

QoS Metrics for Audio Broadcasting

User response time in Audio Broadcasting refers to the time elapsed between transmitting commands and receiving feedback from the audio server. The QoS metrics for audio broadcasting are given in Table 9.

Table 9
QoS Metrics for Audio Broadcasting

Traffic Class	Technology Attributes	QoS Metrics						
		Timeliness			Preciseness			Accuracy
		Response time Expected by Users	Delay (ms)	Jitter (ms)	Data Rate (bps)	Required Bandwidth (bps)	Loss Rate	Error Rate
Audio Broadcasting	Real Time and Highly Asymmetric	2–5 Seconds	<150	<100	56–64K	60–80 K	<0.1%	<0.1%
Network Capacity								
Link: Refer to Appendix I				Router: Refer to Appendix II				

For audio broadcasting we adopt a one-way end-to-end delay for high-quality real time traffic from ITU G.114 specifications of less than 150 milliseconds. For jitter, no more than 400 ms is recommended for telephone-quality speech and no more than 100 ms is recommended for CD-quality compressed sound (Fluckiger, 1995). We suggest a jitter limit of no more than 100 ms for broadcast quality sound.

A typical data stream for audio broadcasting is 56–64 kbps (Kenyon and Nightingale, 1992). The corresponding bandwidth requirement is around 60–80 kbps. The residual bit error rate is recommended to be lower than 10^{-2} for a telephone-quality audio stream and lower than 10^{-4} for a CD-quality audio stream using a compressed format (Fluckiger, 1995). Since there is a higher quality requirement in broadcasting than in telephone, and a lower quality requirement than in a CD audio stream, we recommend a loss rate and error rate of no more than 10^{-3} for broadcast audio streams.

QoS Metrics for Video Broadcasting

In Video Broadcasting, under different codec technologies and quality requirements, there are different requirements for network transmissions. We list the QoS metrics for video broadcasting using some typical codec technologies in Table 10.

Table 10
QoS Metrics for Video Broadcasting

Traffic Class	Technology Attributes	QoS Metrics						
		Timeliness			Preciseness			Accuracy
		Response time Expected by Users	Delay (ms)	Jitter (ms)	n/a	Required Bandwidth (bps)	Loss Rate	Error Rate
Video Broadcasting	Real Time and Highly Asymmetric	2–5 Seconds Lip-synch: <100ms	<150	<100				
Typical Application		Coding Standard						
VCR Quality		MPEG-1		<100		1.2–1.5M	<0.001%	<0.001%
Video Quality slightly superior to that of broadcast TV (NTSC or PAL) with bit rate of 4M		MPEG-2				4–60 M		
HDTV requiring bit rate from 15–34 M				<50			<0.0001%	<0.0001%
Multimedia on Web		MPEG-4		<150		28.8–500K	<0.001%	<0.001%
Network Capacity								
Link: Refer to Appendix I					Router: Refer to Appendix II			

For VCR quality MPEG-1 video streams, the required bandwidth is around 1.2–1.5 Mbps (Silveira et al, 1996). Jitter is recommended to be less than 100 ms for a broadcast-quality video stream (Fluckiger, 1995). The residual bit error rate is recommended to be lower than 10^{-5} for a broadcast TV quality video stream using a compressed format (Fluckiger, 1995). The loss and error rates should be less than 10^{-5}.

For HDTV quality MPEG-2 video streams, the required bandwidth is around 40 Mbps (Silveira et al, 1996). Jitter is recommended to be less than 50 ms for HDTV quality video streams (Fluckiger, 1995). The residual bit error rate is recommended to be lower than 10^{-6} for HDTV quality video streams using a compressed format (Fluckiger, 1995). The loss rate and error rate should be less than 10^{-6}.

