Mechanisms of MPEG Stream Synchronization

G J Lu, H K Pung and T S Chua
Department of Information Systems and Computer Science
National University of Singapore
Singapore 0511

Abstract

Media synchronization is an important issue in developing multimedia applications. MPEG is an international standard for coding moving pictures and associated audio for multimedia applications. Coded audio, video and other data streams are multiplexed into an MPEG stream. We introduce the syntax of the multiplexed MPEG stream and explain the mechanisms used to maintain media synchronization in a hypothetical model, system target decoder, in which it is assumed that data transfer and decoding are carried out instantaneously. Then we propose extensions to these mechnisms to achieve MPEG stream synchronization in a practical decoder.

1. Introduction

MPEG (Motion Picture Expert Group) is an international standard for coding of moving pictures and associated audio for digital storage media up to about 1.5 Mbit/s, prepared by ISO/IEC JTC 1/SC 29/WG11[1]. The draft standard [1] was published as ISO/IEC 11172. When referring to this standard, MPEG and ISO 11172 are used interchangeably. The standard consists of two parts. Part 1 is divided into three sections. Section 1, Systems, specifies the system coding layer. It defines a multiplexed structure for combining elementary streams, including coded audio, video and other data streams, and specifies means of representing the timing information needed to replay synchronized sequences in real-time. Section 2, Video, specifies the coded representation of video data and the decoding process required to reconstruct pictures. Section 3, Audio, specifies the coded representation of audio data. Part 2 , conformance testing, specifies the procedures for determining the characteristics of coded bit streams and for testing compliance with the requirements stated in Part 1.

MPEG Video is a very important aspect of the ISO 11172 standard and has been discussed by Gall [2]. This paper looks into the MPEG Systems, particularly the aspects related to media synchronization. Since coded MPEG stream consists of continuous media streams, i.e. video and audio, it is necessary to synchronize these media for real-time playback. There are two aspects of continuous media synchronization: intra-medium synchronization and inter-media

synchronization. While intra-medium synchronization ensures the continuity for smooth playback of each medium, intermedia synchronization ensures synchronization between associated media. Thus in this paper media synchronization is defined as occurence in which each medium is played out at its fixed rate determined by the type of medium and/or the application concerned and the specified/required temporal relationships among the associated media are maintained.

The ISO 11172 standard specifies syntax and semantics based on a conceptual decoder model, system target decoder (STD). This model assumes that the multiplexed MPEG stream is stored on a constant latency digital storage medium, and data transfer and decoding within the decoder are instantaneous. ISO 11172 stream is designed such that the STD will be able to decode and display elementary streams synchronously.

However, in a practical decoder, data transfer and decoding cannot take place instantaneously. Since ISO 11172 stream is specified based on STD model, it is the responsibility of decoder to compensate for these data transfer and decoding delays to ensure synchronization.

In ISO 11172, the storage medium has a broad meaning, including CD-ROM, magnetic harddisk, digital audio tape and computer networks etc. These storage media are of indeterministic nature in terms of delay and transmission rate instead of constant latency as assumed in STD model. In order to feed the ISO11172 decoder with a stream of constant latency, a medium specific decoder is required. An appropriate buffer must be used in the medium specific decoder to smooth out the transmission jitter in order to provide data to the ISO11172 decoder at the required rate.

The rest of the paper is organized as follows. Section 2 discusses the MPEG Systems in general. Section 3 presents the principle and mechanisms used in the ISO11172 Systems to support media synchronization in STD. Section 4 discusses compensation of decoding delays in practical decoders required in order to maintain synchronization. Section 5 discusses buffer requirement to overcome the bursty nature of digital storage medium.

