Transmitting and Receiving Real-Time Data

The invention is in the field of handling of time-sensitive data over packet switched networks, and more particularly transmitting and receiving video data over the Internet.

The invention relates to a method of providing a streaming video service to a client across a packet network whilst reducing the start-up delay usually associated with preparing a buffer of data while maintaining the use of a buffer. The invention also relates to a method of controlling the transmission rate of the streaming video to adapt to congestion in the network.

Traditionally the Internet has supported traffic such as FTP, e-mail and web-surfing, where the overall delay does not intrinsically detract from the final presentation of the media. The advent of faster processing multimedia PC's has driven the delivery of multimedia, including video, over the Internet. Time-sensitive applications however require continuous, quality of service guaranteed, high bandwidth data channels, which is seemingly at odds with the packet-based nature of the Internet and has the potential to disrupt transmissions with unacceptable packet jitter, i.e. the variation in the inter-arrival times of packets caused by variable routing and changeability of delivery rates owing to congestion. Currently, commercial streaming technologies overcome jitter by constructing a large buffer (5-30 seconds) before starting to playback video material. This start-up delay is non-optimal for a user, who may have to wait for this period, before realising that the content requested is incorrect; and generally detracts from the users experience of the multimedia presentation.

According to a first aspect of the present invention there is provided a method of operating a real-time communication apparatus comprising a real-time data sender, a real-time data display device having a store and a network connecting said sender and said display device, said method comprising the steps of:

operating said sender to transmit a plurality first-encoding-rate data packets representing a first part of a real-time presentation to said display device, said transmission rate being higher than said encoding rate;

operating said display device to:

receive said first-encoding-rate data packets into said store;

remove first-encoding-rate data packets from said store at said first encoding rate for decoding to present said real-time presentation to said user at a first level of quality;

on said store being filled with said first-encoding-rate data to a predetermined level, sending an indication that said level has been reached to said sender;

operating said sender, on receipt of said indication, to send second-encoding-rate data packets representing subsequent parts of said real-time presentation to said display device, said second encoding rate being higher than said first encoding rate;

operating said display device to:

receive second-encoding-rate data packets representing a subsequent part of real-time presentation into said store;

remove second-encoding-rate data packets from said store at said second encoding rate for decoding to present said real-time presentation to said user at a second level of quality higher than said first level of quality.

According to another aspect of the present invention there is provided a method of presenting time-sensitive data at a client while constructing a buffer of time-sensitive data, said method comprising receiving time-sensitive data which has been transmitted to said client, passing said time-sensitive data to a data buffer; and, monitoring the quantity of time-sensitive data in the data buffer; reading said time dependent data out of the data buffer to be processed for viewing; wherein the method is characterised in that the rate at which the time-sensitive data is read out of the data buffer is lower than the rate at which the time-sensitive data is passed to the data buffer; and the time-sensitive data is read out of the data buffer when it

arrives in the data buffer, such that there is substantially no delay between the client receiving the time-sensitive data and making the time-sensitive data available; and, presenting the time-sensitive data.

There will come a point when the data buffer becomes sufficiently full. The rate of transmission can then be reduced to equal the rate of consumption by the viewing means which will bring the quantity of data in the buffer to an equilibrium. However, in this situation the bandwidth of the connection may not be employed to full capacity.

In a further aspect of the present invention there is provided a method of presenting time-sensitive data at a client, wherein, time-sensitive data encoded at a first bit-rate is received until a pre-determined quantity of data fills the data buffer, whereupon time-sensitive data encoded at a second bit-rate is received, wherein said second bit-rate is higher than the first bit-rate.

A still further aspect of the present invention provides a method of providing time-sensitive data to a client is taught, wherein time-sensitive data encoded at a first bit-rate is read from a first data buffer at a first transmission rate to be transmitted to the client; and, upon request, time-sensitive data encoded at a second bit-rate is read from a second data buffer at a second rate.

