System-Level Issues for a Multi-Rate Video VOD System

Sheila S. Hemami
Interactive Video Initiative Laboratory
HPL-95-22
February, 1995

video-on-demand,
multi-rate video,
rate negotiation

This document describes system-level issues for incorporation of multi-rate video into a video-on-demand system. The issues are broadly separated into two classes: *rate negotiation*, involving selection of a rate at which the requested video will be provided, and *rate delivery*, involving delivery of the coded stream to the user. Requirements of the server, the network, and the set-top box are developed, and are shown to be interrelated for design of an efficient system. Additionally, constraints are imposed on the video coding algorithm. Specific research topics for further investigation are proposed.

Internal Accession Date Only

System-level Issues for A Multi-Rate Video VOD System

Sheila S. Hemami
hemami@hpl.hp.com
(415) 857-5039
Hewlett Packard Laboratories, Palo Alto

Abstract

This document describes system-level issues for incorporation of multi-rate video into a video-on-demand system. The issues are broadly separated into two classes: *rate negotiation*, involving selection of a rate at which the requested video will be provided, and *rate delivery*, involving delivery of the coded stream to the user. Requirements of the server, the network, and the set-top box are developed, and are shown to be interrelated for design of an efficient system. Additionally, constraints are imposed on the video coding algorithm. Specific research topics for further investigation are proposed.

1.0 Introduction

Current video-on-demand (VOD) systems in the trial phase are homogeneous — all set-tops are identical, the video is delivered to each set-top via the same physical medium, and each video or program is only coded at one rate. However, heterogeneity exists in the viewing device, as there are many sizes and qualities of televisions, and even computers with NTSC cards can display video. Consumers may value the ability to select a quality level for their VOD programs, possibly paying more for higher quality video. In future systems, heterogeneity will extend to the network, and different physical media will provide varying bandwidths. Some users may have fiber to the home, while others may only have a copper link. (For a review of possible community network technologies, see [1].) A VOD system should have the flexibility to service users of both types.

In both examples given above, a VOD system that can provide the same content at different bit rates (and hence different qualities), depending on the user requirements and network capabilities, is desirable. The simplest solution is to separately encode and store each video sequence at each

In both examples given above, a VOD system that can provide the same content at different bit rates (and hence different qualities), depending on the user requirements and network capabilities, is desirable. The simplest solution is to separately encode and store each video sequence at each desired rate. While providing the desired rate selection, this solution quickly becomes unaffordable at the video server due to disk storage requirements. At a rate granularity of 2, each sequence must be coded and stored twice. At a much finer rate granularity of 10 or 20, the storage requirements are too large, and the disk I/O bandwidth utilization decreases because only a fraction of the rates will be in use at any one time.

Therefore, a more efficient technique of providing video at multiple rates without requiring separately coded sequences is desirable. This capability is provided by so called *multi-rate coding techniques*. A multi-rate technique produces one bitstream for a video sequence, from which can be extracted a coded sequence at various bit rates. Such a stream is called a *multi-rate stream* or a *scalable stream*. The terms multi-rate stream and scalable stream will be used interchangeably.

This document examines system-level issues involving incorporation of multi-rate video into a VOD system. The system issues are examined for each of the three components in the system: the *user*, the *network* or *network provider*, and the *video server* or *service provider*. The user is defined as either the consumer viewing the video, or the set-top box that decodes the video. The definition will be stated when it is not clear from the context. The network provider is the entity providing the delivery system, which itself is the network. For the purposes of this document, the placement of the Level 1 and Level 2 gateways [2], [3] inside or outside of the network is irrelevant. The service provider is the entity providing the content, which is stored on and read from the video server. The video server includes its own content directory, which can be accessed by users when selecting a program and is called the Level 2 gateway (L2GW).

Inclusion of the multi-rate capability can be driven by the user, the service provider, or the network provider. From the user's perspective, selection of a video includes selection of a rate, which is billed appropriately, with higher rate (and subsequently higher quality) video costing more. From the service provider's perspective, server throughput can be increased by modifying current rates of videos being delivered. Rates can be incrementally decreased such that the server can accommodate more users, while visual quality does not suffer. Furthermore, when the server has no more capacity for high-rate video, lower-rate video can still be offered. Such an increase in throughput and users corresponds to an increase in revenue. The same rationale applies to the network provider's perspective, but in this case the capacity and throughput refer to that of the network, rather than the server.

