Deploying IP and MPLS QOS for Multiservice Networks

QOS Requirements and Service Level Agreements

LECTURE 4

Lecturer: Associate Professor A.S. Eremenko

Application SLA Requirements

Different applications have different SLA requirements; the impact that different network services with different SLAs have on an application is dependent upon the specific application:

- ✓ Excessive packet loss or delay may make it difficult to support real-time applications although the precise threshold of "excessive" depends on the particular application.
- ✓ The larger the value of packet loss or network delay, the more difficult it is for transport-layer protocols to sustain high bandwidths.

We consider **the most common applications or application types**, which impose the tightest SLA requirements on the network. In practice, most applications that have explicit SLA requirements will fall into one of the following categories, or will have SLA requirements, which are similar to one of those categories described:

- ❑ voice over IP;
- ❑ video streaming;
- ❑ video conferencing;
- ❑ throughput-focused TCP applications;
- ❑ interactive data applications;
- ❑ on-line gaming.

Voice over IP

Voice over IP (VoIP) is most commonly transported as a digitally encoded stream using the Real-time Protocol (RTP) [RFC3550] **over UDP**; RTP is the transport layer protocol, which deals with the delivery of the VoIP bearer stream from sender to receiver. **Signaling protocols such as the Session Initiation Protocol (SIP)** [RFC3261] **may be used to set up the RTP bearer streams and to determine the media formats** (i.e. **codecs**) **that will be used**.

The key factors that determine the impact that variations in networks SLA characteristics such as delay and loss have on VoIP are the codec that is used to encode the signal and the specific details of the end-system implementation. The most widely used codecs are those defined by the **ITU G.71x** and **G72x standards**.

The codecs available for VoIP vary in complexity, in the bandwidth they need, and in the delivered call quality perceived by the end-user. Algorithms that are more complex may provide better perceived call quality, but may incur longer processing delays; **Figure 3** shows the functional components in VoIP end-systems, which contribute to delay. The table compares characteristics of some of the more common VoIP codecs.