For typical MPEG-4 video streams on the World Wide Web, the required bandwidth is around 28.8–500 kbps (Silveira et al, 1996). Jitter less than 150 ms is recommended due to a lower quality requirement than VCR technologies, which require a jitter less than 100ms. On the other hand, MPEG-4 also has a higher compression rate, which usually requires less residual error. Based on considerations of quality and compression rate, we recommend a loss rate and error rate less than 10^{-5}.

QoS Metrics for Interactive Audio on Demand

Interactive audio-on-demand on the World Wide Web is the ability to listen to sound on a personal computer (with a sound card and speakers installed) while the file is passed through an individual modem. Under different codec technologies and quality requirements, there are different requirements for network transmission. We list the QoS metrics for some typical audio-on-demand applications in Table 11.

Table 11
QoS Metrics for Interactive Audio on Demand

Traffic Class	Technology Attributes	QoS Metrics						
		Timeliness			Preciseness			Accuracy
		Response time Expected by Users	Delay (ms)	Jitter (ms)	n/a	Required Bandwidth (bps)	Loss Rate	Error Rate
Interactive Audio on Demand	Real Time and Highly Asymmetric	2–5 Seconds	<150	<100				
Typical Application		Coding Standard						
MP3, 64K for mono, 128K for stereo		MPEG-1				32–448K	<0.1%	<0.1%
Advanced Audio Coding (AAC) providing CD quality audio		MPEG-2				384K	<0.01%	<0.01%
Network Capacity								
Link: Refer to Appendix I				Router: Refer to Appendix II				

For a mono-quality MP3 audio stream, the required bandwidth is 64 kbps. For a stereo-quality MP3 audio stream, the required bandwidth is 128 kbps. Delay less than 150ms is recommended for high-quality real time traffic in the ITU G.114 standard (ITU, 1996). A jitter rate of less than 100 ms for CD-quality compressed sound is recommended (Fluckiger, 1995). The residual bit error rate is recommended to be lower than 10^{-3} for a broadcasting-quality audio stream in a compressed format. Therefore, the loss rate and error rate should be less than 10^{-3}.

For a CD-quality MP2 audio stream, the required bandwidth is 384 kbps. A delay less than 150ms is recommended for high-quality real time traffic in the ITU G.114 standard (ITU, 1992). A jitter rate of less than 100 ms for CD-quality compressed sound is recommended (Fluckiger, 1995). The residual bit error rate is recommended to be lower than 10^{-4} for a CD-quality audio stream in a compressed format. Therefore, the loss rate and error rate should be less than 10^{-4}.

QoS Metrics for Interactive Video on Demand

Interactive video-on-demand refers to a range of applications whereby users can request access to video servers, which provide still and moving pictures on an individual basis (Fluckiger, 1995). Elapsed time in video-on-demand refers to the time elapsed from a request generation to the start of video playback. The QoS metrics for video-on-demand are given in Table 12.

Elapsed time of a few minutes should be acceptable for video-on-demand because users are not particularly concerned with the starting delay, but instead are concerned with the delay during playback. Response to interactive functions, such as play and next, should be less than 2–5 seconds. The typical bandwidth requirement for the user's local network is 1.5 Mbps. The bandwidth requirement for the backbone depends on how many users the video server needs to support simultaneously. In general, backbone bandwidth is hundreds of megabits per second.

Other QoS metrics for video-on-demand are the same as those for video broadcasting.

Table 12
QoS Metrics for Interactive Video on Demand

<table>
<tr><th rowspan="3">Traffic Class</th><th rowspan="3">Technology Attributes</th><th colspan="7">QoS Metrics</th></tr>
<tr><th colspan="3">Timeliness</th><th colspan="3">Preciseness</th><th>Accuracy</th></tr>
<tr><th>Response time Expected by Users</th><th>Delay (ms)</th><th>Jitter (ms)</th><th>n/a</th><th>Required Bandwidth (bps)</th><th>Loss Rate</th><th>Error Rate</th></tr>
<tr><td>Interactive Video on Demand</td><td>Real Time and Highly Asymmetric</td><td>Elapsed Time: a few minutes

