Quality of Service issues in IP/NGN networks

1. Introduction and QoS definitions
2. Main parameters defining QoS in IP networks
3. Estimation of call quality for VoIP
4. IPTV QoS issues
5. QoS guarantees: possible approaches to the problem
6. QoS – concluding remarks

1. Introduction and QoS definitions

a. QoS general definitions

- The collective effect of service performances which determine the degree of satisfaction of a user of the service.

- QoS is defined as the measure of performance for a transmission system that reflects its transmission quality and service availability.

- Until recently were not agreed quantifiable measures that define unambiguously QoS, as perceived by a user. Terms, such as “better”, “worse”, “high”, “medium”, “low”, “good”, “fair”, “poor”, are typically used, but these are subjective and cannot therefore be translated precisely into network level parameters that can subsequently be designed for by network planners.

- On the Internet and in other networks, QoS is the idea that the throughput, losses, delays and other network characteristics can be measured, improved, and guaranteed in advance.

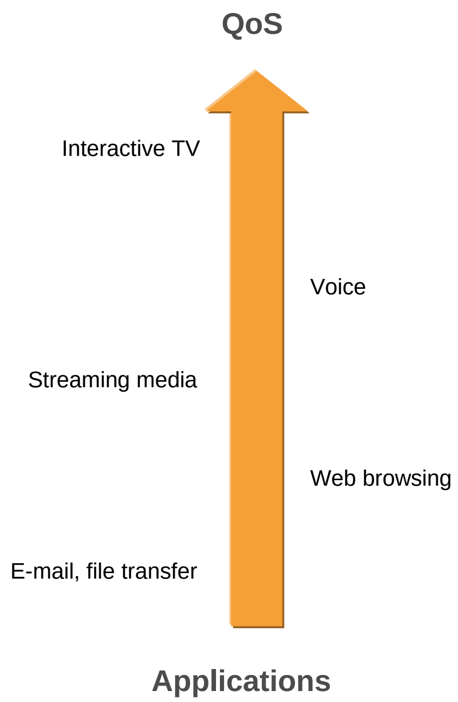
- The end effect at the terminal is also heavily dependent upon issues such as compression algorithms, coding schemes, the presence of protocols for security, data recovery, re-transmission, etc., and the ability of applications to adapt to network congestion.

- However, network providers need performance metrics that they can agree with service providers buying resources from them with certain performance guarantees.

- Nevertheless, there are system performance metrics that are considered as the most important in terms of their impact on the end-to-end QoS, as perceived by a user:

b. QoS parameters - system performance metrics

- Network/Devices Availability
- Network Throughput
- Packet Delay
- Packet Delay Variation (Jitter)
- Packet Loss



2. Main parameters defining QoS in IP networks

- The number one issue operators have is:
  guarantee of Quality of Service

  How to support voice traffic on backbone ?
  Actually, this is the number two issue

- The number one issue is:

  Network availability

- QoS makes a sense only if the network is up and running all the time, hence it’s reliable

#Availability

- Before any QoS can be implemented successfully, the network infrastructure must be designed to be highly available. Service availability is a crucial foundation element of QoS.

- **Availability -** the fraction of time that network connectivity is available between an ingress point and a specified egress point is defined as network availability.

Availability (cnt.)

Availability	Cumulative Downtime per Year
99.000%	3 days, 15 hours, 36 minutes
99.500%	1 day, 19 hours, 48 minutes
99.900%	8 hours, 46 minutes
99.950%	4 hours, 23 minutes
99.990%	53 minutes
99.999%	5 minutes
99.9999%	30 seconds

- Availability in PSTN networks is already for 10s of years equal to the famous 99.999%, also called the 5 nines

- Traditional IP data equipment does not offer 5 nines reliability

#Throughput

- The available user bandwidth between an ingress point of presence (POP) and an egress POP.
- This is the effective data transfer rate measured in bps. It is not the same as the maximum capacity of the network, often erroneously called the network's bandwidth.
- A minimum rate of throughput is usually guaranteed by a service provider (who needs to have a similar guarantee from the network provider).

Other main parameters

- Packet delay
- Packet delay variation (jitter)
- Packet loss

Note: All definition are based on ITU Recommendations Y.1540 and Y.1541

a. Packet delay

- The finite amount of time it takes a packet to reach the receiving endpoint after being transmitted from the sending endpoint. In the case of voice, this delay is defined as the amount of time it takes for sound to leave the speaker's mouth and be heard in the listener's ear.

Y.1540/1541 formal definition of PD

- *IP packet transfer delay* (IPTD) is the time $t_2 - t_1$ between the occurrence of two corresponding IP packet transfer reference events: an ingress event RE_1 at time t_1 and an egress event RE_2 at time t_2, where $t_2 > t_1$ and $(t_2 - t_1) \leq T_{max}$. IPTD is defined for all successful and errored packet transfer outcomes. If the packet is fragmented, t_2 is the time of the final corresponding egress event. *Mean IP packet transfer delay,* the parameter actually specified in Recommendation Y.1541, is the arithmetic average of IP packet transfer delays for a population of interest.

Packet delay (cnt.)

- The average time varies according to the amount of traffic being transmitted and the bandwidth available at that given moment. If traffic is greater than bandwidth available, packet delivery will be delayed.

- Voice and video are a delay-sensitive applications while most data applications are not. When voice packets are lost or arrive late they are discarded; the results are reduced voice quality.

- T_{max} – upper limit of permissible delay; depends from type of traffic and type of network (150 ms – 400 ms)

Components of packet delays

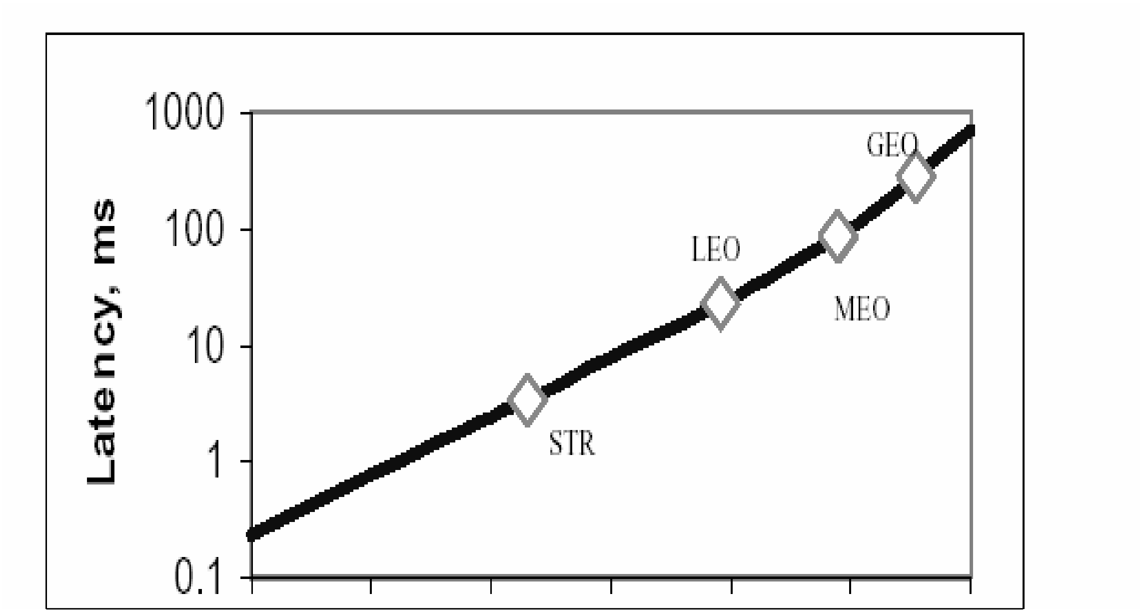
- **Propagation delay**: the time to travel across the network from end to end. It’s based on the speed of light and the distance the signal must travel. Unless satellites are involved, the latency of a 5000 km voice call carried by a circuit-switched telephone network is about 25 ms. The important factor regarding delay is the propagation time along the cable (approx. 15 ms to cross the US and 50 ms to cross Russia).

- **Transport delay**: the time to get through the network devices along the path. Networks with a number of firewalls, routers, queuing or slow WANs introduce more delay than an overprovisioned LAN on one floor of a building.

- **Packetization delay**: the time for the codec to digitize the analog signal and build frames – and undo it at the other end. The G.729 codec has a higher packetization delay (25 ms) than the G.711 codec (1 ms) because it takes longer to compress and decompress the signal.