2. ISO 11172 Stream

While MPEG Video and Audio specify the coding of each individual stream, MPEG Systems specifies the syntax and semantics of information that is necessary in order to reproduce one or more MPEG audio or video compressed data streams in a system. An ISO 11172 stream consists of one or more elementary streams multiplexed together. A elementary stream consists of a number of access units (AU). In the case of compressed audio an access unit is defined as the smallest part of the encoded bitstream which can be decoded by itself. In the case of compressed video an access unit is the coded representation of a picture. A decoded audio access unit or decoded picture is called a presentation unit (PU). In a coded video stream, there are three types of access units: I-pictures, P-pictures and B-pictures. I-pictures are coded without referring to

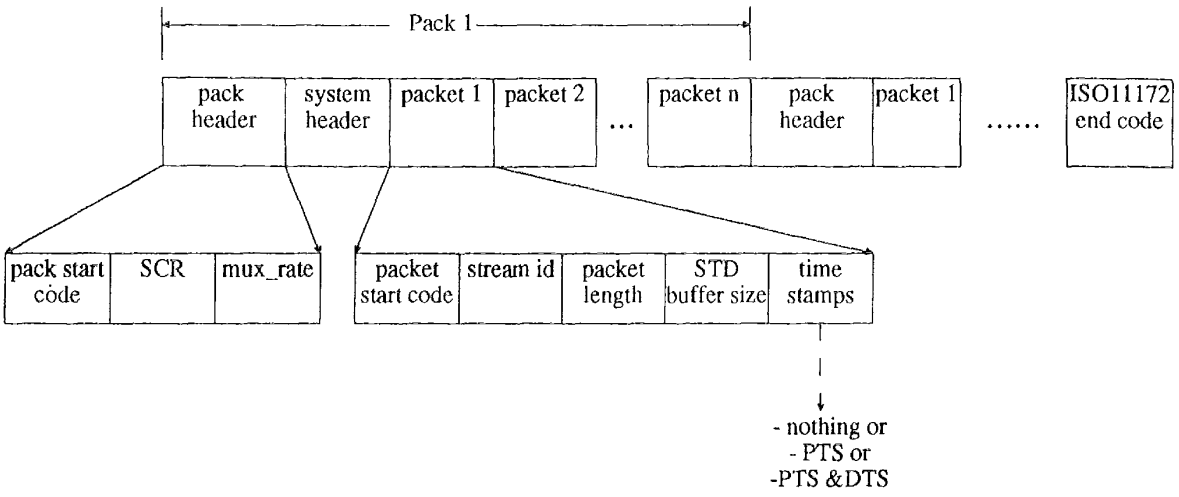

Fig.1 ISO11172 Stream Syntax

other pictures. P-pictures are coded using forward prediction and B-pictures are coded using both forward and backward predictions. Due to the way video stream is coded, picture order in the coded stream may differ from the display order. The decoder must carry out re-ordering if needed [1, 2].

An ISO 11172 stream is organized into two layers: the pack layer and the packet layer. The pack layer is for system operations and the packet layer is for stream specific operations. Fig.1 shows the syntax of an ISO 11172 stream.

An ISO 11172 stream consists of one or more packs. A pack commences with a pack header and is followed by zero or more packets. The pack header begins with a 32-bit start-code. The pack header is used to store system clock reference SCR and bitrate information, mux_rate. The SCR is a 33-bit number, indicating the intended time of arrival of the last byte of the SCR field at the input of the system target decoder. Mux_rate is a positive integer specifying the rate at which the system target decoder receives the ISO 11172 multiplexed stream during the pack in which it is included. The value of mux_rate is measured in units of 50 bytes/second rounded upwards. The value zero is forbidden. The value encoded in mux_rate field may vary from pack to pack in an ISO 11172 multiplexed stream. The mux_rate value together with SCR value defines the arrival time of each byte at the input to the system target decoder.

Data from elementary streams is stored in packets. A packet consists of a packet header followed by packet data. The packet header begins with a 32-bit start code that also identifies the stream to which the packet data belongs. The packet header defines the buffer size required at each elementary decoder for smooth decoding and playback of the elementary stream. The packet header may also contain decoding and/or presentation time-stamps (DTS and PTS) that refer to the first access unit in the packet. The purposes of DTS and PTS are discussed in next section.