Response to Interactive function: 2–5 Seconds

Lip-synch: <100ms</td><td><150</td><td></td><td></td><td>Local:1.5M (typical application)

Backbone: hundreds of mega</td><td></td><td></td></tr>
<tr><td colspan="2">Typical Application</td><td colspan="2">Coding Standard</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="2">VCR Quality</td><td colspan="2">MPEG-1</td><td><100</td><td></td><td>1.2–1.5M / A Single Video</td><td><0.001%</td><td><0.001%</td></tr>
<tr><td colspan="2">Video Quality slightly superior to that of broadcast TV (NTSC or PAL) with bit rate of 4M</td><td colspan="2" rowspan="2">MPEG-2</td><td></td><td></td><td>4–60 M / A Single Video</td><td></td><td></td></tr>
<tr><td colspan="2">HDTV requiring bit rate from 15–34 M</td><td><50</td><td></td><td></td><td><0.0001 %</td><td><0.0001 %</td></tr>
<tr><td colspan="2">Multimedia on Web</td><td colspan="2">MPEG-4</td><td><150</td><td></td><td>28.8–500K / A Single Video</td><td><0.001%</td><td><0.001%</td></tr>
<tr><td colspan="9">Network Capacity</td></tr>
<tr><td colspan="4">Link: Refer to Appendix I</td><td colspan="5">Router: Refer to Appendix II</td></tr>
</table>

QoS Metrics for Telemetry

Telemetry refers to the telemetering and telecontrol of industrial processes. Telemetry is an example of a data service which requires real time streaming performance. Clearly, two-way control implies a tight limit on the allowable delay in telemetry. The QoS metrics for telemetry are given in Table 13.

For telemetry, a delay value of 250 milliseconds is proposed (QoS Performance Requirement for UMTS (Nortel, 2001)). The range of bandwidth requirements is between 2 kbps and 52 Mbps, depending on the specific telemetry system. A key differentiator from voice and video services in this category is the zero tolerance for information error and loss (QoS Performance Requirement for UMTS (Nortel, 2001)). It is clearly understandable that there cannot be data error and loss when controlling an important industrial process.

Table 13
QoS Metrics for Telemetry

Traffic Class	Technology Attributes	QoS Metrics						
		Timeliness			Preciseness			Accuracy
		Response time Expected by Users	Delay (ms)	Jitter (ms)	n/a	Required Bandwidth (bps)	Loss Rate	Error Rate
Telemetry	Real Time and Highly Asymmetric		<250	<100		2K–52M	Zero	Zero
Network Capacity								
Link: Refer to Appendix I				Router: Refer to Appendix II				

QoS Metrics for Teleconferencing

Teleconferencing can occur in three forms: audio conferencing, audiographics conferencing, and video conferencing. Each of these forms requires its own QoS metrics. The metrics for the three forms of teleconferencing are outlined in this section.

QoS Metrics for Audio Conferencing

Audio Conferencing provides an audio link similar to that of a conventional telephone, except that it offers much higher quality audio and enables more than two sites to be linked together. Using hands-free audio units with sensitive microphones and sophisticated echo-cancellation software, audio-conferencing enables communication between groups of participants rather than just individuals (Bleazard, 1985). Voice data transmission dictates the QoS metrics of audio conferencing shown in Table 14.

Table 14
QoS Metrics for Audio Conferencing

Traffic Class	Technology Attributes	QoS Metrics						
		Timeliness			Preciseness			Accuracy
		Response time Expected by Users	Delay (ms)	Jitter (ms)	Data Rate (bps)	Required Bandwidth (bps)	Loss Rate	Error Rate
Audio Conferencing	Real Time and Symmetric		<150	<400			<1%	<1%
Coding Standard								
G.711					64 K	80 K		
G.726					40~16 K	50–22 K		
G.728					16 K	22 K		
G.729					8 K	11 K		
G.723.1					6.3/5.3 K	9 / 8 K		
GSM FR					13 K	18 K		
GSM EFR					12.2 K	17 K		
Network Capacity								
Link: Refer to Appendix I				Router: Refer to Appendix II				