- **Jitter buffer delay:** is introduced for a compensation of a jitter; details see below

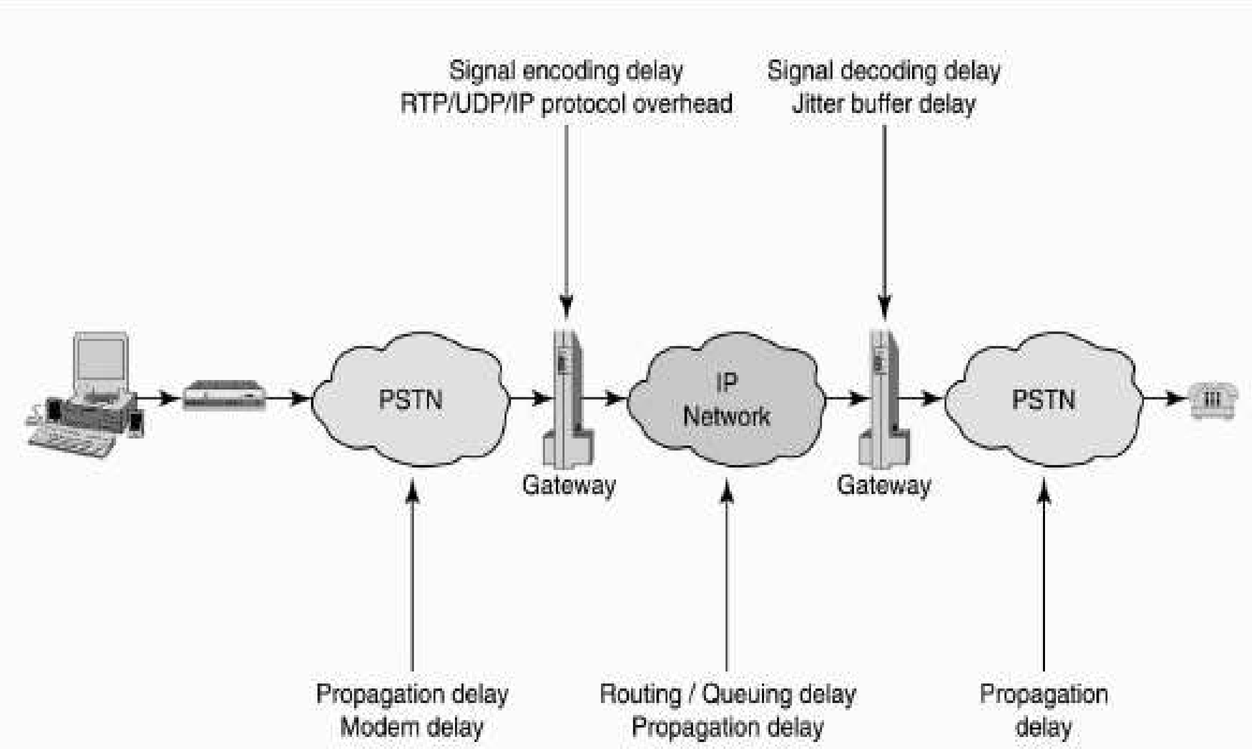
Delays for different satellite communications systems



Distance 10 100 1000 10.000 100.000 km

STR – Stratosphere balloon
LEO – Low-orbit satellite
MEO – Middle-orbit satellite
GEO – Geostationary-orbit satellite

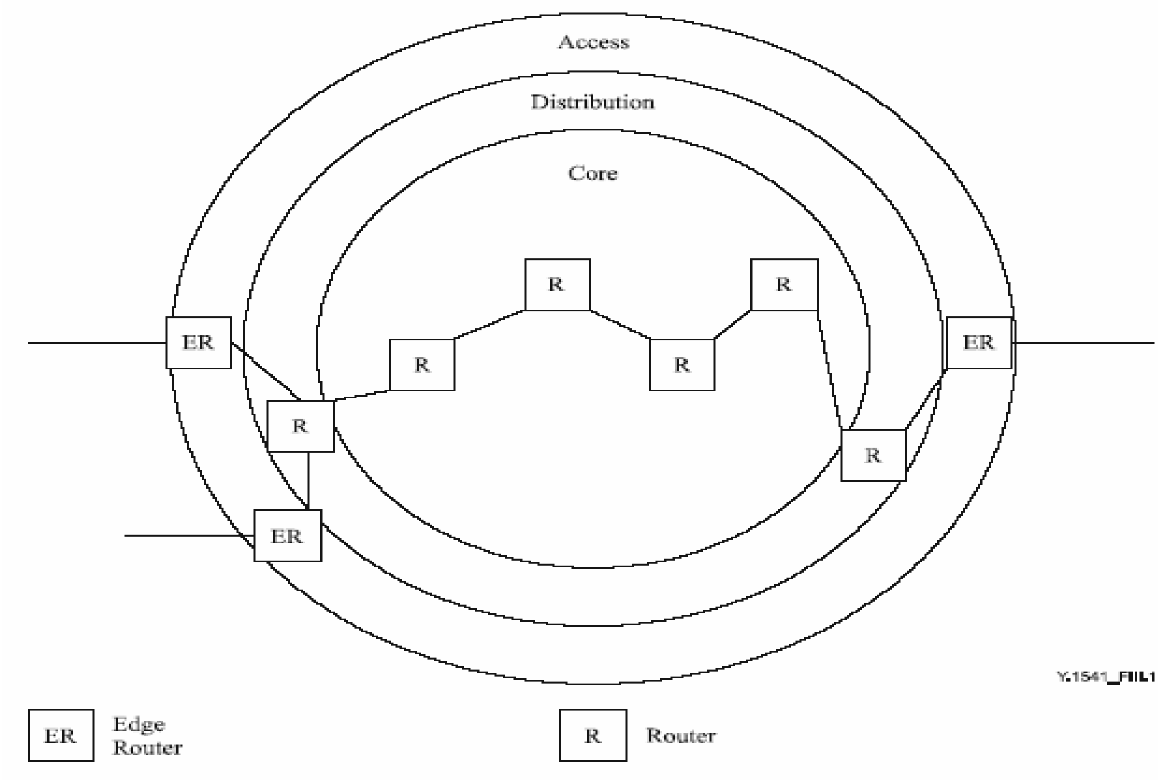
b. Estimation of packet delays for the VoIP



Examples of transmission delays

Transmission or processing system	Contribution to one-way transmission time	Remarks
Terrestrial coaxial cable or radio-relay system: FDM and digital transmission	4 µs/km	Allows for delay in repeaters and regenerators
Optical fibre cable system, digital transmission	5 µs/km	
Submarine coaxial cable system	6 µs/km	
Submarine optical fibre system: – transmit terminal – receive terminal	 13 ms 10 ms	Worst case
Satellite system: – 400 km altitude – 14 000 km altitude – 36 000 km altitude	 12 ms 110 ms 260 ms	Propagation through space only (between earth stations)

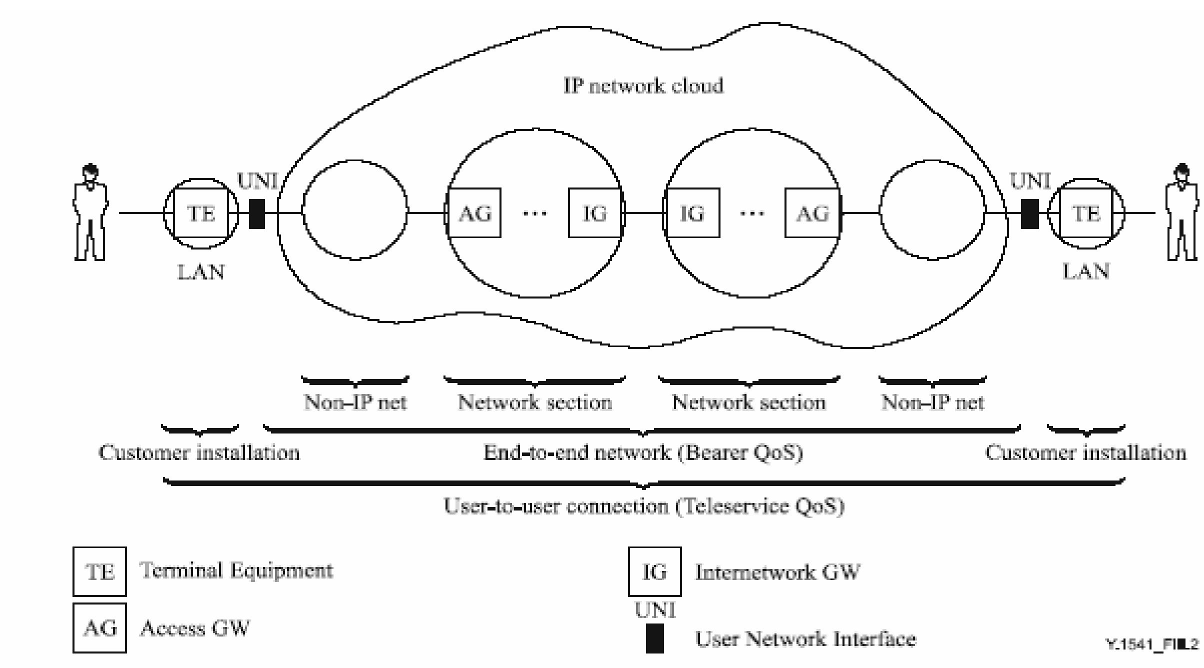
IP nodes in a network section (Y.1541)