Voice over IP

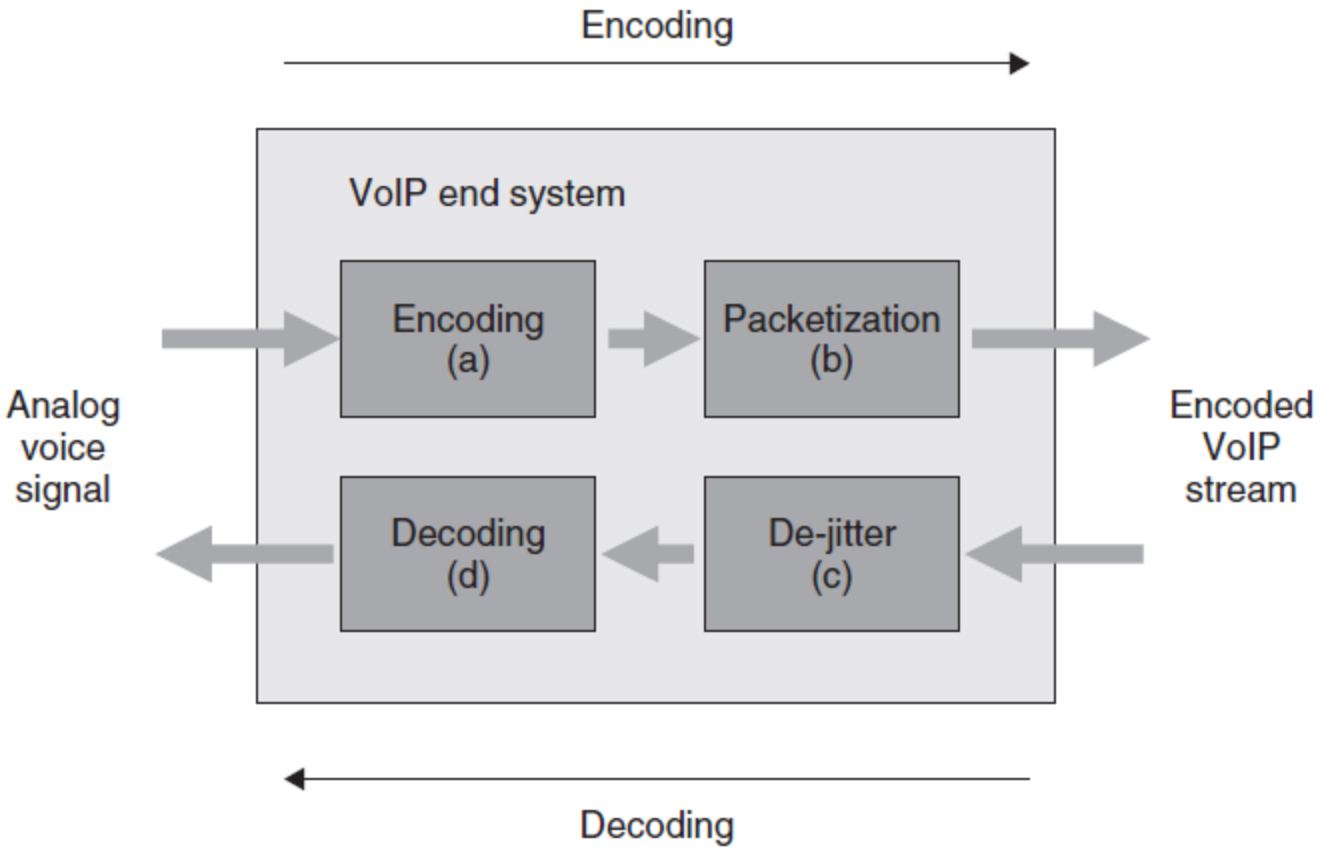

Figure 3 VoIP end-systems components of delay

Voice over IP codec characteristics

ITU-T Codec	Codec type	Maximum codec delay (ms) (a1 d)	Bitrate (bps)	Packetization interval (ms) (b)	pps	Payload size (bytes)	IP pkt size (bytes)[i]	IP bps
G.711	PCM	0.375	64 000	10	100	80	120	96 000
G.711	PCM	0.375	64 000	20	50	160	200	80 000
G.711	PCM	0.375	64 000	30	33.33	240	280	74 659
G.723.1	ACELP	97.5	5 300	30	33.33	20	60	15 998
G.723.1	ACELP	97.5	5 300	15	16.67	40	80	10 669
G.726.16	ADPCM	0.375	16 000	10	100	20	60	48 000
G.726.16	ADPCM	0.375	16 000	20	50	40	80	32 000
G.726.16	ADPCM	0.375	16 000	30	33.33	60	100	26 664
G.726.24	ADPCM	0.375	24 000	10	100	30	70	56 000
G.726.24	ADPCM	0.375	24 000	10	50	60	100	40 000
G.726.24	ADPCM	0.375	24 000	10	33.33	90	130	34 663
G.726.32	ADPCM	0.375	32 000	10	100	40	80	64 000
G.726.32	ADPCM	0.375	32 000	20	50	80	120	48 000
G.726.32	ADPCM	0.375	32 000	30	33.33	120	160	42 662
G.726.40	ADPCM	0.375	40 000	10	100	50	90	72 000
G.726.40	ADPCM	0.375	40 000	20	50	100	140	56 000
G.726.40	ADPCM	0.375	40 000	30	33.33	150	190	50 662
G.728	LD-CELP	1.875	16 000	10	100	20	60	48 000
G.728	LD-CELP	1.875	16 000	20	50	40	80	32 000
G.728	LD-CELP	1.875	16 000	30	33.33	60	100	26 664
G.729A	CS-ACELP	35	8 000	10	100	10	50	40 000
G.729A	CS-ACELP	35	8 000	20	50	20	60	24 000
G.729A	CS-ACELP	35	8 000	30	33.33	30	70	18 665

VoIP: Impact of Delay

For VoIP the **important delay metric** is the **one-way end-to-end** (i.e. **from mouth-to-ear**) **delay**, in each direction. The main impact that end-to-end delay has on VoIP is to the interactivity of conversational speech. If the delay is too high, participants find it difficult to discern the difference between natural pauses in speech and the delays introduced by the system. Excessive end-to-end delay can also impair the effectiveness of mechanisms used for echo-cancellation.

The goal commonly used in designing networks to support voice over IP (VoIP) is the target specified by **ITU-T recommendation G.114**, which uses the **E-model** to estimate the effects of delay on mouth-to-ear speech transmission quality. Recommendation G.114 suggests that ***150 ms*** of end-to-end one-way delay is sufficient to ensure that users will be very satisfied for most applications of telephony.

Ear-to-mouth delay (D)	R factor	Objective MOS
$D < 150$ ms	80–89	5
150 ms $< D <$ 250 ms	70–79	4
250 ms $< D <$ 325 ms	60–69	3
325 ms $< D <$ 425 ms	50–59	2
$D >$ 425 ms	90–100	1

Figure 4 ITU G.114 Determination of the effects of absolute delay by the E-model

VoIP: Impact of Delay

Having determined what the maximum acceptable ear-to-mouth delay is for a particular VoIP service, **a network QOS design should take this budget and apportion it to the various components of network delay** (***propagation delay through the backbone, scheduling delay due to congestion, and the access link serialization delay***) **and end-system delay (due to VoIP codec and de-jitter buffer)**.

The example timeline in **Figure 5** shows the components of delay.

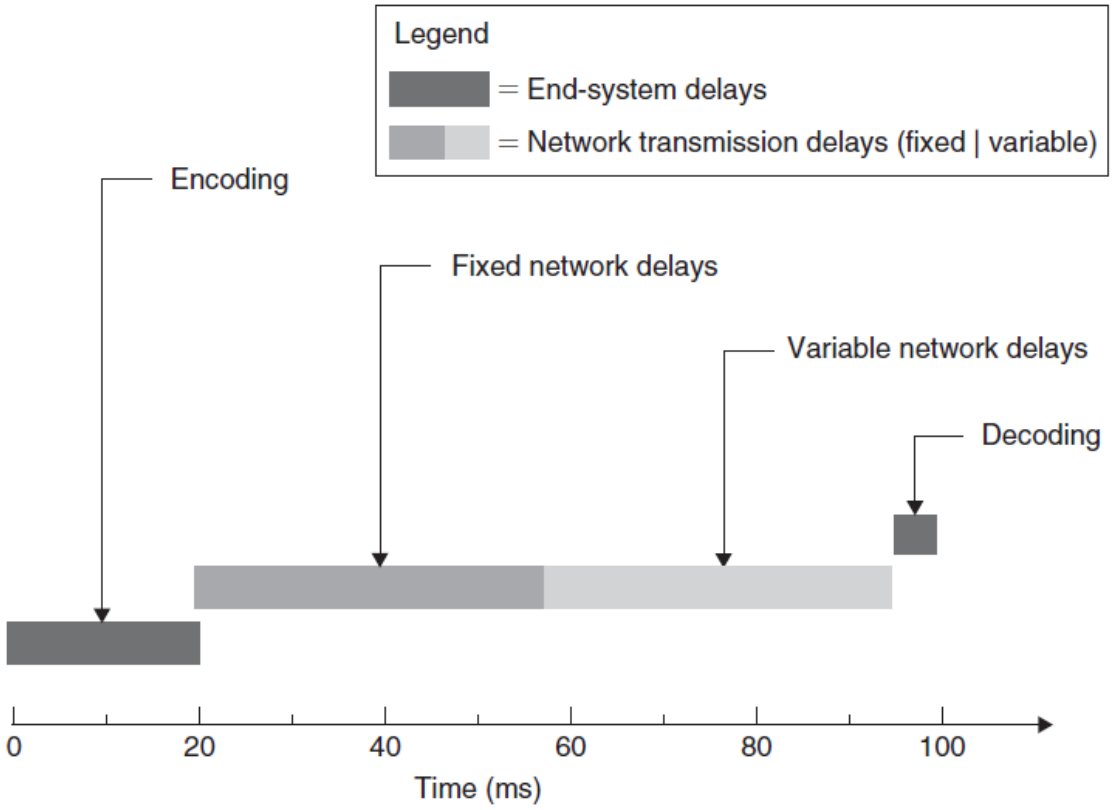

Figure 5 VoIP: components of ear-to-mouth delay

VoIP: Impact of Delay-jitter

It is a common misconception that jitter has a greater impact on the quality of VoIP calls than network delay. Applications which are susceptible to jitter, such as VoIP, use **de-jitter buffers** (also known as jitter buffers and play-out buffers) to compensate for jitter in packet arrival and for out-of-order packets. **De-jitter buffers remove delay variation by turning variable network delays into constant delays at the destination end-systems**.