The packet data contains a variable number of contiguous bytes from the same elementary stream. A data packet never contains data from more than one elementary stream and byte ordering is preserved. Thus, after removing the packet headers, packet data from all packets with a common stream identifiers are concatenated to recover a single elementary stream. The multiplex of different elementary streams is constructed in such a way (in terms of packet size and the relative placement of packets from different streams) as to ensure that the specified STD buffers do not overflow or underflow.

The system header is a special packet that contains no elementary stream data. Instead it indicates decoding requirements for each of the elementary streams. It indicates a number of limits that apply to the entire ISO 11172 stream, such as data rate, the number of audio and video streams, and the STD buffer size limits for the individual elementary streams. A decoding system may use these limits to establish its ability to play the stream. The system header also indicates whether the stream is encoded for constant rate delivery to the STD. The system header must be the first pack of the ISO 11172 stream. It may be repeated within the stream as often as necessary. Repeat of the system header will facilitate random access. Real-time encoding systems must calculate suitable limits for the values in the header before starting to encode. Non-real-time encoders may make two passes over the data to find suitable values.

Up to 32 ISO 11172 audio and 16 ISO 11172 video streams may be multiplexed simultaneously. Two private data streams of different types are provided. One type is completely private and the other follows the same syntax as audio and video streams. It may contain stuffing bytes, a buffer size field, and PTS and DTS fields. The use of these fields is not specified in ISO11172.

The system specification does not specify the architecture or implementation of encoder or decoders. However, bitstream properties do impose functional and performance requirements on encoders and decoders.

The MPEG Systems coding specification provides data fields and semantic constraints on the data stream to support the necessary system functions. These include the synchronized presentation of decoded information, the management of buffers for coded data, and random access. Random access is made possible by repeated appearance of the information needed to start decoding, such as SCR, PTS and system headers, and use of I-pictures (pictures coded without referring to other pictures). Other functions are all related to the smooth and synchrony playback of coded streams and are discussed in the next section.

3. Synchronization in the System Target Decoder

An ISO 11172 stream is constructed such that elementary streams will be synchronously decoded and presented by the STD and elementary input buffers in STD will never overflow or underflow. This section explains the mechnisms used to achieve this.