Examples of typical delay contribution by router (Y.1541)

Role	Average total delay (sum of queuing and processing)	Delay variation
Access gateway	10 ms	16 ms
Internetworking gateway	3 ms	3 ms
Distribution	3 ms	3 ms
Core	2 ms	3 ms

Hypothetical reference path for QoS (Y.1541)



Estimation of IPTD

IPTD (in ms) **= PrD + TD + PD + JBD**
PrD + TD = $(R_{km} \times 5x10\text{-}3) + (N_A \times D_A) + (N_D \times D_D) + (N_C \times D_C) + (N_I \times D_I)$
where:

- R_{km} represents the route length assumption computed above.
- N_A, N_D, N_C, and N_I represent the number of IP access gateway, distribution, core and internetwork gateway nodes respectively; consistent with the network section example in Figure
- D_A, D_D, D_C, and D_I represent the delay of IP access gateway, distribution, core and internetwork gateway nodes respectively; consistent with the values for Class X (e.g., Table for typical delays by router).

PD
JBD

c. Packet delay variation (jitter)

- Packet delay variation (IPDV) - the variability in packet arrival times at the destination
- The difference in the end-to-end delay between packets. For example, if one packet required 100 milliseconds (ms) to traverse the network from the source-endpoint to the destination-endpoint and the following packet required 125 ms to make the same trip, then the delay variation would be calculated as 25 ms.

- In general - voice packets must compete with non real-time data traffic

  # bursts structure of data traffic inside of the network

  # congestion problem

Results are in varied arrival times for voice packets.

- When consecutive voice packets arrive at irregular intervals, the result is a distortion in the sound, which, if severe, can make the speaker unintelligible.

- Jitter has many causes, including:

  # variations in queue length

  # variations in the processing time needed to reorder packets that arrived out of order because they traveled over different paths

  # variations in the processing time needed to reassemble packets that were segmented by the source before being transmitted.

Y.1540/1541 formal definition of PDV

- *IP packet delay variation* (IPDV) is defined based on observations of corresponding IP packet arrivals at ingress and egress MPs (e.g., MP1, MP2). The packet delay variation, v_k , for an IP packet *k* between MP_1 and MP_2 is the difference between the absolute IP packet transfer delay, x_k , of the packet and a defined reference IP packet transfer delay, $d_{1,2}$ between those same MPs: $v_k = x_k - d_{1,2}$.

  The reference IP packet transfer delay, $d_{1,2}$, between SRC and DST is the absolute IP packet transfer delay experienced by the first IP packet between those two MPs.

d. Packet loss

• Network devices, such as switches and routers, sometimes have to hold data packets in buffered queues when a link gets congested.

•If the link remains congested for too long, the buffered queues will overflow and data will be lost.

•The lost data packets must be retransmitted, adding, of course, to the total transmission time. In a well-managed network, packet loss will typically be less than 1% averaged over, say, a month.

• When data packets are lost, a receiving computer can simply request a retransmission. When voice packets are lost or arrive too late they are discarded of retransmitted. The result is in the form of gaps in the conversation (like a poor cell phone connection).

Packet loss (cnt.)

- **Loss**—A comparative measure of packets faithfully transmitted and received to the total number of packets that were transmitted. Loss is expressed as the percentage of packets that were dropped. Loss is typically a function of availability. If the network is highly available, then loss (during periods of non-congestion) would essentially be zero. During periods of congestion, however, QoS mechanisms would determine which packets would be suitable to drop.
- Y.1540/1541 formal definition of PL

  IP packet loss ratio (IPLR) is the ratio of total lost IP packet outcomes to the total transmitted IP packets in a population of interest.

e. IP QoS class definitions and network performance objectives (Y.1541)

Network performance parameter	Nature of network performance objective	QoS classes					
		Class 0	Class 1	Class 2	Class 3	Class 4	Class 5 (unspecified)
IPTD	Upper bound on the mean IPTD	100 ms	400 ms	100 ms	400 ms	1 s	U
IPDV	Upper bound on the min IPDV	50 ms	50 ms	U	U	U	U
IPLR	Upper bound on the packet loss probability	$1*10^{-3}$	$1*10^{-3}$	$1*10^{-3}$	$1*10^{-3}$	$1*10^{-3}$	U
IPER	Upper bound	$1*10^{-4}$	$1*10^{-4}$	$1*10^{-4}$	$1*10^{-4}$	$1*10^{-4}$	U

Note: U = unspecified.

f. Applications according with IP QoS classes (Y.1541)

QoS class	Applications (examples)
0	Real-time, jitter sensitive, high interaction (VoIP, VTC)
1	Real-time, jitter sensitive, interactive (VoIP, VTC).
2	Transaction data, highly interactive (Signalling)
3	Transaction data, interactive
4	Low loss only (short transactions, bulk data, video streaming)
5	Traditional applications of default IP networks

g. QoS: Voice transport requirements (ETSI)

- **Delay**
  - E2E delay (Customer to Customer) < 250ms (no echo canceling is required)
  - objective is < 150ms
    - human ear starts to notice response delay above 150 ms
  - 400 ms is unacceptable, except for satellite links
- **Delay variation or jitter**
  - E2E should be < 40ms
  - Delay variation: example of ETSI TIPHON
    - <10 ms class 1 = gold
    - 10 ms to 20 ms class 2 = silver
    - 20 to 40 ms class 3 = bronze

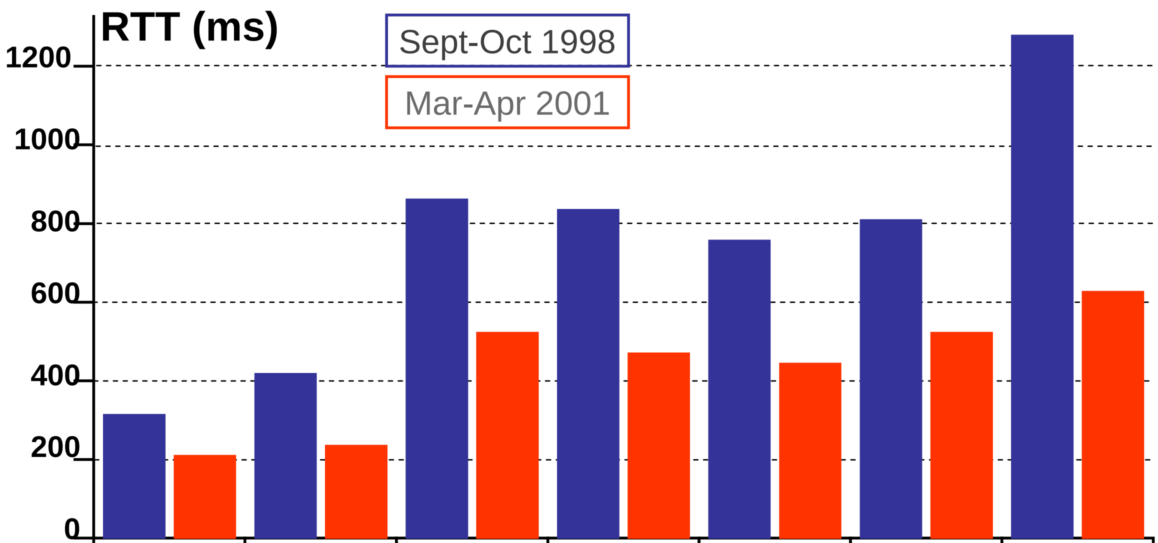
QoS: Voice transport requirements (Cntd)