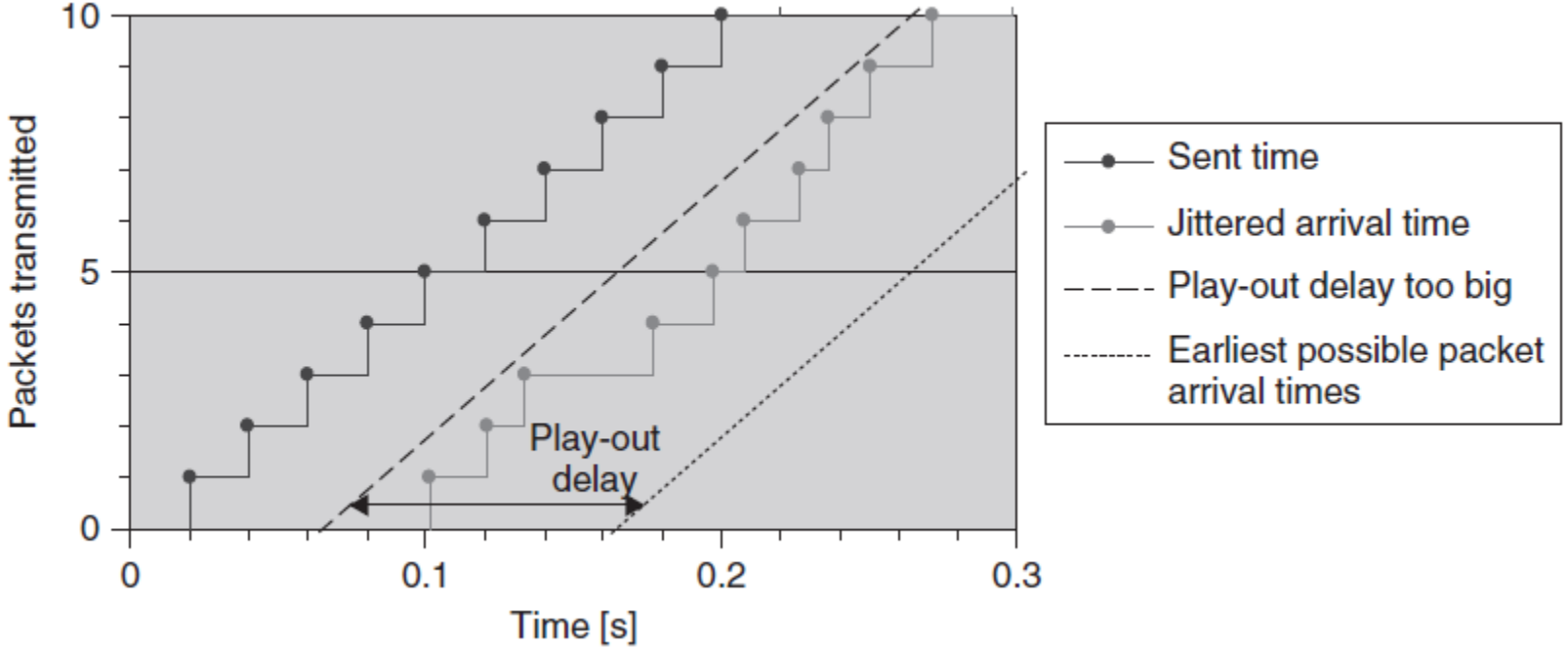

Figure 6 VoIP play-out delay unnecessarily large

VoIP: Impact of Delay-jitter

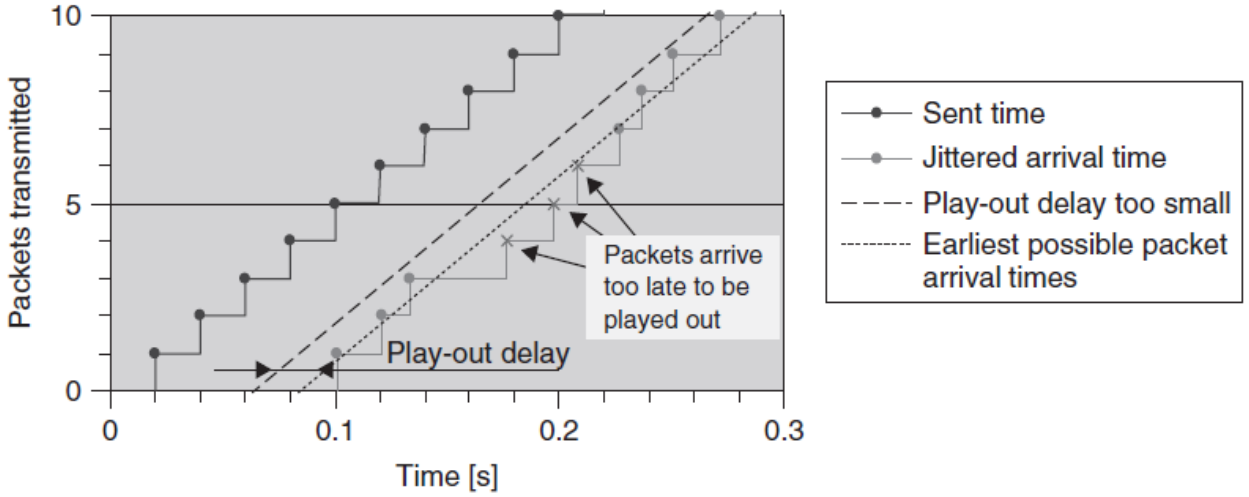

Figure 7 VoIP play-out delay too small

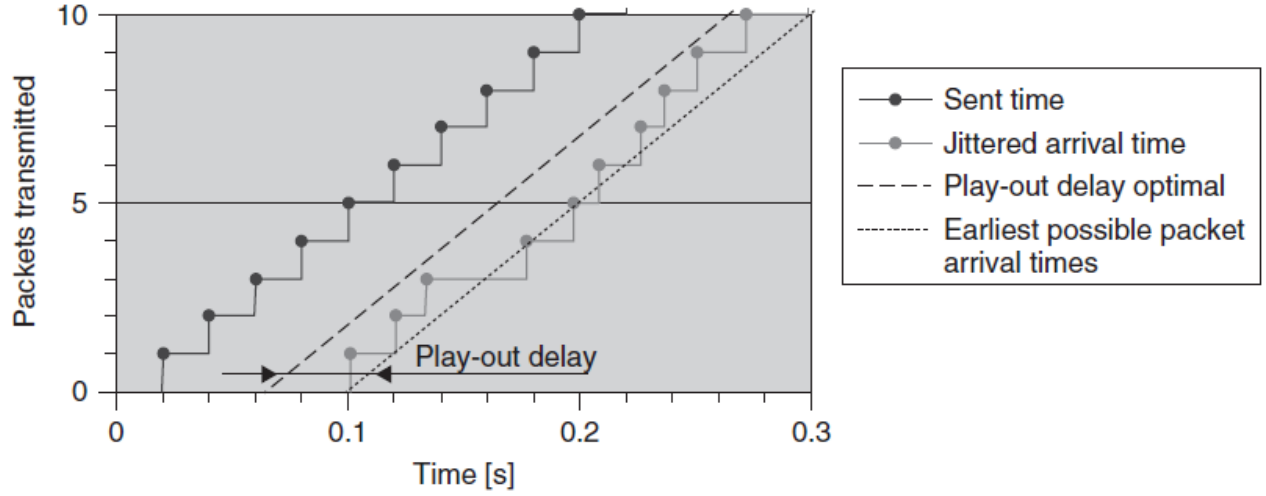

Figure 8 Optimal VoIP play-out delay

VoIP: Impact of Delay-jitter

Well-designed adaptive de-jitter buffer algorithms should not impose any unnecessary constraints on the network design if they display the following characteristics:

- increasing the play-out delay to the current measured jitter value following an underflow, and using packet loss concealment to interpolate for the “lost” packet and for the play-out delay size increase;

- if the play-out delay can decrease then it should do so slowly when the measured jitter is less that the current buffer play-out delay.

Where such adaptive de-jitter buffers are used, they dynamically adjust to the maximum value of network jitter. In this case, the jitter buffer does not add delay in addition to the worst-case end-to-end network delay.

VoIP: Impact of Loss

Packet Loss Concealment (**PLC**) **is a technique used to mask the effects of lost or discarded VoIP packets**. The method of packet loss concealment used depends upon the type of codec used.

A simple method of packet loss concealment, used by waveform codecs like **G.711**, is to replay the previously received sample; the concept underlying this approach is that, except for rapidly changing sections, the speech signal is locally stationary. This technique can be effective at concealing the loss of up to approximately **20 ms** of samples.

Low bit rate frame-based codecs, such as **G.729** and **G.723**, use more sophisticated PLC techniques, which can conceal up to **30–40 ms** of loss with "tolerable" quality, when the available history used for the interpolation is still relevant.

Hence, to summarize the impact that packet loss has on VoIP, with an appropriately selected packetization interval (**20–30 ms** depending upon the type of codec used) **a loss period of one packet may be concealed but a loss period of two or more consecutive packets may result in a noticeable degradation of voice quality**.

VoIP: Impact of Loss

Possible causes of packet loss:

- *Congestion;*
- *Lower layer errors;*
- *Network element failures;*
- *Loss in the application end-systems.*

Therefore, in practice, networks supporting VoIP should typically be designed for very close to zero percent VoIP packet loss. QOS mechanisms, admission control techniques and appropriate capacity planning techniques are deployed to ensure that no packets are lost due to congestion with the only actual packet loss being due to layer 1 bit errors or network element failures. Where packet loss occurs, the impact of the loss should be reduced to acceptable levels using PLC techniques.