Besides the obvious challenge of defining a coding algorithm that provides a scalable stream, incorporation of the multi-rate capability into a complete VOD system provides many technical issues that must be addressed. In this document, these issues are broadly separated into two classes: *rate negotiation* and *rate delivery*. Rate negotiation is defined as the formation of an agreement between the user, server, and network on the rate of the video that will be provided. The user's role may only be implicit in this negotiation. For example, a set-top box may be con-

figured to receive only the lowest-rate video. In this case, the consumer has no direct input at the time of selection, but the server and network must still be informed of the rate. Rate delivery is defined as the process by which a coded stream is read from the server, transmitted over the network, and delivered to the set-top box. In the following, rate negotiation and rate delivery are examined from the perspectives of the user, network provider, and service provider.

The organization of this document is as follows. Section 2 first reviews current and proposed multi-rate coding techniques for video, and then discusses rate granularity for a multi-rate VOD system. Rate negotiation is discussed in Section 3, where the key idea of maintaining a downstream flow of rate information is introduced. Section 4 presents rate delivery issues, which are concerned predominantly with the server, and to a lesser extent with network processing and storage. Finally, Section 5 presents conclusions and outlines research topics for further investigation.

2.0 Multi-Rate Video Coding

A multi-rate video coding technique produces a scalable stream, from which multiple streams so different bit rates can be extracted. These streams can vary in spatial resolution, temporal resolution, picture quality, or any combination of the three. This section reviews current multi-rate capabilities of MPEG and several multi-resolution techniques proposed in the literature. Finally, the granularity of rate offerings for a VOD system is discussed.

2.1 MPEG-2 Multi-Rate Capabilities

To date, the only non-proprietary digital video coding standard that exists is MPEG [4], and as a result its use is widespread. The second generation of the standard, known as MPEG-2, roughly incorporates the concept of scalability. Video is coded in a mandatory *base layer* and in optional *enhancement layers*, of which there can be two, and hence a maximum of 3 bit rates per coded sequence is possible. Scalability is provided by extra information in the enhancement layer(s) to increase the spatial resolution (called *spatial scalability*), to increase the temporal resolution (called *temporal scalability*), or to increase the quality of the coded video (called *SNR scalability*). Because the standard was developed to be backwardly compatible with the earlier MPEG-1, the incorporation of scalability was limited by the compatibility. As a result, the quality of an MPEG-2 stream coded with 2 Mb/s in the base layer and 2 Mb/s in the enhancement layer is less than the quality had the stream been coded with 4 Mb/s in the base layer alone and no enhancement layer.

SNR scalability is achieved through use of finer-grain quantization. Because MPEG is inherently a block-based coding technique, temporal and spatial scalability are achieved through prediction and subsequent coding of the resulting difference information. However, other source coding techniques exist that explicitly provide streams at different spatial and temporal resolutions without resorting to predictive coding. Two such coding techniques use pyramid and subband decompositions, and are discussed in the next section.

2.2 Multi-Resolution Video Coding Techniques

Image and video coding techniques known generically as multi-resolution techniques are better suited to providing spatial scalability because they operate on the entire image rather than on individual blocks, and decompose visual information into a frequency-based hierarchy in two dimensions. In three dimensions, temporal information is separated into high- and low- frequency information as well. The frequency separation also includes subsampling steps, and the result is an image or video sequence that can be represented in increasing spatial resolution, with corresponding higher spatial frequency resolution, and also in increasing temporal resolution with corresponding higher temporal frequency resolution.

One such example of video sequence coding using a pyramid decomposition is presented in [5]. A video sequence is first filtered and subsampled both spatially and temporally in multiple stages, forming a so-called pyramid. The coded information consists of the low resolution sequence, and the motion interpolation and difference information required to interpolate the low resolution sequence back to full size and frame-rate. Lower-rate sequences are provided by the low-resolution sequences, and rates corresponding to any intermediately upsampled and interpolated sequences before full size and full frame-rate are achieved. Reconstruction of the pyramid is illustrated in Figure 1. Motion estimation is performed on a block basis (block size 8×8). The resulting coding technique provides as many different rates as there are levels in the pyramid (i.e., stages of filtering and subsampling).