It is desirable to use as much of the available bandwidth of a link as possible to transmit data because with a higher bit-rate of video data comes better quality reproduction. However, loss of data in the network causes severe degradation of service – far outstripping the benefits of increased bit-rate. For example, with predictive coding schemes such as H.263 and MPEG, receiving half of a 500 $kbits^{-1}$ video stream is likely to give a much worse quality than all of a 250 $kbits^{-1}$ stream. It is therefore important to reduce transmission rate in a controlled way, rather than letting data be lost to the network. The Internet protocol TCP has a built-in control mechanism whereby the data transmission rate is steadily increased until packet loss is detected, whereupon the data rate is reduced. The data rate is then increased again until packet loss reoccurs. A variable transmission rate is said to be elastic and applications which are able to control the transmission rate of data in response to network conditions are said to be TCP-friendly. It is desirable to provide video data in

a TCP-friendly way so that the as much of the bandwidth available at any particular time is utilised. A further benefit of TCP-friendly data delivery is that congestion in the network is managed as individual applications themselves reduce data rates until each has a fair share of the bandwidth.

Standard compression technologies, such as MPEG4 or H.263 can be managed to exhibit TCP-friendly behaviour, see for example the applicant's co-pending patent application number GB 9928023.2. This solution, however, requires a high-speed, dedicated PC per video stream. Transcoding an encoded data stream from a high bit-rate to a low bit-rate when network congestion is detected also suffers from the problem of being computationally demanding. Another approach is to use layering of video streams, whereby quality adaption is achieved by adding or dropping layers of the video stream. The disadvantage of this method is that it is inefficient, as a certain proportion of the available bandwidth must be allocated to instructions for integrating the layers.

The present invention further provides a method wherein the rate at which time-sensitive data is read out from first and second buffers may be dynamically varied in dependence upon the condition of a link to the client, and further, time-sensitive data encoded at a first bit-rate is read from a first data buffer at a first transmission rate to be transmitted to the client; or, time-sensitive data encoded at a second bit-rate is read from a second data buffer at a second rate, in dependence upon the condition of a link to the client, wherein said first bit-rate is lower than the second bit-rate.

Embodiments of the invention will now be described, by way of example only, with reference to the figures, where:

Figure 1 is a schematic overview of the relationship between encoder, video streamer and clients;

Figure 2 shows the arrangement of the video streamer;

Figure 3 shows the arrangement of a client; and

Figure 4 shows the stepwise operation of one embodiment of the present invention.

As shown in Figure 1, a first embodiment of the present invention consists of a source of compressed video data, encoder 1, which encodes data both at a low bit-rate R_L, which may have a value of for example 500 $kbits^{-1}$, and a high bit-rate R_H, of

for example 1500 $kbits^{-1}$. The compression codec used is H.263 but equally may be any other codec, such as MPEG4. Encoder 1 takes 'live' video data as its input, for instance a broadcast of a sporting event.

The two encoded data streams are transmitted via separate logical connections to the video streamer 2 at a transmission rate T_E. The video streamer 2 may be on the same premises as the encoder 1 and linked via an intranet. The video streamer 2 runs on a server computer, for instance one comprising a Pentium III 700 MHz, 256MB RAM which has access to the Internet.

A video viewer, hitherto referred to as the client, running on a PC (a,b,c etc in Figure 1) suitably configured to have access to the Internet, may connect to the video streamer 2 via the Internet and thus the client is able to access content. A suitable PC terminal is a 266 MHz Pentium II laptop PC. The video streamer 2 can support a large number of clients (typically up to 1000) viewing the same stream.

For a live broadcast, the encoder 1 will transmit at a transmission rate T_E which is real-time. The two streams of data R_L and R_H coded at different bit-rates offer different quality video reproduction, but each data stream has the same transmission rate, T_E. The data must be decoded at this rate for the program to play back in real-time.