- **Packet loss**
  - E2E packet loss for voice should be < 2%
  - E2E 64k transparent should be more stringent < x %
  - ETSI TIPHON (voice)
    - <0.5% class 1 = gold
    - 0.5% to 1% class 2 = silver
    - 1% to 2% class 3 = bronze
  - Provided the E2E delay < 150 ms all above classes are acceptable

Summary of network QoS requirements (ETSI)

Optimal network QoS parameters		Limits of network QoS parameters	
Delay – one way	<= 100ms	Delay – one way	<= 150ms
Jitter	<= 40ms	Jitter	<= 75ms
Packet loss	<= 1%	Packet loss	<= 3%

h. Examples: Internet measurements of RTT (from Belgium to a specific region)



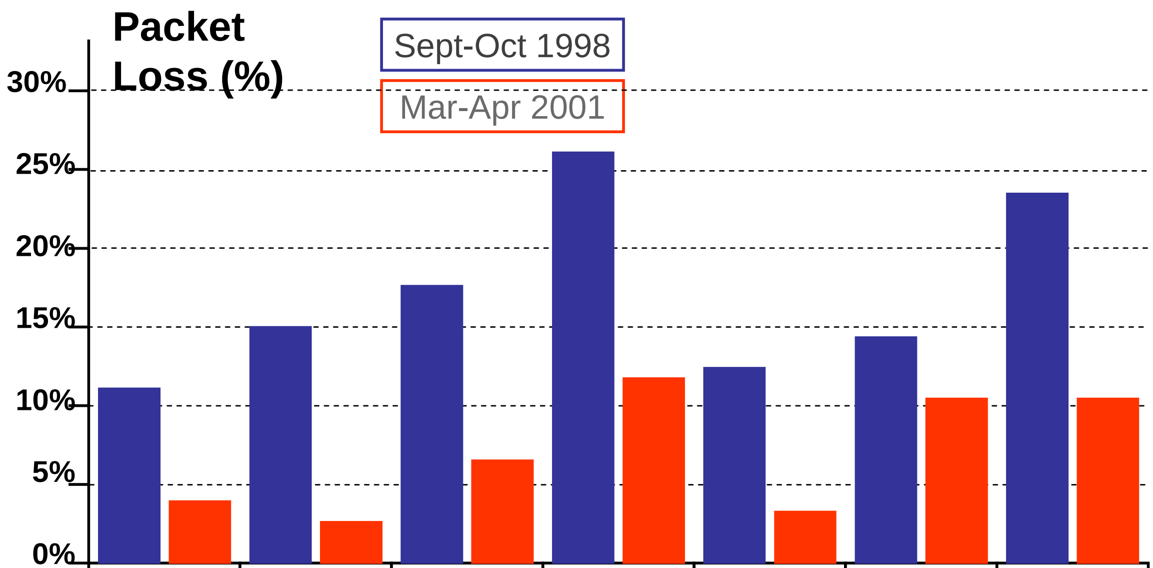
	Europe	North-America	South-America	Asia	Oceania	Africa	Middle-East
1998	353.3	417.3	882.6	841.3	738.8	808.4	1270.6
2001	204.4	219.7	509.6	461.8	441.0	521.4	620.9

RTT – round-trip time

Source: Alcatel

Internet measurements of packet loss

(from Belgium to a specific region)



	Europe	North-America	South-America	Asia	Oceania	Africa	Middle-East
1998	11.2%	15.3%	17.0%	26.6%	12.6%	14.4%	23.4%
2001	3.7%	2.4%	5.8%	12.1%	3.0%	10.1%	10.2%

Source: Alcatel

3. Estimation of call quality

A. Data and Voice network performance requirements (in general)

DATA

• File transfer applications - big volumes, big resources,
• E-mails - small volumes, tolerance to losses
• Tolerance to delay
• TCP

VoIP applications

Relatively little bandwidth, but can’t tolerate large delays, variations, losses.

• Intolerance to delay
• Packets are sent at different rates
• RTP/UDP for voice
• Packets are buffered and delivered to the destination differently

Delays caused by other applications, overloaded routers, or faulty switches may be inevitable for VoIP apps

B. Standards for measuring call quality

•Quality goal for a VoIP call is the PSTN level of quality (“toll” quality)

•But what is in IP networks? We need to understand some of the different measurement standards for voice quality

•The leading subjective measurement of voice quality - Mean Opinion Score (MOS) – Recommendation ITU P.800.

•The Mean Opinion Score (MOS) described in ITU P.800 is a subjective measurement of call quality as perceived by the receiver. (Estimation by group of listeners).

•MOS can range from 5 down to 1, using the following rating scale (see Table).

This mapping between audio performance characteristics and a quality scores makes the MOS (Mean Opinion Score) standard valuable for network assessments, benchmarking, tuning, and monitoring

MOS	Quality Rating
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

The MOS is measured
on a scale from 5 down to 1

Appropriateness of MOS in VoIP apps

• MOS and other methods (P.861- PSQM, P.862 - PESQ)*, are based in older telephony approaches. These approaches are not very well suited to assessing call quality on a data network, as they can't take into account the network issues of delay, jitter, and packet loss.

• Models don't take into account E2E delay between the telephone speaker and listener. Excessive delay adversely affects MOS.

• Models show quality in one direction at a time.

* PSQM - Perceptual Speech Quality Measure
  PESQ - Perceptual Evaluation of Speech Quality

C. E-model - Impairment factor method

- Recommendation ITU G.107 introduced the E-model. The E-model is better suited for use in data network call quality assessment because it takes into account impairments specific to data networks.

- The output of an E-model calculation is a single scalar, called an “R-value” or R-factor derived from delays and equipment impairment factors. Once an R value is obtained, it can be mapped to an estimated MOS.

E-model (cntd.)

- IFM is based on the assumption that transmission impairments can be transformed into psychological factors and that these psychological factors are additive on the "Psychological scale“.
- An appropriate mathematical algorithm is provided by the E-Model, with which the different transmission parameters can be transformed into different "impairment factors".
- With the E-Model, a very useful tool is available, which provides a simplified and easy-to-handle method for practical planning purposes.
- The final result of any calculation with the E-Model is the E-Model Rating R. The relation between the different impairment values and R is given by the equation:

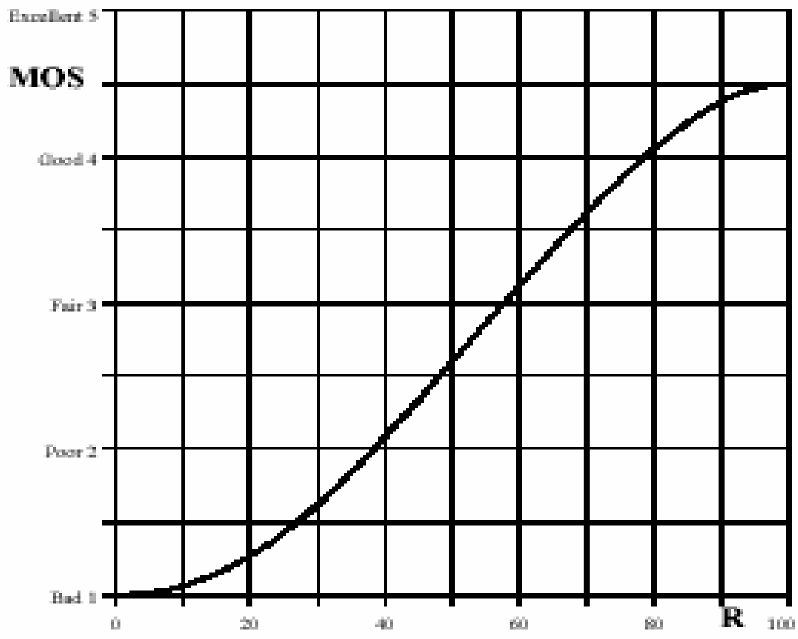
$$R = Ro - Is - Id - Ie + A$$

E-model (cntd.)

R_0	the basic signal-to-noise ratio based on sender and receiver loudness ratings and the circuit and room noise
I_s	the sum of real-time or simultaneous speech transmission impairments, e.g. loudness levels, sidetone and PCM quantizing distortion
I_d	the sum of delay impairments relative to the speech signal, e.g., talker echo, listener echo and absolute delay
I_e	the equipment impairment factor for special equipment, e.g., low bit-rate coding (determined subjectively for each codec and for each % packet loss and documented in ITU-T Recommendation G.113)
A	the advantage factor adds to the total and improves the R-value for new services.