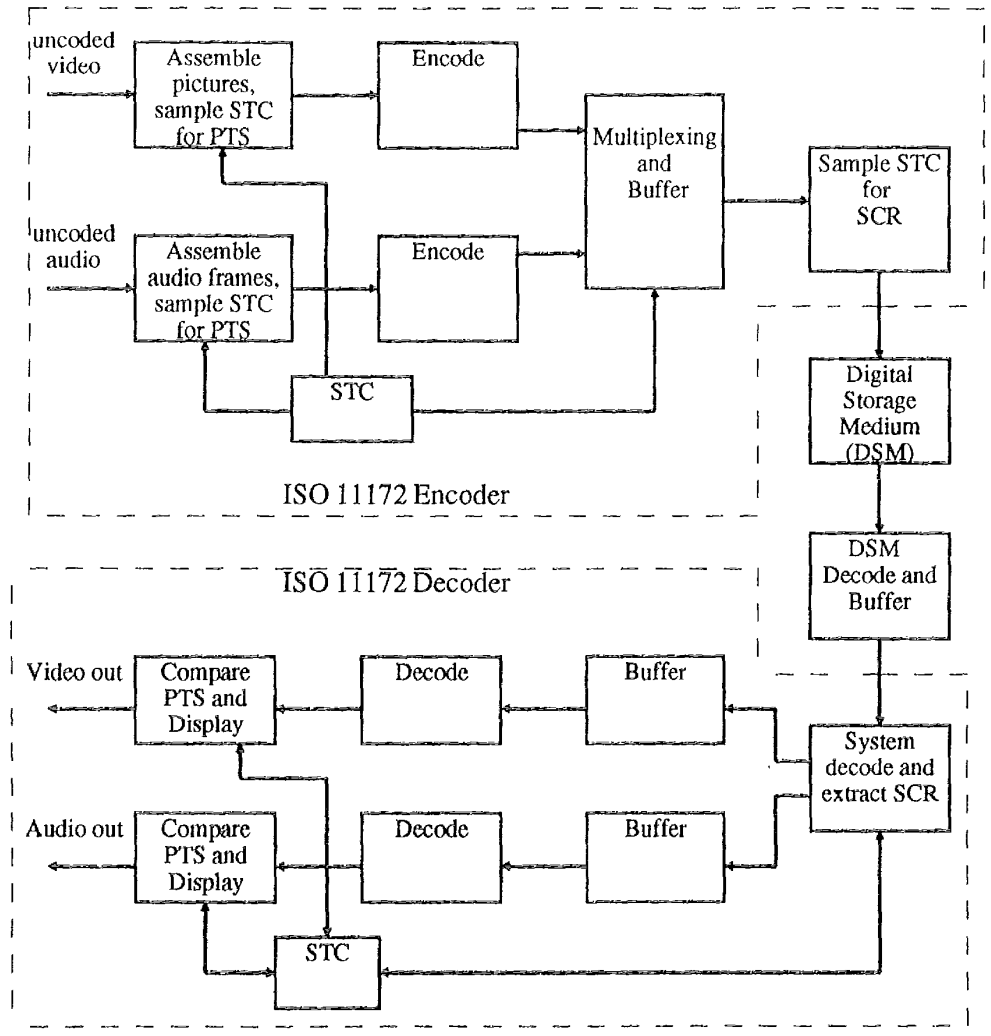

Fig.2 Protypical Encoder and Decoder

In STD, playback of N streams is synchronized by adjusting the playback of all streams to a master time base rather than by adjusting the playback of one stream to match that of another. The master time base may be one of the N decoders' clock, the DSM or the channel clock, or it may be some external clock. The similar synchronization principle has been used in a number of synchronization schemes, such as synchronization marker[3], logical time system [4] and relative time system [5]. In these scheme, each presentation unit has a time stamp and presentation units with the same time stamps are displayed at the same time to achieve synchronization.

MPEG Systems provide for end-to-end synchronization of complete encoding and decoding process. This is achieved by use of time stamps, including system clock reference (SCR), presentation time stamp (PTS), decoding time stamp (DTS). This end-to-end synchronization is illustrated in Fig.2, which includes a protypical encoder (upper part) and a protypical decoder

(lower part). These protypical encoding and decoding systems are not normative, they illustrate the functions expected of real systems.

In the protypical encoding system, there is a single system time clock (STC) which is available to the audio and video encoders. Audio samples entering the audio encoder are organized into audio present units. Some, but not necessary all, of the audio PUs have PTS values associated with them, witch are samples of the STC at the time the first sample of the PU is input to the encoder. Likewise, the STC values at the times when video pictures enter the video encoder are used to create video PTS fields. SCR values represent the time when the last byte of the SCR field leaves the encoder. DTSs specify the time the access units are decoded, that is the time at which all the bytes of an access unit are removed from the buffer of an elementary stream decoder in the STD model. The STD model assumes instantaneous decoding of access units. In audio streams, and for B-pictures in video streams, the decoding time is the same as the presentation time and so only the PTSs are encoded; DTS values are implied. In video streams, for I-pictures and P-pictures the DTS values are nominally equal to the PTS values minus the number of picture periods of video reordering delay multiplied by the picture period.

In the protypical decoder system, the ISO 11172 stream arrives according to the arrival schedule specified by SCR and mux_rate fields in the pack header. The first SCR value is extracted and used to initialize the STC in the decoder. The correct timing of the STC is maintained by ensuring that STC is equal to the subsequent SCR values at the time the SCRs are received. The STC may be maintained either by updating the STC with the value of the SCRs or via a control loop, using the SCR values as reference inputs. Elementary access units are decoded at times specified by their DTSs and PUs are presented when their PTS values are equal to the STC value. In this way both intra-stream and inter-stream synchronization is maintained. Intra-stream synchronization is maintained by ensuring the STCs at the encoder and the decoder running at the same rate. Inter-stream synchronization is maintained by present each PU at their specified PTS relative to STC. So long as each PU is presented at its specified time, the inter-stream temporal relationship is maintained.