Figure 2 shows the arrangement of the video streamer 2. Low quality encoded video data encoded at a low bit-rate R_L and high quality encoded video data encoded at a high bit-rate R_H from the encoder 2 is received at the input connections 21 and 22 respectively and fed to buffers 23 and 24 respectively. It should be noted that there is provided one buffer per channel of encoded video data that is received by the video streamer 2. Encoded video data is read out from each buffer 23, 24 via a switch 26 which selects which encoded video data stream is to be sent to the output connection 27. There is provided a buffer manager 25 which is capable of controlling the rate at which data is read out from each of the buffers 23, 24 and thus defines the transmission rate T_S of the video streamer 2. The buffer manager is also in connection with the switch 26 and is further capable of receiving signals from connection 28. T_S is selected by varying the time delay between the transmission of each packet, such that T_S may be less than, equal to or greater than the encoder transmission rate T_E. Those skilled in the art will realise that the limiting factor on the sustainability of transmission where $T_S > T_E$ is the size of the buffer 23, 24 such that

a buffer of size S kbits will be able to sustain a transmission rate of $T_S = 2T_E$ for twice as long as a buffer of size S/2 kbits. Through the control of both switch 26 and the transmission rate T_S the buffer manager is able to control the bit-rate which is output from the video streamer 2 on two scales; by adjusting the transmission rate T_S fine control of the bit-rate is achieved, and by switching between the two encoded data streams encoded at bit-rates R_L and R_H control of the bit-rate on a coarse scale may be achieved. The buffer manager 25 makes adjustments to T_S or switches the output between buffers in response to signals received from connection 28.

Figure 3 shows the arrangement of the client running on a PC 3a, b, c etc. The encoded video data that is sent from the video streamer 2 is received at the client via a connection 27 and checked for completeness by a packet loss detector 31. The data is then sent into a client buffer 32 which is of a size suitable to absorb fluctuations in network throughput. The client buffer 32 is connected directly to a decoder 33 and from there decoded data is sent to be displayed at the client screen (not shown). A client status monitor 34 is connected to the packet loss detector 31 and client buffer 32. The client status monitor 34 is able to send signals via connection 28.

The packet loss detector 31 monitors incoming packets. If packet loss is detected then a signal is sent to the client status monitor 34, which is informs the buffer manager at the video streamer 2 via connection 28. Missing packets can be retransmitted. The buffer manager 25 steadily increases the transmission rate T_S until a consistent pattern of packet loss occurs, indicating that the maximum bandwidth is being utilised. In the interest of maintaining a congestion free network, the transmission rate T_S may then be exponentially reduced. The client status monitor 34 monitors the volume of data in the client buffer 32 such that a signal is sent via connection 28 to the buffer manager 25 at the video streamer 2 when the client buffer 32 becomes sufficiently full of data.

The system of video streamer 2 and client 3 as described above allows user-friendly video streaming, i.e. the client buffer 32 enables the quality of the video to be despite variations in network conditions, which might otherwise have a detrimental effect on the overall perceived quality of the media.

The operation of the present embodiment of the invention will now be described with reference to Figure 4.

The video streamer 2 is initialised, which involves filling the buffers 23, 24 with a quantity of data from the encoder 1. For a live broadcast, data is constantly fed into the buffers 23, 24 and is subsequently discarded after an amount of time defined by the size of the buffer and the quality of data being received.

A PC running browser software to browse web pages on the Internet may be used to select a link to, for example, a live broadcast on a site hosted by the entity providing streamed video. Being interested in viewing the particular clip or broadcast, the user clicks (selects) the link. The browsing software detects that streamed video data has been requested and launches the video viewing client software which embodies the client 3. The client 3 issues a "send data" command via connection 28 to the buffer manager 25, which sets switch 26 to read encoded video data from the low bit-rate data buffer 23 and requests a transmission rate of $T_S = 2T_E$. The data is transmitted to the data connection 27 and thence to the client 3. Using the example encoding bit-rate cited above of 500 $kbits^{-1}$ for R_L , data flows into the network to the client at a rate of 1000 $kbits^{-1}$.