Range of R-factor is from 100 to 0

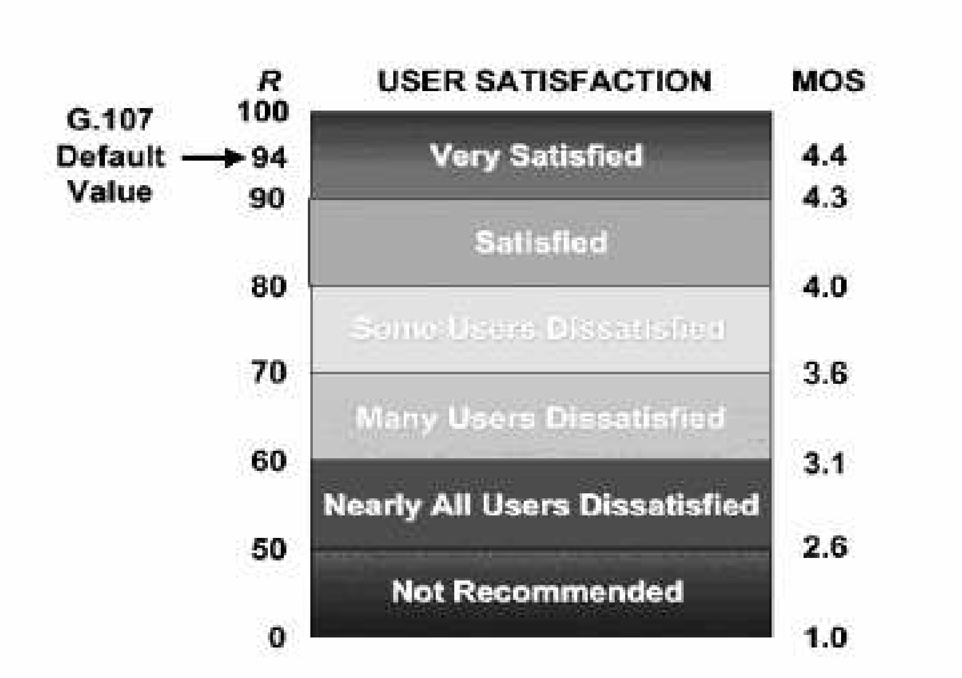
D. R-factor values from the E-model and corresponding MOS values



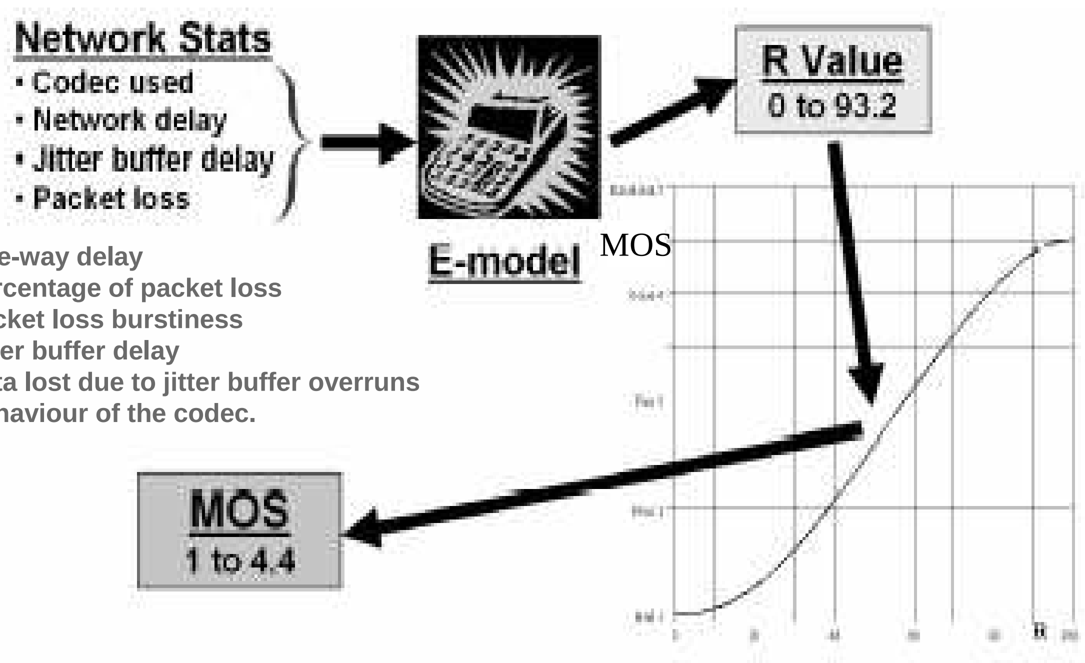
E-model

The R value, the output from the E-model, ranges from 100 down to 0, where 100 is excellent and 0 is poor. The calculation of an R value starts with the undistorted signal.

R-factor values from the E-model and corresponding MOS values (Cntd)



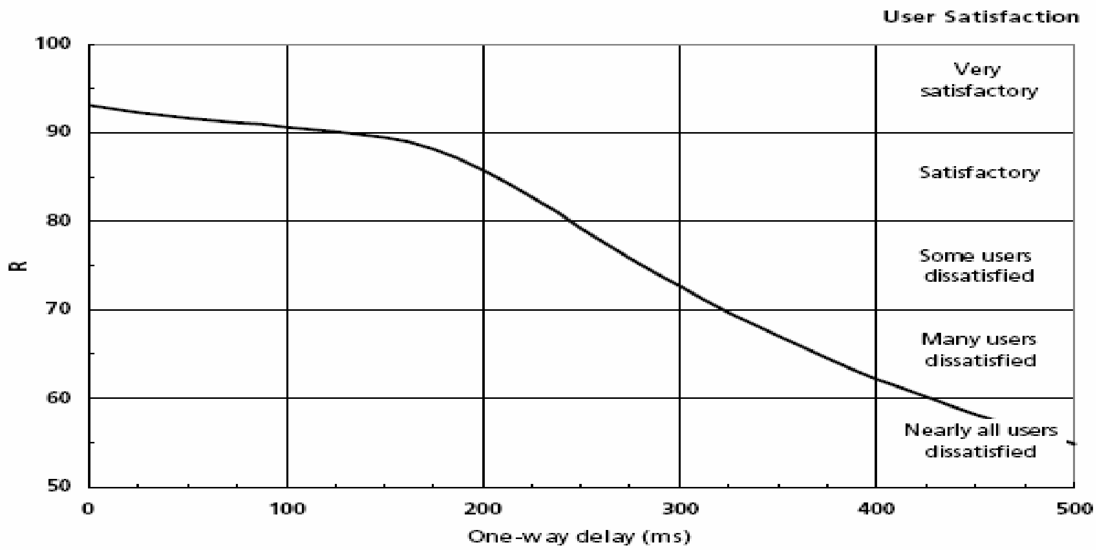
R-factor values from the E-model and corresponding MOS values (Cntd)



- One-way delay
- Percentage of packet loss
- Packet loss burstiness
- Jitter buffer delay
- Data lost due to jitter buffer overruns
- Behaviour of the codec.

E. Impact of delay, type of codecs and packet loss on the R-factor

1. Delay Impairment of Reference Connection (I_d)



Delay (ms)	0	50	100	150	200	250	300	350	400
R	93.19	91.74	90.65	89.53	85.79	79.17	72.66	67.02	62.24

2. Speech codecs and their I_e values (G.113)

Codec Type	Codec	Bit Rate (Kbps)	Ie Value
PCM	G.711	64	0
ADPCM	G.726 G.726 G.726	40 32 24	2 7 25
LD-CELP	G.728	16	7
CS-ACELP	G.729-A + VAD	8	11
RPE-LTP	GSM-Full Rate	13	20
VCELP	GSM-Half Rate	5.6	23
ACELP	GSM-EFR	12.2	5
MP-MLQ	G.723.1	5.3	19
MP-MLQ	G.723.1	6.3	15

Types of codecs

- **G.726 – ADPCM -** Adaptive Differential Pulse Code Modulation (32, 24, 16 kbit/s).
- **G.728 - LD-CELP** - Low-Delay Code-Excited Linear Prediction (16, 12.8, 9.6 kbit/s)
- **G.729 - CS-ACELP -** Conjugate Structure Algebraic Code-Excited Linear Prediction (8 kbit/s)
- **G.723-1 - ACELP** - Algebraic Code-Excited Linear Prediction (5.3 kbit/s) and **MP-MLQ -** Multipulse Maximum Likelihood Quantization (6.3 kbit/s)
- **GSM-FR - RPE-LTP** - Regular Pulse Excitation Long Term Prediction (13 kbit/s)
- **GSM-HR - VSELP** - Vector Sum Excited Linear Prediction (5.6 kbit/s)