VoIP: Impact of Throughput

VoIP codecs generally produce a constant bit rate stream; that is, unless silence suppression is used. **Silence suppression**, which is also known as **voice activation detection** (**VAD**), **prevents the transmission of packets carrying "silent" samples**. Silence suppression becomes active when it detects periods of silence from the microphone that exceed defined thresholds; when silence suppression is active it prevents the encoder output from being sent to the far end. When silence suppression is active for a leg of a VoIP call, the bandwidth used for that leg of the call is almost zero. As most conversational speech contains approximately 50% silence, this can significantly reduce the average bandwidth used for a call; however, the peak bandwidth used for the call remains unchanged.

Networks supporting VoIP should typically be designed for very close to zero percent VoIP packet loss, and hence are designed to be congestionless from the perspective of the VoIP traffic. **This means that the available capacity for VoIP traffic must be able to cope with the peak of the offered VoIP traffic load**. This peak load must be able to be supported without loss while maintaining the required delay and jitter bounds for the VoIP traffic. But even if VoIP capacity is provisioned to support the peak load, the VoIP service may be statistically oversubscribed.

VoIP: Impact of Packet Re-ordering

VoIP traffic is not commonly impacted by packet re-ordering, as the magnitude of re-ordering would need to be very significant to affect a VoIP flow whose inter-packet gap is a multiple of 20 ms, for example. It is, however, noted that in addition to the impact that it has on application throughput, per-packet load balancing, which is a common cause of packet re-ordering, can also increase the jitter that is experienced within a flow due to the different delays of alternate paths; this effect can impact VoIP services.

Video. Video Streaming

With video streaming applications, a client requests to receive a video that is stored on a server; the server streams the video to the client, which starts to play out the video before all of the video stream data has been received. Video streaming is used both for "**broadcasting**" video channels, **which is often delivered as IP multicast**, and for **video on demand** (**VOD**), **which is delivered as IP unicast**.