The client 3 receives the encoded video data and sends it via the packet loss detector 31 to the client buffer 32 which is supplied at the rate $2T_E$. When data is detected in the buffer 32 the encoded video data is promptly read out to the decoder 33 at a rate of T_E. Therefore the buffer 32 fills at a rate T_E while the decoded data from the decoder 33 is displayed. Thus the user is provided with video pictures without having to wait for the client buffer 32 to fill.

The client monitor 34 waits for the quantity of R_L data in the client buffer 32 to reach a specified level, upon which a "switch buffer" command is sent to the buffer manager 25 at the video streamer 2 via the connection 28. The buffer manager 25 then switches the flow of data from the low bit-rate data buffer 23 to the high bit-rate data buffer 24 and instructs transmission at a rate $T_S = T_E$. Using the example encoding rate cited above, data is transmitted on the network at 1500 $kbits^{-1}$.

The client buffer 32 will then begin filling with high quality data which will be placed behind the low quality data. After a length of time the R_H data will begin to be read into the decoder 33, whereupon the user will perceive an increase in the picture quality. At this point, the client 3 has a full buffer and the user is watching images of a quality which is consistent with the capacity of the network link.

The video streamer 2 can support a number of clients (typically 1000). Each client is initially given a unique read-out point for the start-up phase, whereupon, after equilibrium of the client buffer 32 has been reached and the video streamer 2 is supplying high bit-rate data from the buffer 24, the read-out point can be amalgamated with other client read-out points. Read-out points may have to be devolved as discrepancies in network capacity demand increasing or decreasing the transmission rate for a particular client.

The skilled person will appreciate that the low bit-rate data buffer 23 should be of a size which will allow data to be read from it at a rate $2T_E$ for a period of time which is long enough to provide the client buffer 32 with a suitable quantity of data. For example, in order to buffer 5 seconds worth of 500 $kbits^{-1}$ data at the client 3, the video streamer 2 must supply 1000 kbits of data for 5 seconds, 500 kbits of which will be consumed by the decoder 33 per second and 500 kbits will build up in the buffer per second until 5 seconds has elapsed. Therefore the low bit-rate data buffer must be able to hold at least 5Mbits of data (5x1000kbits), or just over 0.5Mb.

The skilled person will appreciate that there are problems associated with 'tapping into' a stream of encoded data when data is initially read out of a buffer. The compression technology typically employed by the encoder 1 involved coding a frame of video data, termed an anchor frame or an I-frame and from this frame an estimate is made as to what the next frame will look like, this estimated frame being termed a B-frame. In this way the quantity of data representing a series of frames may be greatly reduced. However, if the first frame to be read from either of the data buffers 23, 24 is a B-frame then the first few frames of decoded data may be unintelligible as the decoder tries to reconstruct frames based on an estimate. In a further embodiment of the invention, an extra buffer of data is supplied in parallel with the data buffers 23, 24 consisting solely of I-frames. The first frame to be transmitted is read from the I-frame buffer and thus gives the decoder a reliable point from which to start decoding. Data is then switched to be read from either of the data buffers 23, 24.

The system allows user-friendly video streaming, i.e. the quality of the video does not fluctuate rapidly as network conditions vary, which can have a detrimental affect on the overall perceived quality of the media. In the event of packet loss being reported by the client, the system can exponentially reduce its transmission rate. This need

not result in an immediate switching of the video source, as there may be data buffered at the client. Immediately after the packet loss it is possible that the transmission rate is lower than the encoding rate, and the client is supplementing received data with buffered data in order to meet the demands of the video decoder, with the result that the client's buffer is emptying. In the event of isolated packet loss, the system can again ramp up the transmission rate, initially slowing the rate at which the client's buffer is emptying before eventually returning to a state of filling it.