3. Packet loss impairment and their I_e values (G.113)

Packet loss (%)	G.711 w/o PLC	G.711 + PLC RPL	G.711 + PLC BPL	G. 729A 8 kbs	G.723.1 6.3 kbs
0	0	0	0	11	15
1	25	5	5	15	19
2	35	7	7	19	24
3	45	10	10	23	27
4	–	–	–	26	32
5	55	15	30	–	–

4. Examples of Advantage Factor A

Communications system	A value
• Usual wired phone	0
• Cellular in building	5
• Cellular in moving vehicle	10
• Access to hard-to-reach geographical zones (many satellite hops)	20

Estimation of R-factor

- R_0
- I_d
- I_e

F. Codec's selection

• In audio processing - a codec is the hardware or software that samples the sound and defines the data rate of digital output. There are, each with different characteristics

•Dozens of available codecs

•Types of codecs correspond to the certain ITU standards

• First codecs - G.711a/G.711 - 64 kb/s (PCM) – ADC with no compression and high quality

• New generation of codecs based on new compression algorithms New codecs provide intelligible voice communications with reduced bandwidth consumption.

•The lower-speed codecs
 # G.726-32 (32 kb/s)
 # G.729 (8 kb/s)
 # G.723.1 family (6.3/5.3 kb/s)
• New codecs consume less network bandwidth – bigger number of concurrent calls
• BUT - bigger impairment on the quality of the audio signal than high-speed codecs, the compression reduces the clarity, introduces delay, and makes the voice quality very sensitive to a packet loss

Parameters of VoIP codecs

- **MOS and R value include Pack delay and Jitter buffer delay**
- **Common bandwidth – real bandwidth consumption:**
  # Payload = 20 bytes/p (40 bytes/s)
  # Overhead includes 40 bytes of RTP header (20 IP + 8 UDP + 12 RTP)
- **G.723.1 – Quality is "Acceptable" only**

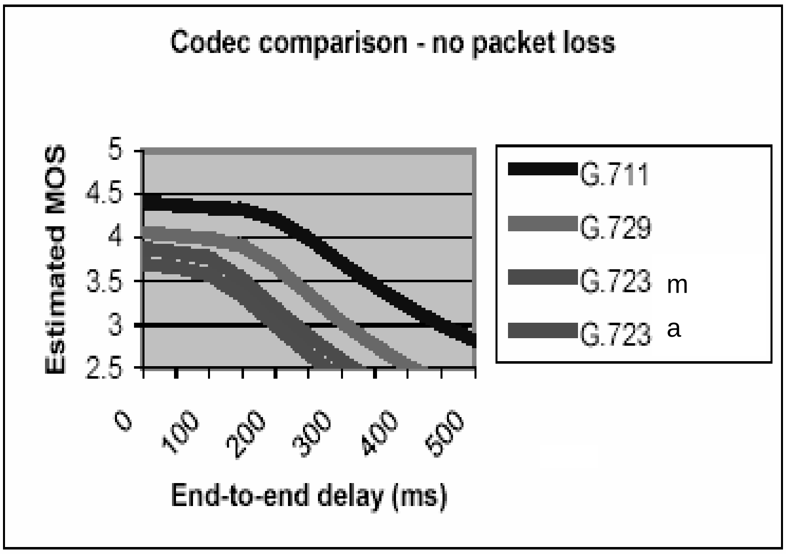
Codec	Data Rate	Typical Datagram Size	Packeti -zation Delay	Combined Bandwidth for 2 Flows	Typical Jitter Buffer Delay	Theoretical Maximum MOS
G.711u	64.0 kbps	20 ms	1.0 ms	174.40 kbps	2 datagrams (40 ms)	**4.40**
G.711a	64.0 kbps	20 ms	1.0 ms	174.40 kbps	2 datagrams (40 ms)	**4.40**
G.726-32	32.0 kbps	20 ms	1.0 ms	110.40 kbps	2 datagrams (40 ms)	**4.22**
G.729	8.0 kbps	20 ms	25.0 ms	62.40 kbps	2 datagrams (40 ms)	**4.07**
G.723.1 m	6.3 kbps	30 ms	67.5 ms	43.73 kbps	2 datagrams (60 ms)	**3.87**
G.723.1 a	5.3 kbps	30 ms	67.5 ms	41.60 kbps	2 datagrams (60 ms)	**3.69**

Common voice codecs and corresponding audio quality

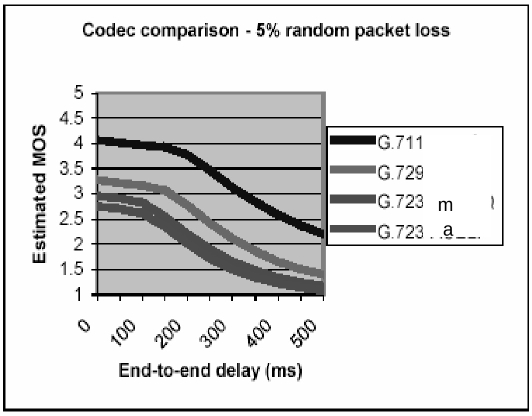
Codec	Nominal Data Rate (kbps)	Amount subtracted from the R-factor
G.711	64.0	0
G.729	8.0	11
G.723.1m	6.3	15
G.723.1a	5.3	19

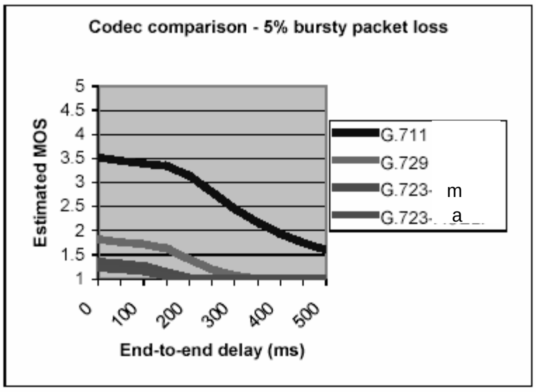
Codec	R-factor	MOS
G.711	93.2	4.4
G.729	82.2	4.1
G.732.1m	78.2	3.9
G.723.1a	74.2	3.75

Codecs' comparison



Codec	R-factor	MOS
G.711	93.2	4.4
G.729	82.2	4.1
G.732.1m	78.2	3.9
G.723.1a	74.2	3.75





Codecs' comparison (Cntd)

• Any lost datagram impairs the quality of the audio signal. Data loss is thus a key call-quality impairment factor in calculating the MOS.

• Random loss –simplest loss model

One datagram is lost or two datagrams are lost time by time

Small effect inside of delay limit (<=150 ms)

• Bursts of loss

Quality degrades most significantly

More than two consecutive datagrams are lost

• 5% random packet loss (upper Figure)

MOS starts at around 4 for the G.711 codec

• 5% bursty packet loss (Figure below)

MOS starts at around 3.5 for the same codec

• The effect of bursty loss is even greater on the other codecs

Codec	R-factor	MOS
G.711	93.2	4.4
G.729	82.2	4.1
G.732.1m	78.2	3.9
G.723.1a	74.2	3.75

G. List of VoIP network design tips

Main factors QoS of VoIP - delay, jitter and packet loss. Following design tips could be useful during VoIP deployment process

Use the G.711 codec on E2E if a capacity is enough

G.711 codec gives the best voice quality - no compression, minimum delay, less sensitive to packet loss

Other codecs - G.729 and G.723 use compression. Results – economy of a bandwidth, but delay is introduced and the voice quality very sensitive to lost packets

Keep packet loss well below 1% and avoid bursts of consecutive lost packets

Sources of packet loss - channel noise, traffic congestion and jitter buffer size

Tools - Increased bandwidth and TE can often reduce network congestion, which, in turn, reduces jitter and packet loss

Use a small speech frame size and reduce the number of speech frames per packet

Using small packets/datagrams (in ms) - impact of the packet loss is less than losing a big packets

One of standard size - 20ms of speech frame per datagram. Of course, using small packets increases an overhead conditions, because each packet requires its own fixed-size header

Always use codecs with packet-loss concealment (PLC)

PLC masks the loss of a packet or two by using information from the last good packet

PLC helps with random packet loss

List of VoIP network design tips (Cntd)

Actively minimize one-way delay, keeping it below 150ms

E2E Delay = PrD + TD + PcD + JBD <= 150ms

• PrD – physical distance (4-5 mcs/km)
 # Routing – network path is ADAP

• TD – all network devices (routers, gateways, TE tools, firewalls)
 # Factors – number of hopes, software/hardware processing

• PD - fixed time needed for the AD conversion
 # G.711 - adds 1ms
 # G.723 – adds 67.5ms
 # E2E – the same type of codecs

• JBD - to decrease variations in packet arrival rates
 # Larger jitter buffer than in a network where the delay is already high.