IP-based streaming video is most commonly transported as a data stream encoded using standards defined by the **Motion Picture Expert Group** (**MPEG**) and transported using **RTP over UDP**. MPEG defines the encoding used for the actual video stream, while [**RFC2250**, **RFC 2343**, and **RFC3640**] define how real-time audio and video data are formatted for RTP transport. RTP is the transport layer protocol, which deals with the delivery of that stream from sender to receiver. Protocols such as the **Real-time Streaming Protocol** (**RTSP**) [**RFC2326**] may be used to set up the RTP streams.

Video Streaming

An MPEG encoder converts and compresses a video signal into a series of pictures or frames; as there is generally only a small amount of change between one frame and the next it is possible to compress the video signal significantly by transmitting only the differences. Three different **types of MPEG frames**:

- **"I"-frames**. Intra or "I"-frames carry a complete video frame and are coded without reference to other frames. An I-frame may use spatial compression; spatial compression makes use of the fact that pixels within a single frame are related to their neighbors. Therefore, by removing spatial redundancy, the size of the encoded frame can be reduced and prediction can be used in the decoder to reconstruct the frame. A received I-frame provides the reference point for decoding a received MPEG stream.
- **"P"-frames**. Predictive coded or "P"-frames are coded using motion compensation (temporal compression) by predicting the frame to be coded from a previous "reference" I-frame or P-frame. P-frames can provide increased compression compared to I-frames with a P-frame typically 10–30% the size of an associated I-frame.
- **"B"-frames**. Bidirectional or "B"-frames use the previous and next I- or B-frames as their reference points for motion compensation. B-frames provide further compression, still with a B-frame typically 5–15% the size of an associated I-frame.

Video Streaming

Frames are arranged into a Group of Pictures or GOP. Unlike with VoIP where codec implementations are very specifically defined, with streaming video there is significant scope for variation in the specific way that an MPEG stream may be encoded, even for a single type of encoding. **The specific GOP structure used to encode a video stream can have a major impact on the effect that network loss, latency and throughput have on the video reproduction at the receiver**.

Video Streaming: Impact of Delay

For video streaming, the important delay metric is the **one-way end-to-end delay from streaming server to client**. The main constraint that end-to-end network delay and jitter have on streaming video is on end-user "interactivity," or the "finger-to-eye" delay.

Broadcast Video Services

Broadcast television services delivered over IP (also known as **IPTV**) commonly use ***IP multicast***. Assuming a broadcast video service being delivered using IP multicast to a receiver – which could be a **set-top box** (**STB**) for example – where each channel is a separate multicast group, the overall channel change time is made up of a number of components:

- Remote control and STB processing.
- Network transmission delay.
- Multicast processing.
- Network transmission delay.
- STB Buffering/processing (De-jitter buffer; FEC or real-time retransmission delay; Decryption delay; MPEG decoder buffer; IBB frame delay).

Broadcast Video Services

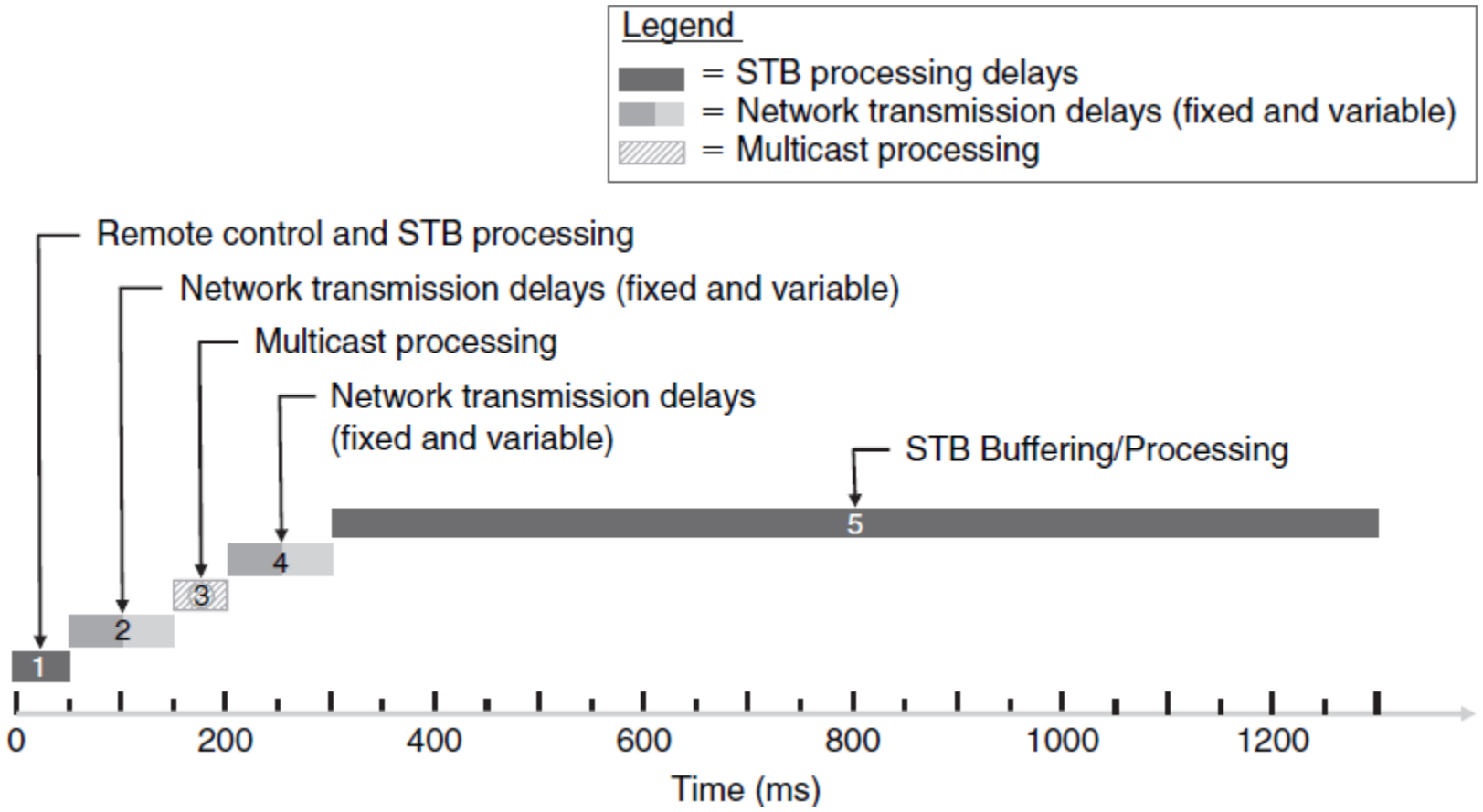

Figure 9 Broadcast video channel change time delay components (example)

Video-on-demand Services

Video-on-demand (**VOD**) and network **personal video recorder** (**PVR**) services are commonly delivered as **unicast**. For VOD services the end-to-end delay impacts the finger-to-eye delay, i.e. the response time it takes for user requests to be translated into actions visible to the end-user; for example, how long it takes after pressing play for a VOD to start. Typically, response times of approximately **1 second** are targeted.