As mentioned in the last section, SCR is encoded in the pack headers. PTS and /or DTS are encoded in the packet headers. The PTS and /or DTS are associated to the first access unit commencing in the packet. More than one access units can commence in a packet. It is not necessary for each packet to contain PTS and /or DTS. PTS and DTS fields are not necessarily encoded for each picture or audio PU. They are required to occur with intervals not exceeding 0.7 seconds. This bound allows the construction of a control loop using the PTS values which has guaranteed stability with a known bandwidth. For those PUs for which PTS is not encoded, the decoder can approximate the correct value as the sum of the most recent PTS and an increment. The increment is the nominal number of system_clock_frequency cycles per PU times the number of PUs since the last PTS.

The value of the system clock frequency is measured in Hz and shall meet the following constraints:

90 000 - 4.5 <= system_clock_frequency <= 90 000 + 4.5

rate of change of system_clock_frequency with time <=250 * 10^{-6} Hz/s

All timing information (SCR, PTS, DTS) is specified in terms of a 90 kHz clock, which provides sufficient accuracy for audio interchannel phase alignment. 90 kHz is a multiple of the various video frame rates under consideration and is also a submultiple of the CCIR 601 video sampling frequency of 13.5 MHz. The time stamps are encoded into 33 bits which are long enough to support absolute program durations of at least 24 hours.

So far, we have discussed about synchronization and time stamps. Buffer management is also an important aspect of synchronization because if there is data starvation or buffer overflow, synchronization will not be maintained. In ISO 11172, it is specified that for all multiplexed streams the delay caused by system target decoder input buffering shall be less than or equal to one second. The input buffering delay is the difference in time between a byte entering the input buffer and when it is decoded. This STD buffering delay, together with elementary stream bitrate, bounds the size of STD buffers. These buffer size values are specified in the system header and packet header. The STD model precisely specifies the times at which each data byte enters and leaves the buffer of each elementary stream decoder in terms of a common system time clock. It is guaranteed that the input buffers with specified sizes will not overflow or underflow during decoding as long as the data stream conforms to the specification, and the complete decoding system is synchronized in terms of SCR, DTS and PTS.

To Summarize, MPEG Systems specifies syntax and semantics of multiplexed stream based on STD model such that STD can decode and present elementary streams synchronously.

4. Compensation of Actual Decoding Delays

In the last section, we discussed synchronization provision for STD. But practical system does not conform to STD model, because it takes time to transfer data from buffer and decode the data. This section looks at how to compensate for these delays.

In a practical decoder, presentation time and decode time are not the same any more even for audio PUs and B-pictures, because in practical decoding system decoding takes some time instead of zero time as assumed in STD model. Furthermore, it takes different time to decode access units of different types of elementary streams. To compensate these differences, time stamps PTS and DTS have to be modified in a practical decoder. Assume there are two elementary streams, one audio and one video, in an ISO 11172 stream. Let the decoding time for an audio access unit

be t_a and decoding time for a video access unit t_v. Also assume t_v is greater or equal to t_a. Then DTS and PTS should be adjusted in the decoder as follows:

(1) In the ISO 11172 stream, DTSs are implied for audio access units and B-pictures. To make this explicit, a DTS is inserted immediately after each PTS with the same value. These DTSs indicate commencing times access units are being decoded. For those access units without DTSs associated, the decoding commencing time is equal to the most recent value of DTS plus time corresponding to access units between them.

(2) To compensate for the decoding delay for video access units, the effective presentation time should be equal to the value of associated PTS plus decoding time t_v.

(3) For audio access units, if we choose the effective presentation time to be the value of associated PTS plus decoding time t_a, the audio presentation units will be presented t_v-t_a second ahead of video stream. This is not desirable. Therefore, in order to maintain the desired temporal relationship between audio and video, the effective presentation time for audio access units should be equal to the value of associated PTS plus t_v. Decoded access units will be buffered at the output buffer until their effective presentation times. The required output buffer is equal to the audio display rate times (t_v-t_a).

Above discussion is concerned with decoding time and presentation time. Since it takes time to decode an access unit, the data corresponding to an access unit is not removed from the elementary stream input buffer instantaneously. Therefore, in order to avoid buffer overflow, an additional buffer size equivalent to the largest access unit in the elementary stream should be added to the specified STD buffer size for each elementary stream.