The skilled person will appreciate that the ability to transmit data at variable rates for a period of time enables the streamed data to be elastic and allows TCP-friendly transmission. Detection of sustained packet loss by the packet loss detector 31 is indicative of network congestion. The buffer manager 25 at the video streamer 2 reacts to notification of packet loss by instructing a reduction in the transmission rate of data from the high bit-rate data buffer 24. The high bit-rate data buffer 24 should be appropriately sized to cope with such an event. If packet loss persists at the reduced transmission rate for longer than the high bit-rate data buffer can sustain, then the buffer manager 25 will switch to supply data from the low bit-rate data buffer 23. Effective management protocols are necessary to prevent rapid switching between data buffers 23 and 24 as the data capacity of the network fluctuates, because this will cause changes in the perceived quality of the played back video. While a user will tolerate low quality playback, rapid changes in quality can be irritating to a user.

There is no limit to the number of encoded data streams that may be provided to the video streamer. Maximum bandwidth utilisation may be achieved thus: starting by reading data from a low bit-rate data buffer, the transmission rate is increased. Finding that no packets are lost at this transmission rate, the output is switched to a higher bit-rate data buffer, whereupon the transmission rate is increased. If this transmission rate encounters no obstacle then a higher still bit-rate data buffer can be switched in, and so on until the maximum bandwidth is employed.

The buffer manager 25 located at the video streamer 2 is enabled to decide how to adjust the transmission rate T_S and when to switch buffers. Equally, instructions may be sent from the client 3 to the video streamer 2 about transmission rate T_S and which buffer to feed data from. The location of the buffer manager 25 in the

embodiments described has been chosen because it is practical to situate the control centre close to the centre which is responsible for charging for the service, which in this case is the ISP.

The example of video data is chosen as an example of multimedia data to illustrate the above embodiments. The invention is equally suited to any other form of time-sensitive data, such as audio data or a multimedia presentation.

In the embodiment described above, data is supplied by the encoder 1. Equally, compressed video data may be held in a library of program data files, for example a library of feature films, which may be accessed when required.

The video streamer 2 may be remote from the encoder 1, such that the video streamer 2 and the encoder 1 are connected via the Internet. It is likely that the video streamer 2 would be operated by an Internet Service Provider (ISP) and remote connection of the video streamer 2 and encoder 1, would allow the ISP to make content available to the client from many encoders.

CLAIMS

1. A method of operating a real-time communication apparatus comprising a real-time data sender, a real-time data display device having a store and a network connecting said sender and said display device, said method comprising the steps of:

operating said sender to transmit a plurality first-encoding-rate data packets representing a first part of a real-time presentation to said display device, said transmission rate being higher than said encoding rate;

operating said display device to:

receive said first-encoding-rate data packets into said store;

remove first-encoding-rate data packets from said store at said first encoding rate for decoding to present said real-time presentation to said user at a first level of quality;

on said store being filled with said first-encoding-rate data to a predetermined level, sending an indication that said level has been reached to said sender;

operating said sender, on receipt of said indication, to send second-encoding-rate data packets representing subsequent parts of said real-time presentation to said display device, said second encoding rate being higher than said first encoding rate;

operating said display device to:

receive second-encoding-rate data packets representing a subsequent part of real-time presentation into said store;

remove second-encoding-rate data packets from said store at said second encoding rate for decoding to present said real-time presentation to said user at a second level of quality higher than said first level of quality.

2. A method of presenting time-sensitive data and constructing a buffer of time-sensitive data, said method comprising

receiving time-sensitive data;

reading said time-sensitive data into a data buffer; and,

reading said time-sensitive data out of the data buffer for presentation; and,

presenting the time-sensitive data;

wherein the method is characterised in that

the rate at which the time-sensitive data is read out of the data buffer is lower than the rate at which the time-sensitive data is read into the data buffer; and,

reading the time-sensitive data out of the data buffer is initiated when the time-sensitive data first arrives at the data buffer, such that there is substantially no delay between the client first receiving the time-sensitive data and making the time-sensitive data available for presentation.