Assuming a video-on-demand service being delivered over IP unicast to a receiver, which could be a set-top box (STB) for example, the overall response time is made up of a number of components:

- Remote control and STB processing.
- Network transmission delay.
- Middleware processing.
- Network transmission delay.
- STB buffering/processing (de-jitter buffer; FEC or real-time retransmission delay; decryption delay; MPEG decoder buffer).

Video-on-demand Services

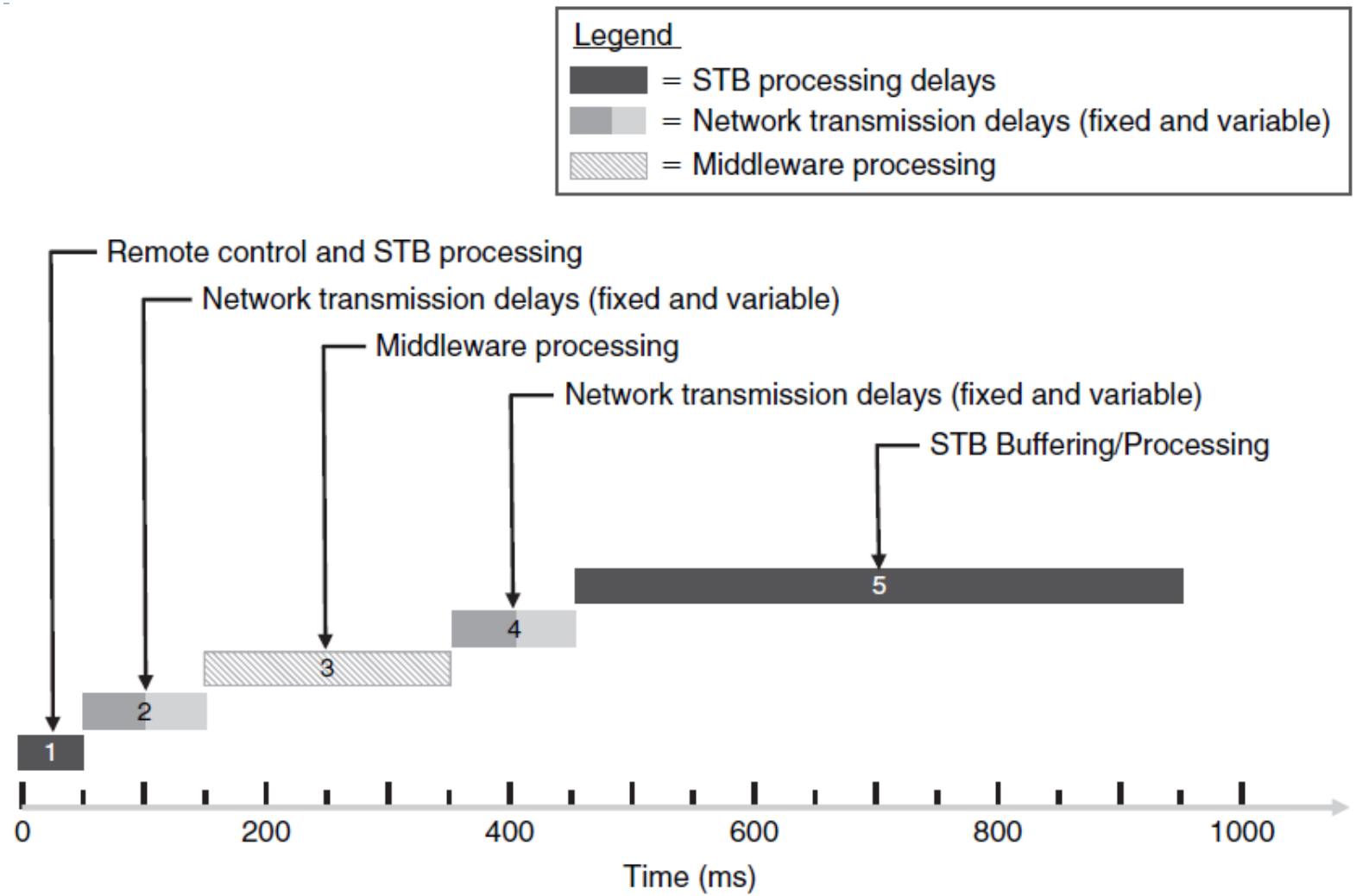

Figure 10 VOD response time delay components (example)

Video Streaming

Video Streaming: Impact of Delay-jitter

Digital video decoders used in streaming video receivers need to receive a synchronous stream, typically with jitter tolerances of only 500 ns, in order to decode without visible impairments. Such jitter tolerances are not achievable natively in IP networks, hence as for VoIP, broadcast video services use **de-jitter buffers** (also known as **play-out buffers**) in receivers to remove delay variation caused by the network and turn variable network delays into constant delays such that the tolerances required by the decoder can be met.

Video Streaming: Impact of Loss

Causes of packet loss:

- Congestion.
- Lower layer errors. There are two main techniques for loss concealment for streaming video:
  - ✓Forward error correction (FEC).
  - ✓Real-time retransmission.
- Network element failures.
- Loss in the application end-systems.

Therefore, in practice, networks supporting video streaming services should typically be designed for very close to zero percent video packet loss.

Video Streaming

Video Streaming: Impact of Throughput

The bandwidth requirements for a video stream depend upon the video format, the encoder and the specific GOP structure. There are four main video formats used for IP-based video services:

1. Standard definition (SD).
2. High definition (HD).
3. Common interchange format (CIF) – low definition (LD) format.
4. Quarter CIF (QCIF).