Avoid using slow speed links

Use RTP header compression for links with the lack of capacity
 # CRTP can reduce the 40-byte RTP headers to a tenth of their original size
 # Decreasing the bandwidth consumption
 # BUT - it adds latency.

List of VoIP network design tips (Cntd)

Use call admission to protect against too many concurrent calls
#Call Admission Control

Use priority scheduling for voice traffic

DiffServ (EF)
Queuing mechanisms - giving VoIP higher priority

Get your data network ready for VoIP
In general, unsatisfactory data networks
Network should be fully upgraded and tuned, before starting a VoIP deployment

4. IPTV QoS issues

A. Bandwidth Dimensioning

IPTV and VoD services require high bandwidth capacities and predictable performance, placing additional requirements on the network. Depending on the compression and coding technology the following transmission rates should be considered:

• H.264 (MPEG-4 part 10) coded SD VoD video streams or IPTV stream per one TV channel: up to 2 Mbit/s
• HD signals will need 8-12 Mbit/s coded with H.264
• MPEG-2 coded SD VoD video streams or IPTV stream per one TV channel: 3,5 – 5 Mbit/s

:

B. QoS issues for multimedia traffic

- **Voice traffic** is smooth, drop-sensitive, and delay-sensitive, and is typically UDP-based. Bandwidth per call depends on the particular codes adopted, sampling rate, and Layer 2 media employed. Voice quality is directly affected by all three QoS quality factors (loss, delay, and delay variation).

- **Data traffic** is much more varied. It can be smooth or bursty, benign or greedy, or drop- and delay-insensitive, and involves Transmission Control Protocol (TCP) for send/receive acknowledgment and retransmit. Traffic patterns vary by application, and data classes must support several different priorities or application categories.

- **Video traffic** is bursty, bandwidth-greedy, drop-sensitive, and delay-sensitive. IP-based videoconferencing has some of the same sensitivities as voice traffic.

Traffic attributes of different services

	Data Service	Digital Video Distribution Service	Digital Video Communication Service	Voice Service
Directionality	Asymmetrical	Asymmetrical	Symmetrical	Symmetrical
Burstiness	High	Medium/High	Medium/High	Low
Time Sensitivity	Non-real time	Non-real time	Real time	Real time
Holding Time	Medium/Long	Long	Short/Medium	Short
Bandwidth Requirement	Medium	High	High	Low
Performance Metrics	Throughput	Throughput, Video quality	Delay, Video quality	Delay, Voice quality

Model for QoS categories (G.1010)

	Interactive (delay <<1 s)	Responsive (delay ~2 s)	Timely (delay ~10 s)	Non-critical (delay >>10 s)
Error tolerant	Conversational voice and video	Voice/video messaging	Streaming audio and video	Fax
Error intolerant	Command/control (e.g. Telnet, interactive games)	Transactions (e.g. E-commerce, WWW browsing, Email access)	Messaging, Downloads (e.g. FTP, still image)	Background (e.g. Usenet)

Performance targets for audio and video applications (G.1010)

Medium	Application	Degree of symmetry	Typical data rates	Key performance parameters and target values			
				One-way delay	**Delay variation**	**Information loss (Note 2)**	**Other**
Audio	Conversational voice	Two-way	4-64 kbit/s	<150 ms preferred (Note 1) <400 ms limit (Note 1)	< 1 ms	< 3% packet loss ratio (PLR)	
Audio	Voice messaging	Primarily one-way	4-32 kbit/s	< 1 s for playback < 2 s for record	< 1 ms	< 3% PLR	
Audio	High quality streaming audio	Primarily one-way	16-128 kbit/s (Note 3)	< 10 s	<< 1 ms	< 1% PLR	
Video	Videophone	Two-way	16-384 kbit/s	< 150 ms preferred (Note 4) <400 ms limit		< 1% PLR	
Video	One-way	One-way	16-384 kbit/s	< 10 s		< 1% PLR	

NOTE 1 – Assumes adequate echo control.

NOTE 2 – Exact values depend on specific codec, but assumes use of a packet loss concealment algorithm to minimise effect of packet loss.

NOTE 3 – Quality is very dependent on codec type and bit-rate.

NOTE 4 – These values are to be considered as long-term target values which may not be met by current technology.

Performance targets for data applications (G.1010)

Medium	Application	Degree of symmetry	Typical amount of data	Key performance parameters and target values		
				One-way delay (Note)	Delay variation	Information loss
Data	Web-browsing – HTML	Primarily one-way	~10 KB	Preferred < 2 s /page Acceptable < 4 s /page	N.A.	Zero
Data	Bulk data transfer/retrieval	Primarily one-way	10 KB-10 MB	Preferred < 15 s Acceptable < 60 s	N.A.	Zero
Data	Transaction services – high priority e.g. e-commerce, ATM	Two-way	< 10 KB	Preferred < 2 s Acceptable < 4 s	N.A.	Zero
Data	Command/control	Two-way	~ 1 KB	< 250 ms	N.A.	Zero
Data	Still image	One-way	< 100 KB	Preferred < 15 s Acceptable < 60 s	N.A.	Zero
Data	Interactive games	Two-way	< 1 KB	< 200 ms	N.A.	Zero
Data	Telnet	Two-way (asymmetric)	< 1 KB	< 200 ms	N.A.	Zero
Data	E-mail (server access)	Primarily one-way	< 10 KB	Preferred < 2 s Acceptable < 4 s	N.A.	Zero
Data	E-mail (server to server transfer)	Primarily one-way	< 10 KB	Can be several minutes	N.A.	Zero
Data	Fax ("real-time")	Primarily one-way	~ 10 KB	< 30 s/page	N.A.	$<10^{-6}$ BER
Data	Fax (store & forward)	Primarily one-way	~ 10 KB	Can be several minutes	N.A.	$<10^{-6}$ BER
Data	Low priority transactions	Primarily one-way	< 10 KB	< 30 s	N.A.	Zero
Data	Usenet	Primarily one-way	Can be 1 MB or more	Can be several minutes	N.A.	Zero

NOTE – In some cases, it may be more appropriate to consider these values as response times.

Traffic classes

In general, enterprises should restrict themselves to about five main traffic classes, such as:

- **Mission-critical and real-time** - Interactive applications with high business priority;
- **Transactional/interactive** - Client-server applications, messaging applications
- **Bulk** - Large file transfers, e-mail, network backups, database synchronization and replication, and video content distribution
- **Best-effort** - Default class for all unassigned traffic; typically at least 25 percent of bandwidth is reserved for best-effort traffic
- **Scavenger** (optional)—Peer-to-peer media sharing applications, gaming traffic, and entertainment traffic

QoS requirements for video applications

QoS requirements for interactive video traffic:

- # Packet loss should be no more than 1 percent.
- # One-way latency should be no more than 150 ms.
- # Jitter should be no more than 30 ms.
- # The minimum priority bandwidth guarantee is the size of the video session plus 20 percent. (For example, a 384 kbps video conferencing session requires 460 kbps of guaranteed priority bandwidth.)

QoS requirements for for streaming video traffic:

Loss should be no more than 2 percent.
Latency should be no more than 4-5 seconds (depending on video application's buffering capabilities).
There are no significant jitter requirements.
Guaranteed bandwidth requirements depend on the encoding format and rate of the video stream.

Scavenger Class

The *Scavenger* class is intended to provide “less-than Best-Effort” services, to certain applications.

Applications are typically entertainment-oriented and include:

- Peer-to-peer media-sharing applications (KaZaa, Morpheus, Groekster, Napster, iMesh, etc.)
- Gaming applications (Doom, Quake, Unreal Tournament, etc.), and any entertainment video applications.