ITU G.114 specification recommends less than a 150 millisecond one-way end-to-end delay for high-quality real time traffic, such as voice transmissions. Jitter of no more than 400 milliseconds is recommended for telephone-quality speech (Fluckiger, 1995). In the case of real time audio and video, retransmission techniques are not appropriate. G.729 codec requires packet loss far less than 1 percent to avoid audible errors. In the case of presentation to human users without recording for further processing, the residual bit rate error rate of a telephone-quality audio stream should be lower than 10^{-2} (Fluckiger, 1995). Differing Codec technologies lead to different bandwidth requirements.

QoS Metrics for Audiographics Conferencing

Except for the audio link, audiographics conferencing enables participants at two or more sites to have a shared workspace on their computer desktops. This might be a shared “whiteboard” where they can draw, write, or import and manipulate images collaboratively in real time. Audiographics conferencing might also be “application sharing”, where a piece of software can be run and controlled by both users (G B Bleazard, 1985). This type of conferencing is useful when users at different sites want to work together on documents such as reports or statistical data in spreadsheets. The largest QoS concern for audiographics is voice quality. The QoS metrics for audiographics conferencing are given in Table 15.

Table 15
QoS Metrics for Audiographics Conferencing

Traffic Class	Technology Attributes	QoS Metrics						
		Timeliness			Preciseness			Accuracy
		n/a	Delay (ms)	Jitter (ms)	n/a	Required Bandwidth (bps)	Loss Rate	Error Rate
Audiographics Conferencing	Real Time and Symmetric		<150	<400		9.6–19.6K	<1%	<1%
Network Capacity								
Link: Refer to Appendix I				Router: Refer to Appendix II				

ITU G.114 specification recommends less than a 150 millisecond one-way end-to-end delay for high-quality real time traffic, such as voice. For jitter, a rate of no more than 400 milliseconds is recommended for telephone-quality speech (Fluckiger, 1995). The G.729 codec requires a packet loss rate of far less than 1 percent to avoid audible errors. In the case of presentation to human users without recording for further processing, the residual bit error rate of a telephone-quality audio stream should be lower than 10^{-2} (Fluckiger, 1995). Typical bandwidth requirements for audiographics conferencing are 9.6–19.6 kbps (Bleazard, 1985).

QoS Metrics for Video Conferencing

Video conferencing enables real time communication over a distance by allowing people at two or more sites to communicate with each other. In video conferencing, in addition to hearing each others voices, as with a conventional telephone, people can see video pictures of each other from their various sites (Janssen et al, 2001). Each site has one or more cameras, microphones, loudspeakers and monitors, as well as a codec. This type of conferencing aims to create a sense of people from distant sites being in the same room, an effect that has been called Virtual Presence. Video conferencing is not only concerned with voice quality, but is also concerned with video quality. The QoS metrics for video conferencing are given in Table 16.

Table 16
QoS Metrics for Video Conferencing

Traffic Class	Technology Attributes	QoS Metrics						
		Timeliness			Preciseness			Accuracy
		Response time Expected by Users	Delay (ms)	Jitter (ms)	Data Rate (bps)	Required Bandwidth (bps)	Loss Rate	Error Rate
Video Conferencing	Real Time and Symmetric	Lip-synch: <100 ms	<150	<400			<0.01%	< 0.01%
Coding Standard								
H.320					64–1920K	80K–2M		
H.323					64X K	80X K		
H.324					<64K	<80K		
Network Capacity								
Link: Refer to Appendix I					Router: Refer to Appendix II			

Since humans are much more sensitive to alteration of audio signals than visual signals, we adopt the recommendations for delay and jitter from ITU G.114: less than a 150 millisecond one-way end-to-end delay for high-quality real time traffic and no more than a 400 millisecond jitter for telephone-quality speech (Fluckiger, 1995). However, the requirement for loss rate and error rate differ. Recall that compressed streams are more sensitive to errors than uncompressed streams since one lost packet may influence a number of subsequent packets. Similarly, streams with a higher compression ratio are more sensitive to errors than streams with a lower compression ratio since more packets will be impacted. Therefore, the end-to-end network bit error rate for compressed streams, before possible error recovery, between end-systems is recommended not to exceed 10^{-4} for videoconference quality (Fluckiger, 1995).