MPEG allows for streaming video to be encoded either as variable bit rate streams, where the quality of the resultant video is constant, or as constant bit rate streams where the quality of the resultant video is variable. The table in **Figure 11** gives indicative average bit rates for LD, SD and HD video stream rates using MPEG-2 and MPEG-4 AVC.

Video Streaming

Format	MPEG-2	MPEG-4 AVC
LD QCIF	100–200 kbps	50–100 kbps
LD CIF	0–5–1 Mbps	0.25–0.5 Mbps
SD 4CIF	~3–4 mbps	~2–3 Mbps
HD	~15–20 Mbps	~10–15 Mbps

Figure 11 Typical broadcast quality video stream IP rates

Video Streaming: Impact of Packet Re-ordering

Many real-time video end-systems do not support the re-ordering of received frames, hence packet re-ordering effectively results in higher packet loss and should be avoided.

Video Conferencing

Video conferencing sessions are typically set up using the signaling protocols specified in ITU recommendation **H.323** or **SIP**. Whichever method is used to establish the connections, from an SLA perspective, the fundamental requirements and principles remain the same.

The audio streams will typically use codecs such as those defined by the ITU **G.71x**/**G72x** standards.

The video formats and encoding used for video conferencing applications are less constrained than for broadcast quality video services. Codecs such as **MPEG-2/H.262** or **MPEG-4 AVC/H.264** are typically used; where bandwidth is constrained, lower definition (e.g. CIF or QCIF) and lower frame rates (e.g. 10 fps), potentially reduce the bandwidths required significantly compared to broadcast video services.

As for discrete voice and video services, in practice networks supporting video conferencing services should typically be designed for very close to zero percent packet loss for both the VoIP and video streams.

Data Applications

QOE requirements for data application, which in turn drive network level SLAs, are less well defined than for voice or video applications. While there are multiple types of data applications that exist, from a QOS perspective they can be broadly divided into **interactive data applications** and **applications that are targeted at data transfer with no requirements on interactivity**.

Throughput focussed applications in general use **TCP** as their transport layer protocol, due to the reliability and flow control capabilities that it provides.

Interactive applications depend on providing responses to an end-user in real-time. As the specific implementations of interactive data applications can vary, the impact that network characteristics such as delay have on them can also vary.
For **client/server applications** which require a network transaction, network delay is but one aspect of the total transactional delay, which may be comprised of the following components:

- Client-side processing delays.
- Server-side processing delays.
- Network delays.

Interactive Data Applications

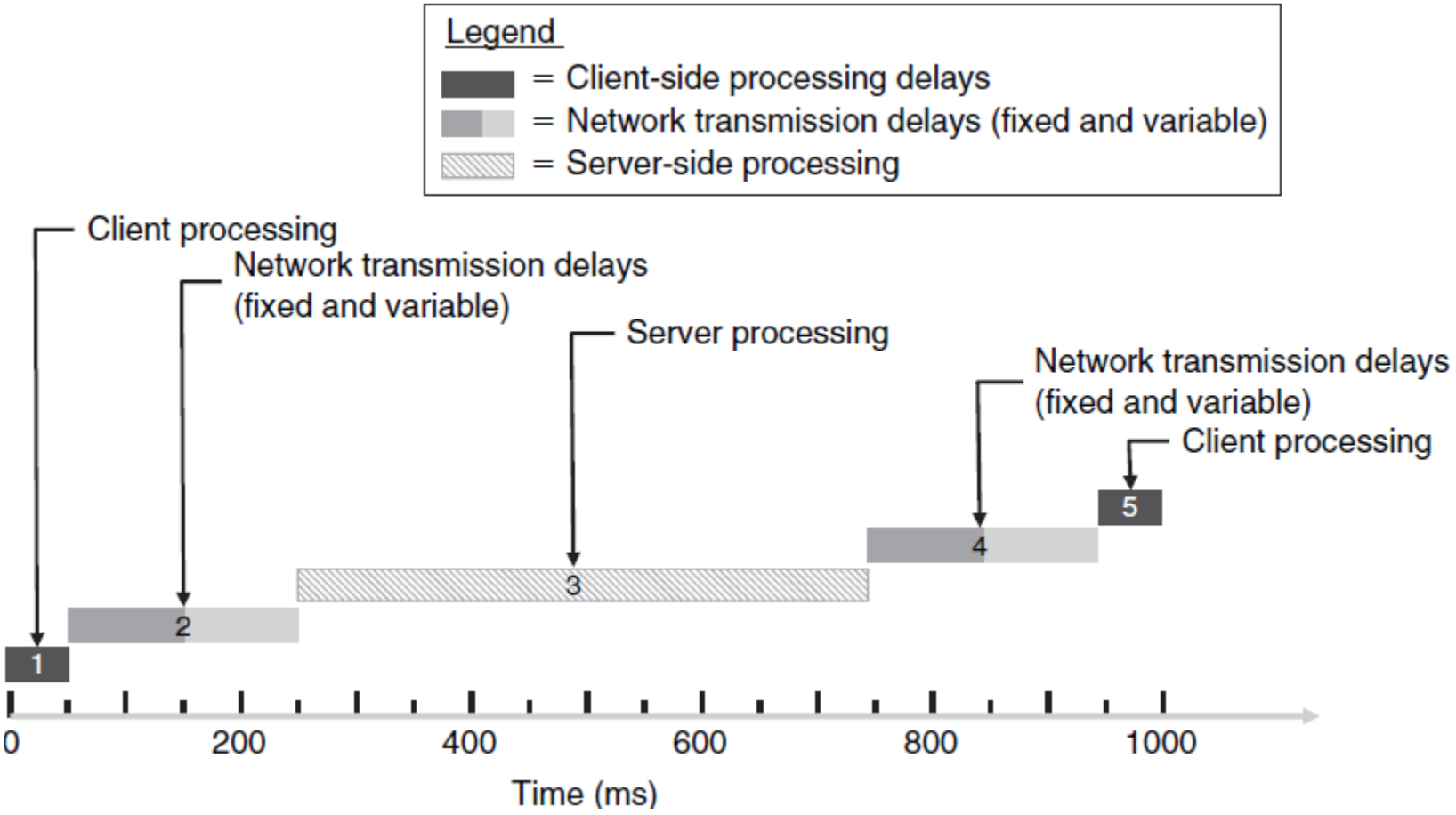

Figure 12 Delay components: example interactive data application #1

Interactive Data Applications

Another example: an application with the same total client-side, and server-side processing delays, but which instead required two network transactions (a DNS query and an HTTP GET for example) per user transaction, a network RTT of approximately 200 ms or less would be required in order to meet the target.