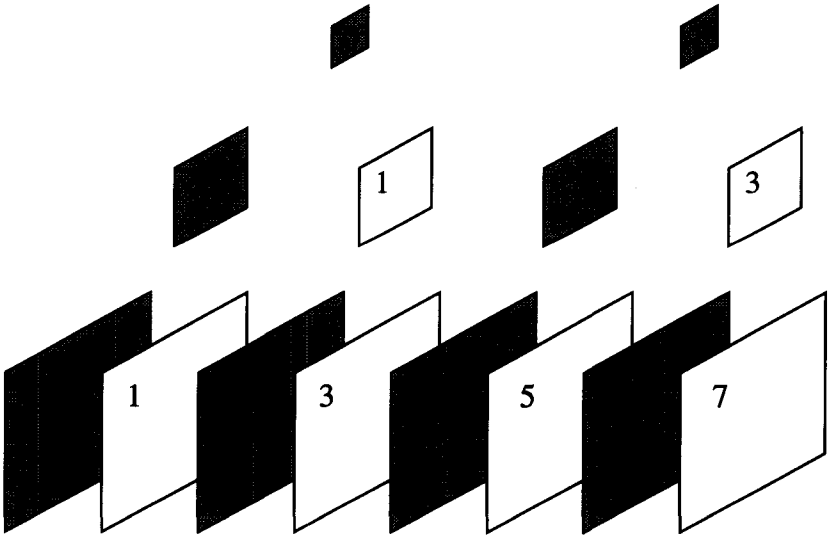

Figure 1 Reconstruction of a pyramid. Shaded frames are spatially coded and interpolated; white frames are temporally interpolated. (from [5])

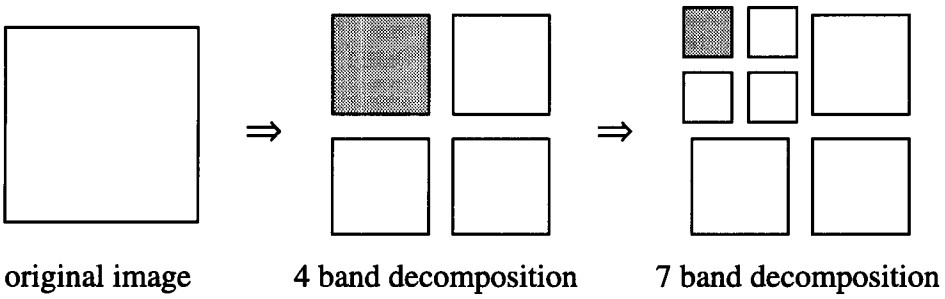

Figure 2 A two-dimensional hierarchical subband decomposition. The image is first filtered and subsampled vertically and horizontally, resulting in four bands. The shaded band is the low frequency representation, and resembles the original image. The low frequency band can be recursively decomposed, resulting in a smaller low frequency band. A multi-rate stream is generated by progressively transmitting the subbands.

More common are subband decompositions for scalable video. A subband decomposition consists of filtering and subsampling a signal recursively into frequency bands of equal or unequal width, depending on the hierarchy of filtering/subsampling operations. A seven band decomposition is illustrated in Figure 2. When appropriate filters are used, the original signal can be perfectly or nearly perfectly reconstructed from the frequency bands through upsampling and filtering. When the subband decomposition is performed on each frame in a video sequence, spatial scalability is provided. In [6], three levels of spatial resolution are provided by performing a 19-band subband decomposition, consisting of a 16-band full-band decomposition, and then a 4-band decomposition of the low frequency band. Each subband is separately motion compensated.

Quality scalability and temporal scalability with subband coding are introduced in [7], in which three-dimensional subband decomposition provides the temporal scalability. A sequence that is subband filtered temporally is decomposed into two half-rate sequences, one containing the low frequency band (and resembling the original sequence), and one containing the high frequency band. Multirate quantizers are used for the subband coefficients to provide quality resolution. This concept was extended in [8], in which better temporal pre-processing is used. These two algorithms provide a truly scalable bit stream, in which a fine granularity of bit rates is provided for a wide range of spatial and temporal resolutions.

As previously mentioned, these techniques are well suited to providing multi-rate streams; however, they are not without drawbacks. While multi-resolution coding via pyramid and subband decompositions provides excellent scalability properties, it has several drawbacks from an implementation standpoint. Much larger amounts of memory are required at the decoder because entire frames must be processed at once, rather than in a completely localized fashion as allowed by block-based coding. Secondly, the inclusion of temporal subband filtering requires more memory, and produces a latency of at least the length of the FIR filter used.

2.3 Rate Offerings in a Multi-Rate System

Depending on the coding technique selected, the granularity of rates may range from 2 or 3 (as in MPEG) to more than 10 (as in the subband techniques). A higher granularity of rates is more desirable from the standpoints of the service and network providers, while simple rate selection is required by the consumer. These two features can coexist.