The H.320-series governs the basic video-telephony concepts of audio, video and graphical communications by specifying requirements for processing audio and video information, providing common formats for compatible audio/video inputs and outputs, and defining protocols that allow a multimedia terminal to utilize the communication links and synchronization of audio and video signals. H.320 standards address ISDN Videoconferencing. The H.323 standards address Video communications on Local Area Networks, and the H.324 standards address video and audio communications over low bit rate connections such as POTS modem connections. From Table 16, we can see that different technology standards lead to different bandwidth requirements.

It is worthy to note that the bandwidth requirements for video conferencing and videophony (described in the next section) may be less stringent than that of general video. This is based on the assumption that these media will likely have less motion and little or no scene changes in the picture.

QoS Metrics for Videophony

Videophony is a companion audiovisual service of videoconferencing. It may be distinguished by the following features (Kenyon and Nightingale, 1992): it is primarily for person-to-person rather than group-to-group audiovisual communication, and it is usually an on-demand service provided on customer-switched networks, whereas videoconferences can be scheduled in advance. Compared with video conferencing, videophony is a more interactive application, which requires instantaneous responses from dialogists. The QoS metrics for videophony are given in Table 17.

Videophony requires a shorter end-to-end delay of 100 ms to make the difference between audio and real speech imperceptible to users (Fluckiger, 1995). Beyond 100 ms, although speech is still comprehensible, users may become irritated with the service. Since videophony is a special application of video conferencing, it complies with the same technology standards and has the same QoS metrics as video conferencing.

Table 17
QoS Metrics for Videophony

Traffic Class	Technology Attributes	QoS Metrics						
		Timeliness			Preciseness			Accuracy
		Response time Expected by Users	Delay (ms)	Jitter (ms)	Data Rate (bps)	Required Bandwidth (bps)	Loss Rate	Error Rate
Videophony	Real Time and Symmetric	Phone to Phone Delay: <150 ms Lip-synch: <100 ms	<100	<400			<0.01%	< 0.01%
Coding Standard								
H.320					64–1920K	80K–2M		
H.323					64X K	80X K		
H.324					<64K	<80K		
Network Capacity								
Link: Refer to Appendix I				Router: Refer to Appendix II				

Table 18
QoS Metrics for Voiceover IP

Traffic Class	Technology Attributes	QoS Metrics						
		Timeliness			Preciseness			Accuracy
		Response time Expected by Users	Delay (ms)	Jitter (ms)	Data Rate (bps)	Required Bandwidth (bps)	Loss Rate	Error Rate
Voiceover IP	Real Time and Symmetric	Phone to Phone Delay: <150 ms	<100	<400			< 1%	< 1%
Coding Standard								
G.711					64 K	80 K		
G.726					40~16 K	50–22 K		
G.728					16 K	22 K		
G.729					8 K	11 K		
G.723.1					6.3/5.3 K	9 / 8 K		
GSM FR					13 K	18 K		
GSM EFR					12.2 K	17 K		
Network Capacity								
Link: Refer to Appendix I				Router: Refer to Appendix II				

QoS Metrics for Voiceover IP

The QoS metrics for Voiceover IP are given in Table 18.