5. QoS guarantees
Possible approaches to the problem

a. Over-provisioning the core network
b. Reservation and service differentiation - IP QoS mechanisms
c. Congestion avoidance mechanisms by reservation
d. Service differentiation using IP QoS mechanisms

a. Over-provisioning the core network

Assumption: physical bandwidth is available to scale and cheap bandwidth will be plentiful (based on FOC networks). The cost of bandwidth in the FOC backbones is decreasing, since:

@ The supply of long-distance fiber in the ground currently exceeds the demands for it

@ DWDM technology → the specific cost of a capacity and the specific cost of a transmission is almost zero

Provisioning can be planned

The capacity of the access tributaries is known, and the combined data rate cannot exceed the sum of the access links. As orders for faster access links are received, a decision can be made (taking also into account the current measured traffic load) whether or not it is necessary to upgrade the backbone capacity.

Over-provisioning the core network (Cntd)

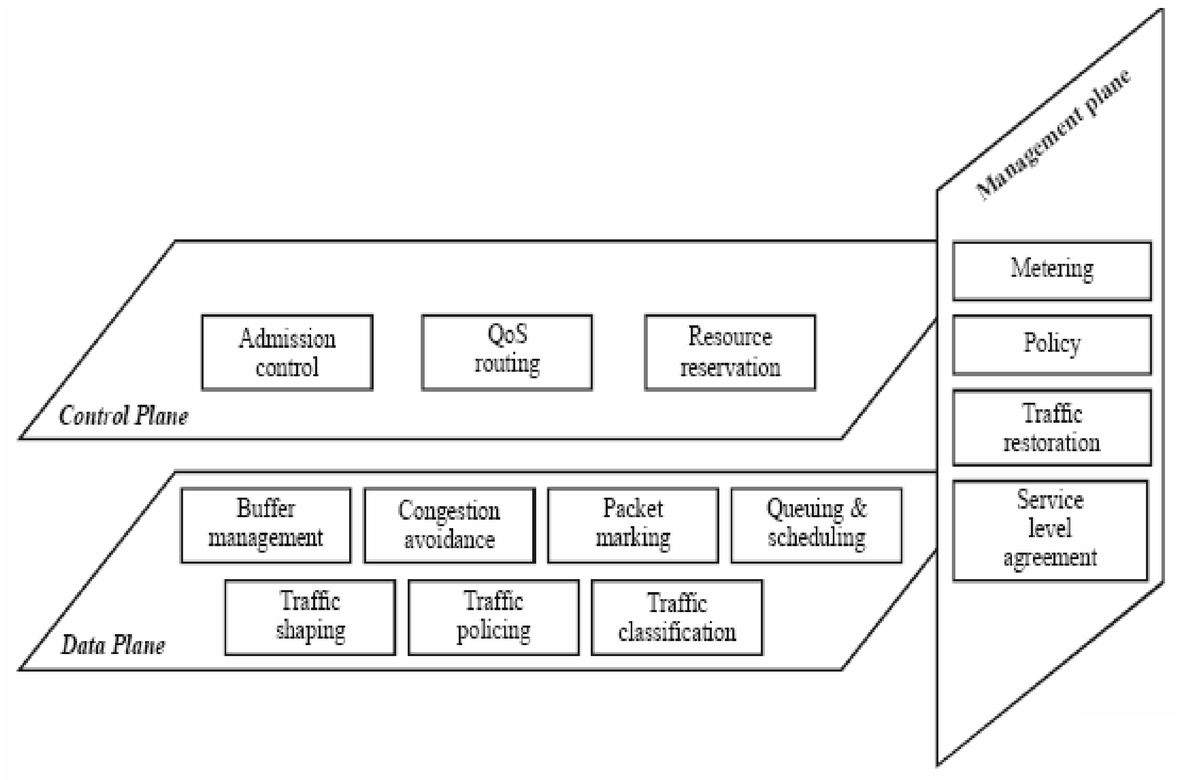
– Ultimately, the main argument for the QoS decision via over-provisioning - the availability of fiber. So this does not apply to all networks, and, of course, not to the edges of the network

– Over-provisioning the core is a short-term solution. As access capacity progressively increases, backbone networks will become susceptible to congestion and overloading

b. Reservation and service differentiation - IP QoS mechanisms

- QoS on IP can be delivered on the base of mechanisms:
  - IntServ (Integrated Services)
  - DiffServ (Differentiated Services)

QoS building blocks (Y.1291)



c. Reservation mechanisms

•Integrated Services (IntServ)

IETF Integrated Services (IntServ) Working Group developed a service model based on the principle of integrated resource reservation.

The group of IntServ mechanisms (first of all, *RSVP)* refers to a group of methods providing a “hard” QoS.

RSVP (Resource ReSerVation Protocol) mechanism is the most well known representative of the IntServ mechanisms (RFC 2205, 1997).

RSVP is a signaling protocol according to which reservation and resource allocation is carried out to guarantee “hard” QoS. Reservation is accomplished for the certain IP packet flow before the main flow transmission start up.

Hard requirements to network resources

Integrated Services (IntServ)

- Flow = stream of packets with common Source Address, Destination Address and port number
- Requires router to maintain state information on each flow; router determines what flows get what resources based on available capacity

IntServ components

- Traffic classes
  - best effort
  - controlled load (‘best-effort like’ w/o congestion)
  - guaranteed service (real-time with delay bounds)
- Traffic control
  - admission control
  - packet classifier
  - packet scheduler

IntServ components (cont.)

- Setup protocol: RSVP
- “Path” msg from source to destination collects information along the path; the destination gauges what the network can support, then generates a “Resv” msg
- If routers along the path have sufficient capacity, then resources back to the receiver are reserved for that flow; otherwise, RSVP error messages are generated and returned to the receiver
- Reservation state is maintained until the RSVP “Path” and “Resv” messages stop coming

IntServ/RSVP problems

- Scalability (processing of every individual flow on core Internet routers)
- Lack of policy control mechanisms

d. Service differentiation using IP QoS mechanisms

Differentiated Services (DiffServ)

- DiffServ concept and mechanisms

  # Necessity to develop more flexible mechanisms for providing QoS

  # The detailed specifications of DiffServ (RFC 2475) - in the middle 1999.

  # As against IntServ group the DiffServ methods provide a “relative” or “soft” QoS.

- The main idea of DiffServ mechanisms to provide differentiated services to a set of traffic classes characterized by various requirements to QoS parameters

- One of the central point of DiffServ model is the ***Service Level Agreement (SLA)***

  # SLA – the contract between the user and the service provider

  # SLA - basic features of users’ traffic and QoS parameters ensured by providers

  # SLA - static or dynamic contract

Differentiated Services (DiffServ) - Cntd

- Main issues of QoS - priorities
  The support of a satisfactory QoS:
  - means for labeling flows with respect to their priorities
  - network mechanisms for recognizing the labels and acting on them

- According the IETF Differentiated Services model the network architecture includes two areas - edge segment and core segment

- In the edge routers a short tag is appended to each packet depending on its Class of Service (CoS)

- DS byte - ToS (IPv4) or TC (IPv6)

Differentiated Services (DiffServ) - Cntd

Network mechanisms

- Edge routers

  #*Traffic Classification* mechanism (to select the packets of one flow featured by common requirement to QoS)

  # *Conditioning* mechanism If necessary a part of packets can be discarded.

  # *Shaping* mechanism (if required)

- Backbone routers

  # Packets forwarding in compliance with the required QoS level

  # Two forwarding classes are specified - Expedited Forwarding (EF) and Assured Forwarding (AF).

  # EF class provides the *Premium Service* (apps requiring forwarding with minimum delay and jitter)

  # AF class maintains a lower QoS than EF class, but higher than BES

  # AF class identifies 4 classes of traffic and three levels of packet discarding – 12 types of traffic depending on the set of the required QoS

Differentiated Services (DiffServ) - Cntd

- Queuing mechanisms
  # Target - a control of a packet delay and jitter and elimination of possible losses
  # Based on priority level and type of traffic
  # Mechanisms
- *Priority Queuing*
- *Weighted Fair Queuing*
- *Class-Based Queuing*

- In the past - QoS planners supported both IntServ and DiffServ. At present - DiffServ supplemented by RSVP at the edges. At the edges of the network, resources tend to be more limited, and there are not so many flows to maintain

6. QoS - Concluding remarks

- Real-time applications should be supported by manufacturers' products due to reliability and Quality of Service capabilities
- QoS demanding applications come from:
  - introduction of multimedia
  - bypass of voice networks (e.g. Long-Distance Bypass)
  - growth in the voice networks
  - migration of voice to data networks