Presenting the user with a selection of one of many rates (where many may be, say, more than 5) can prove confusing. For example, understanding the relative qualities of "low", "medium", and "high" quality video is easier than understanding the relative qualities of 10 streams, from 1 to 10 Mb/s. Therefore, the user may best be able to select a desired quality if the qualities are relatively few; for example, "L(ow), M(edium), H(igh), HD(TV)". Alternatively, the qualities can be listed as suggested topics, avoiding negative connotations associated with low or medium quality: "News, Dramas, Sports, HDTV."

However, from the network and server standpoints, providing few and fixed rates does not allow for great flexibility in providing quality-of-service (QOS) guarantees and reliable admissions control, and in dealing with traffic or usage bursts. Defining ranges of rates corresponding to each user quality level allows for adjustments within each rate category as traffic and usage change. Providing that the rate ranges are appropriately defined, small changes in rates are imperceptible to the viewer; for example, MPEG-2 videos generated from the same source material and coded at rates of 9 Mb/s and greater are virtually visually indistinguishable. The size of each range will be proportional to the quality level; the lowest quality will have a fixed lower bound and a smaller range to maintain visual equality.

3.0 Rate Negotiation

Rate negotiation is defined as the formation of an agreement between the user, server, and network, made at the time the video is selected, on the rate of the video to be delivered. Rate negotiation from the perspectives of the three components is described in Section 3.1. Ideally, rate negotiation should be performed to avoid *ping-ponging* between the three entities. Ping-ponging is defined as selection of a rate by the user that is unavailable, causing the system to respond with "not available" and requiring the user to select again. It can occur when the server or network load is too great to satisfy a user's request. Ping-ponging cannot be entirely avoided, but rate negotiation can be designed to minimize its probability through use of a *downstream flow of rate information*, which is described further in Section 3.2. Section 3.3 summarizes the rate negotiation process.

3.1 User, System, and Network Perspectives on Rate Negotiation

From the perspectives of each of the three entities, rate negotiation has different meanings.

From the user's perspective. Rate negotiation involves selection of a rate, either actively or passively. Active rate selection involves the consumer, who selects both a video and a rate. Passive rate selection is a function of predetermined hardware and/or software configurations. It can be solely a function of the delivery system to the set-top box. For example, a set-top box with an ADSL input from the network can only receive a maximum rate of 1.5 Mb/s, and if that is the minimum rate offered by the server then no negotiation is necessary; the viewer has no choice of rates at the time of video selection. Alternatively, if the set-top box has only subscribed to a low-rate service, then again the consumer has no choice of rates. However, the server and network must still be made aware of what rate is required when the video selection is made. Whether the viewer is passive or not, the set-top box performs this function (and will be discussed further in Section 3.2).

From the server's perspective. Rate negotiation tasks are twofold. To avoid ping-ponging, the server must make available information about what rates are available. This can be performed by self-monitoring and by updating selection listing information in its L2GW, which is made available to users for program selection. Updating must be performed at a frequency that minimizes ping-ponging caused by outdated information. The availability information can be obtained directly from the server's admission control algorithm, and hence no additional software burden is placed on the server to provide this information, other than communicating it to the L2GW.

From the network's perspective. Including the network in the avoidance of ping-ponging may or may not be possible, depending on the specific instance of the VOD system. We consider two systems, one with a *public network provider* and one with a *private network provider.* A public network provider is defined as one who will only transmit data, and who is either not willing or not able to modify content. A private network provider can modify content. A VOD system could use both; for example, a public network transmits data to a local node, from which point the service provider operates the distribution network. An example of this is video distribution within a hotel: the hotel requests full-rate video, and then internally redistributes it at the appropriate rates.

The role of the public network provider in rate negotiation is limited to refusing or allowing a connection at a requested rate once it has been negotiated between the user and the service provider.

For a private network provider, the network rate negotiation tasks are essentially identical to those of the service provider, but in the context of network load rather than server load. These will be discussed in more detail in the next section.

3.2 Downstream Flow of Rate Information

To avoid the complex exchange of information between the set-top box, the server, and the network, a downstream flow of rate information is proposed. Rate information originates at the server and possibly in the network and is interpreted by the set-top box before being presented to the user. The benefit of maintaining a downstream flow of rate information is that no user-specific

information is required anywhere but at the set-top box (i.e., at the user's location) at the time of selection. Specifically, no knowledge about the network characteristics or the user is required at the Level 2gateway, no knowledge about the set-top box is required in the network, and therefore all user-specific information is limited to the user's location. No rate information travels upstream until the user makes a rate selection. The key to the feasibility of a downstream flow of rate information is intelligence in the set-top box to perform the appropriate filtering functions. This section describes the flow of information and the motivation for such.