If presentation and decoding times and buffer sizes are adjusted as discussed above, a practical decoder system will be synchronized and the buffer will not overflow or underflow, provided that ISO 11172 stream data are retrieved from a constant latency storage medium. However, practical digital storage media, such as harddisk and computer network, cannot deliver constant latency data stream. Therefore it is necessary to have a buffer in the storage medium specific decoder to smooth the burstiness of the delivered data stream.

5. Channel Smoothing and Storage Medium Requirements

An ISO stream can be coded in constant bitrate or variable bitrate. A bit, fixed_flag, in the system header indicates which mode is used. If its value is set to "1" fixed bitrate operation is indicated. If its value is set to "0" variable bitrate operation is indicated. To simplify the discussion in this section, a constant bitrate stream is assumed, although the principle also applies to the variable bitrate stream.

Under the above assumption, an ISO11172 decoder consumes data at a constant rate. However, digital storage media, such as harddisk and computer networks, are bursty in nature. Therefore, a buffer in the storage medium specific decoder has to be used to smooth the burstiness of data stream from the storage medium. The problem which we are interested in is to find what is the buffer size requirement and what requirements should be imposed on the storage medium so that the storage medium specific decoder can output a constant latency constant rate data stream.

Lets first examine requrements on average data arriving rate at the storage medium specific decoder and on the storage medium transmission bandwidth. We introduce data arriving function A(t) and data output function C(t). A(t) indicates the amount of data arrived at the storage medium specific decoder within time interval 0 to t. C(t) indicates the amount of data output from the medium specific decoder within time interval 0 to t. Both A(t) and C(t) are non-decreasing function. C(t) increases with time at a fixed output rate equal to the ISO 11172 decoder consume rate. A(t) will normally not increase at a fixed rate due to burstiness of the arriving rate caused by delay jitter. Assume the first byte of data arrives at the client at time t1 and the storage medium specific decoder

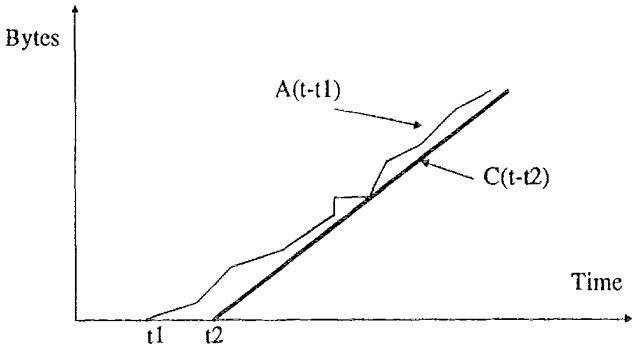

Fig.3 A(t-t1) is close to C(t-t2), buffer requirement is small

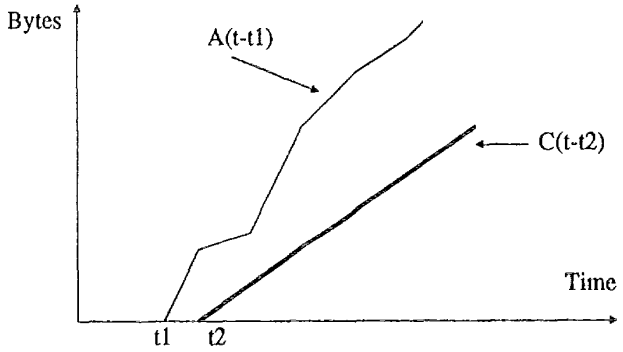

Fig.4 Average arriving rate is larger than the output rate, buffer requirement becomes larger and larger with time