Like videophony, VoIP also requires a shorter end-to-end delay of 100 ms to give the users an imperceptible difference between audio and real speech (Szuprowicz, 1995). A jitter rate of less than 400 milliseconds is suggested for telephone-quality speech (Fluckiger, 1995). We adopt the loss rate requirement, packet loss far less than 1 percent, in the G.729 codec. In the case of presentation to human users without recording for further processing, the residual bit rate error rate of a telephone-quality audio stream should be lower than 10^{-2} (Fluckiger, 1995). Again, different voice coding technologies lead to different bandwidth requirements.

SUMMARY

In order for users to find a certain type of multimedia service acceptable, the application must meet specific QoS requirements. Metrics used to determine QoS include response time, delay, jitter, data rate, required bandwidth, loss rate and error rate. In order to provide QoS on a network, requirements for QoS must be established. Defining the limits of QoS metrics helps to establish necessary requirements.

In this paper, we consider two key factors of QoS requirements: human factor and technology attributes. Based on our survey of human factors and technology features, we propose numeric QoS metrics of network applications in this paper. Up to date, we have not found any literature stating metrics which differed from our findings. Based on the lack of literature which would dispute these findings, we could not definitively state whether or not specific measures are “widely accepted”. However, we believe that metrics proposed in this paper provide a starting point to establish a universally accepted set of QoS requirements for network applications on the Internet. We concede that these proposed measures could be refined through experimental verification. Further work can make use of these metrics to differentiate network traffic for realizing quality assurance in networks.

ACKNOWLEDGEMENTS

This work is sponsored by the Air Force Research Laboratory – Rome (AFRL-Rome) under grant number F30602-01-1-0510, and the Department of Defense (DoD) and the Air Force Office of Scientific Research (AFOSR) under grant number F49620-01-1-0317. The U.S. government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of, AFRL-Rome, DoD, AFOSR, or the U.S. Government.

REFERENCES

Austerberry, D. (2002). *The Technology of Video & Audio Streaming*. Focal Press.

Bleazard, G.B. (1985). *Introducing Teleconferencing.* NCC Publication. pp. 58.

Eslinger, B. & Baker, M. (2002). Advances in Telemetry and Network Transmission, International Test and Evaluation Association (ITEA). Instrumentation Conference.

ETSI TIPHON (2000). End-to-end Quality of Service in TIPHON Systems. Part 2: Definition of Quality of Service (QoS) Classes. *ETS.* 101 329–2.

Fiedler, U. & Plattner, B. *Quality of Service in Business-to-Business E-Commerce Application*. Computer Engineering and Networks Laboratory, Switzerland

Fluckiger, F. (1995). *Understanding Networked multimedia.* Prentice Hall, pP. 242–382.

Janssen, J.,Vleeschauwer, D.D., Buchli, M.J.C. & Kooij, R.E. (2001). *Tuning an IP-based Network Transporting Telephony and Videophony.* Alcatel Bell, Network Strategy Group, Francis Wellesplein 1, B-2018 Antwerp, Belgium.

Kenyon, N.D. & Nightingale, C. (1992). *Audiovisual Telecommunications.* BT Laboratories, Ipswich, UK, Chapman & Hall.

Ko, M. & Koo, I. (1996). *An Overview of Interactive Video On Demand System.* The University of British Columbia.

Menasce, D.A. & Almeida, V.A.F. (1998). *Capacity Planning for Web Performance.* Prentice Hall, PTR.

Nortel Networks (2001). *QoS Performance Requirement for UMTS.* Retrieved November 8, 2001: http://www.3gpp.org/ftp/tsg_sa/WG1_Serv/TSGS1_03HCourt/Docs/Docs-Long_Names/ S199362_QoS_Nortel_Performance_requirements_for_UMTS.doc

Silveira, R.M., Margi, C.B., Gonzalez, L.G., Favero, E., Vilcachagua, O.D., Bressan, G. & Ruggiero, W.V.A. (1996). *Multimedia on Demand System for Distance Education.* LARC-Laboratory of Computer Architecture and Networks, PCS-Department of Computer and Digital System Engineering EPUSP-Polytechnic School, University of San Paulo, Standards: ITU-T G.114.