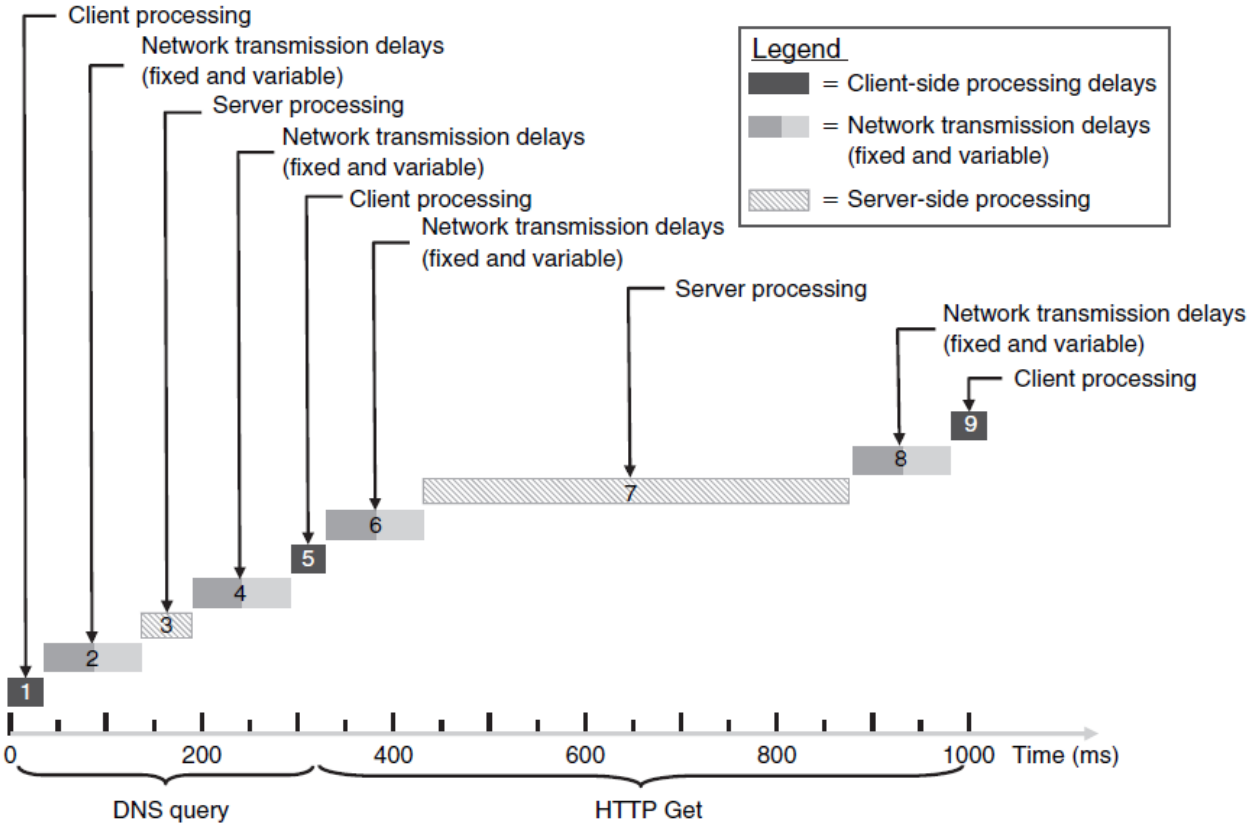

Figure 13 Delay components: example interactive data application #2

Interactive Data Applications

Jitter has no explicit impact on interactive data applications; jitter only has an impact on TCP in that it is a component of network delay. Network loss and packet re-ordering can have an impact on interactive data applications in that lost or re-ordered packets may need to be retransmitted which may probabilistically increase the network component of the total transaction delay. The impact of packet loss and resequencing will depend upon the characteristics of the transport layer protocol that is used.

For **UDP-based interactive data applications**, a detailed knowledge of the specific application implementation is required in order to understand the impact of packet loss and resequencing; this would require analysis on an application-by-application basis.

On-line Gaming

Multiplayer on-line or networked games are the most popular form of a type of application known as Networked Virtual Environments (NVEs); other uses of NVEs include **military simulation**. **Users in NVEs, who may be in geographically separate locations, interact with each other in a virtual world in real-time**. The IEEE **Distributed Interactive Simulation** (**DIS**) [**IEEE1278**] standard covers NVE; however, this is not generally used by the software vendors that produce on-line games who instead use proprietary implementations.

Although there are different types of real-time on-line games – the most common game types being: **First Person Shooter** (**FPS**), **Real-Time Strategy** (**RTS**) and **Multiplayer On-line Role-Playing Game** (**MORPG**) – most use a client-server architecture, where a central server tracks client state and hence is responsible for maintaining the state of the virtual environment. The players' computers are clients, unicasting location and action state information to the server, which then distributes the information to the other clients participating in the game. Most implementations use UDP as a transport protocol.

On-line Gaming

Most on-line gaming implementations have evolved to work over the public Internet and have bandwidth requirements of less than 64 kbps and in-built mechanisms to deal with packet loss.

However, it is noted that these bandwidth requirements may increase over time, with the prevalence of higher bandwidths available to end-users due to broadband access. In addition, some games provide the capability to tweak various network parameters, which can have a significant impact on their bandwidth requirements.

It is commonly cited that **low network delay is a requirement of on-line gaming applications**; players who experience higher delays to/from the server than others may experience a relative “lag” in play as they receive information from the server later than lower delay users, and similarly the server receives information from them later than from the lower delay user. Consequently, **users with lower RTTs (Round-Trip Times) may have a game-playing advantage**.