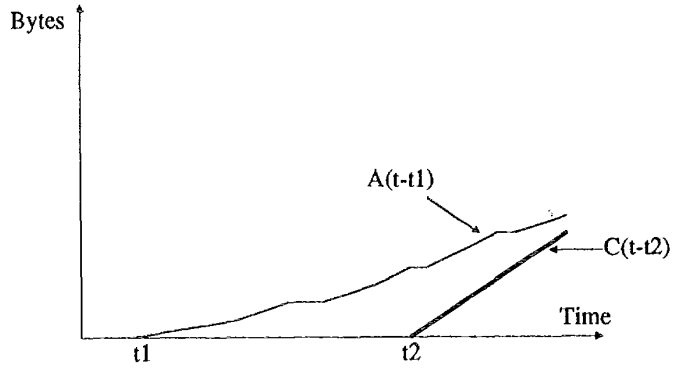

Fig.5 Average arriving rate is smaller than the output rate, there is a long initial delay and large buffer requirement

outputs the first byte of data at time t2, then we have arriving function A(t-t1) and output function C(t-t2), as shown in Fig.3.

In order to meet the continuity requirement, A(t-t1) must be equal to or greater than C(t-t2). The difference, A(t-t1)-C(t-t2), is buffered and represents the buffer occupancy. The slope of A(t-t1) represents the data arriving rate. The average value of the arriving rate should be equal or close to the output rate. If average arriving rate is greater than the output rate, the difference A(t-t1) and C(t-t2), i.e. the buffer occupancy will become larger and larger (Fig.4). This means that in order to successfully play out a stream, either a very large buffer size is required or playout of the stream can only sustain for a short time. This is not desirable in a practical system.

If the average arriving rate is smaller than the output rate, to satisfy the requirement A(t-t1)-C(t-t2)>=0, t2 must be large (Fig.5). This means that the initial delay from the time the first byte arrives to the time the first packet is displayed is very long. This also means that a very large initial buffer is required. The longer the stream to be played, the longer the initial delay and the larger the buffer requirement. All these, long delay and large buffer size requirement, are not desired and practical. Therefore the average arriving rate should be matched closely to the output rate in order to output the data stream at constant rate continuously. To achieve this, the storage transmission bandwidth at least equal to the consume rate must be guaranteed.

Now lets look at the buffer size requirement in the storage medium specific decoder in order to provide a constant latency stream to the ISO 11172 decoder. Suppose a byte can experience delay ranging from the minimum d_{min} to the maximum d_{max} from the data source to the storage medium specific decoder. Then, if a byte experiences a delay of d is buffered for a time d_{max} - d, all bytes will experience the same total delay equal to d_{max} relative to the ISO 11172 decoder. In this case, latency for each byte will be constant. The maximum required buffering time is d_{max} - d_{min}, which is the delay jitter. Therefore the buffer size requirement is equal to the arriving data rate times the channel delay jitter. It should be noted that the larger the delay jitter, the larger the buffer size requirement. To limit the buffer size required at the storage medium specific decoder, delay jitter should be small.

Above discussion shows that a buffer size equal to bitrate times channel delay jitter is required to smooth the channel and a channel bandwidth at least equal to the output rate must be guaranteed in order to play out a MPEG stream successfully.

6. Summary

MPEG Systems defines a multiplexed structure for combining coded audio, video and other data streams, and specifies means of representing the timing information needed to replay synchronized sequences in real-time. The specification is based on a system target decoder model.

We explained how media synchronization is achieved in the system target decoder. Then we discussed the required modification of some time stamps in a practical ISO 11172 decoder and discussed channel smoothing buffer requirement in the storage medium specific decoder.

References

1. Draft International Standard ISO/IEC DIS 11172, 1992.

2. Didier Le Gall, "MPEG: A video Compression Standard for Multimedia Applications", Communications of the ACM, Vol.34, No.4, April 1991, pp.46-58.

3. D. Shepherd and M Salmony, "Extending OSI to support synchronization required by multimedia applications", Computer Communications, Vol. 13, No. 7, September 1990, pp.399-406.

4. D. P. Anderson and G. Homsy, "A continuous Media I/O Server and Its Synchronization Mechanism", Computer Journal, October 1991, pp.51-57.

5. P. V. Rangan, S. Ramanathan, H. M. Vin and T. Kaeppner, "Techniques for Multimedia Synchronization in Network File Systems", Computer Communications, March 1993.