Szuprowicz, B.O. (1995). *Multimedia Networking.* McGraw-Hill, Inc., 161–162.

To, T.P.J. & Hamidzadeh, B. (1998). *Interactive Video-on-demand Systems.* Kluwer Academic Publishers, pp. 9–15.

Marziale, V. & Vitaletti, A. (2001). *A Framework for Internet QoS Requirements.* IST Mobile Communication Summit 2001, Barcelona, Spain.

Ye, N. (2002). QoS-centric stateful resource management in information systems. *Information Systems Frontiers.* 4(2) 149–160.

Yan Chen is a doctoral student in the Department of Industrial Engineering at Arizona State University. She received a B.E. degree and a M.E. degree from Xiamen University, Xiamen, Fujian, P.R.China, both in Systems Engineering, where her interests were primarily in the application of Artificial Neural Network. Her current interests are in Modeling and Analysis of Semiconductor Manufacturing and Scheduling of complex job shops.

Toni Farley is the Assistant Director of the Information and Systems Assurance Laboratory, and a PhD student of Computer Science at Arizona State University (ASU), Tempe, Arizona. She has been awarded a graduate fellowship from the AT&T Labs Fellowship Program (ALFP). She received her BS degree in Computer Science from ASU in 2003. Her research interests include computer networks, network security, and information assurance. She is a member of IEEE and IEEE Computer Society.

Nong Ye is a Professor of Industrial Engineering and an Affiliated Professor of Computer Science and Engineering at Arizona State University. She holds a Ph.D. degree in Industrial Engineering from Purdue University, a M.S. degree in Computer Science from the Chinese Academy of Sciences, and a B.S. degree in Computer Science from Peking University. Her research interests are in assuring process quality and preventing faults and errors in information systems. Dr. Ye is an Associate Editor for *IEEE Transactions on Reliability* and *IEEE Transactions on Systems, Man, and Cybernetics*. She is a senior member of the Institute of Industrial Engineers and a senior member of IEEE.

Appendix I: Link Capacity of Networks

Wide Area Networks (WANs)

Technology used to build WANs	Speed
X.25 ISDN Frame Relay SMDS (Switched Multi-megabit Data Service) ATM	45 M bps–several G bps

Local Network (LANs)

Technology used to build LANs	Speed	Topology	Medium	Access Mechanism
Ethernet	10 M bps 100 M bps	Bus	Twisted Pair	CSMA/CD (Carrier Sense with Multiple Access / Collision Detection)
Token Ring	4 M bps 16 M bps	Ring	Shielded Twisted Cable	Token-passing
FDDI (Fiber Distributed Data Interconnect)	100 M bps	Ring (Two rings)	Optical Fiber	Token-passing

LAN to WAN

Connection	Speed
Dedicated leased line at T1	1.544 M bps
Dedicated leased line at T3	45 M bps

Home to WAN

Connection	Speed
Dial-up Analog Modem	14.4 ~ 56 K bps
ISDN Basic Rate Interface (BRI)	128 K bps
ISDN Primary Rate Interface (PRI)	1.544 M bps
High-Bit-Rate Digital Subscriber Line (HDSL)	1.544 M bps
Asymmetric Digital Subscriber Line (ADSL)	Outgoing: 640 K bps; incoming: 6 M bps
Cable TV Companies	10 M bps

Appendix II: — Router Capacity of Network

Router	Upper Bandwidth
Cisco 12000 Series Internet Router	10 G bps
Cisco 10700 Series Internet Routers	5 G bps
Cisco 10000 Series Internet router	51.2 G bps
Cisco 7600 Series Internet Router	256 G bps
Cisco 7500 Series Internet Router	2 1.067 G bps
Cisco 7300 Series Internet Router	16 G bps
Cisco 7200 Series Internet Router	1.2 G bps
Cisco 7100 Series Internet Router	140 M bps