3. A method of presenting time-sensitive data at a client according to claim 2, including the step of,

monitoring the quantity of time-sensitive data in the data buffer;

wherein the bit-rate of the received time-sensitive data is dependent on the quantity of data in the data buffer.

4. A method of presenting time-sensitive data at a client according to claim 3, wherein,

time-sensitive data encoded at a first bit-rate is received until the quantity of data in the data buffer reaches an upper threshold, whereupon time-sensitive data encoded at a second bit-rate is received, wherein said first bit-rate is lower than the second bit-rate.

5. A method of presenting time-sensitive data at a client according to claim 4, wherein,

time-sensitive data encoded at the second bit-rate is received until the quantity of data in the data buffer reaches a lower threshold, whereupon time-sensitive data encoded at the first bit-rate is received.

6. A method of presenting time-sensitive data at a client according to any one of the preceding claims, wherein,

the rate at which the time-sensitive data is read into the data buffer is dependent on the quantity of data in the data buffer.

7. A method of providing time-sensitive data to a client, wherein

time-sensitive data is encoded at a plurality of bit-rates;

wherein data encoded at one of the plurality of bit-rates is selected for transmission to the client;

transmitting the data to the client.

8. A method of providing time-sensitive data to a client according to claim 7, wherein,

the rate at which time-sensitive data selected for transmission to the client is transmitted to the client may be dynamically varied.

9. A method of providing time-sensitive data to a client, according to claim 7, wherein,

the time-sensitive data encoded at one of the plurality of bit-rates is selected in dependence upon a received signal.

10. A method of providing time-sensitive data to a client according to claim 8, wherein,

the rate at which the time-sensitive data selected for transmission is transmitted is dependent upon a received signal.

11. A method of providing time-sensitive data to a client according to claims 9 or 10, wherein,

the received signal is representative of the condition of the link to the client.

12. A method of providing time-sensitive data to a client according to claims 9 or 10, wherein,

the received signal is representative of the quantity of data in a data buffer at the client.

13. A method of providing time-sensitive data to a client according to claim 11, wherein,

the condition of the link to the client is determined by the quantity of data in a data buffer at the client.

14. A method of providing time-sensitive data to a client in accordance with any one of the preceding claims, wherein the time-sensitive data represents video data or may represent audio data.

15. Apparatus for receiving time-sensitive data, said apparatus comprising,

buffer means for storing received time-sensitive data for a period of time,

means for reading said time-sensitive data into the buffer means,

means for reading said time dependent data out of said buffer means,

monitoring means for monitoring the quantity of data held in the buffer means; and

means for transmitting data relating to the quantity of time dependent data held in said buffer means,

said apparatus being characterised in that means for initiating reading the time-sensitive data out of the data buffer when the monitoring means detects the arrival of time-sensitive data at the data buffer, such that there is substantially no delay between the client first receiving the time-sensitive data and the time-sensitive data being read out of the buffer.

16. Apparatus for supplying encoded video data to at least one client, said apparatus comprising,

receiving means to receive a plurality of data streams, wherein each of the plurality of data streams comprises video data encoded from the same source with differing quantisation parameters,

a plurality of buffer means for storing each stream of encoded video data for a period of time,

at least one reading means for reading encoded video data from one of the plurality of buffer means,

buffer management means arranged to control the at least one reading means,

characterised in that,

said reading means is arranged to read, under control of the buffer management means, encoded video data from any of the buffer means at a frame rate which may be less than, equal to or greater than the frame rate at which the video data is encoded.

17. Apparatus for supplying encoded video data to a client as claimed in claim 16, including,

means for receiving data relating to the quantity of encoded video data in the buffer of the client, and

means for determining the current capacity of the link to the client,

wherein the buffer management means is arranged to control the reading means in dependence on the quantity of encoded video data in the buffer of the client and on the capacity of the link to the client.