At the server and Level 2 gateway. The L2GW maintains menus of server offerings, with the rate availability information updated periodically as discussed in the previous section. When requested, a menu or menus are transmitted from the server's L2GW to the set-top box via the network. Transmitting all the rate information is not a tremendous overhead, as it may simply consist of 3 bits for each selection, representing the availability of low-medium-high rate video.

In the network. The network contribution to the downstream flow of rate information is a function of the complexity of tasks that the network can perform, as described below.

In a public network. In the simplest case, the network simply delivers the transmitted data without regard to its contents. If the network is to provide availability information for rate selection, it must be transmitted to the set-top box separately. Because a set-top box may be accessing a server that is separated from it by several network segments, including all the network information, and hence providing accurate information for rate selection, may not be possible.

In a private network. If the network is able to identify and modify data that it transmits (for example, in the switches), the network information can be appended to the menu data en route; that is, as it passes through switches between network segments, it is updated.

In order of complexity of network requirements, the network contribution to the downstream flow of rate information is:

1. None; the network merely transmits the data and provides no input.
2. The network provides last-segment information by separate transmission to the set-top box.
3. The network provides complete bandwidth information by appending availability information to the menus as they are switched.

The third contribution involves network modification of application-layer information, the possibility of which depends on particular network and service providers. The second merely involves communications of the network with the set-top box, and hence may fit better into the standard OSI seven layer model.

At the set-top box. The menus arrive at the set-top box, which uses the network rate information and its own characteristics to produce menus with only available videos and available rates

shown, thereby disallowing the user from selecting an unavailable rate. Such menu generation requires intelligence in the set-top box, where intelligence is defined as the capability to render menus based on both transmitted information and internal information. The internal information is stored in a *profile* in the set-top box, containing user information including the maximum rate. This concept can also be extended to service selection from the Level 2 gateway, in which the user is prevented from selecting services at rates that are too great.

Ping-ponging may still occur in one of three ways.

1. Outdated information is transmitted to the set-top.
2. Information at the set-top becomes outdated while the user is selecting.
3. Many users make program requests to the server within a small period of time, thereby rendering the information outdated by the time it arrives.

However, an appropriate admissions control algorithm at the server can minimize such ping-ponging.

The number of navigation states that are stored locally at the set-top box can also affect ping-ponging. If a large hierarchy of menus are transmitted at once, then the user may spend a long time working through them, and finally make a selection based on outdated information. In this case, the L2GW can perform an update function, in which updated rate information only is transmitted periodically to a set-top box in the selection process. Given the high data rates and update rates required for interactive games, the rate required for rate update are negligible and can be easily handled. However, the L2GW must then store the states of each set-top box requesting menu information. This adds complexity to the gateway and is hence undesirable.

3.3 The Rate Negotiation Process

This section lists the rate negotiation process. Note that the majority of systems issues involving rate negotiation are included in Step 1 with the downstream flow of rate information.

1. The downstream flow of rate information presents the user with available rate information while minimizing user-specific data at the server and in the network.
2. The user selects a video and a rate from the menus rendered at the intelligent set-top, in which rate and video availability information have been filtered using possibly available network data and the set-top box profile. Rate selection is either active or passive.
3. Call/connection set-up proceeds. If set-up fails due to network or server unavailability, the user is so informed and selects another rate and/or video.

4.0 Rate Delivery

Once a rate has been negotiated between the user, network, and server (and a connection or call has been set up), the video must be delivered. Rate delivery encompasses reading the requested information from the server, transmitting it reliably over the network, and decoding it at the set-top box. Each delivery task should be performed efficiently, to maximize the usage of the system resources, and reliably. This section describes rate delivery issues for the server, the network, and the set-top box. By far the most issues concern the server, because efficient server use (and hence a cost effective server) is imperative to the economic success of a VOD system.

4.1 Rate Delivery at the Server

The ease and efficiency of rate delivery at the server are affected by the video coding algorithm, and the algorithm should be designed with the server's requirements in mind. Two server design issues arise when considering storing and reading multi-rate video streams: data layout and the combined issue of data block sizing and read scheduling.

4.1.1 Multi-Rate Coded Video and Efficient Data Layout

A fundamental conflict exists between efficient decoding of a scalable stream and efficient data layout of such a stream on a server. Efficiency in decoding corresponds to the amounts of memory and data processing required at the decoder, while efficient data layout refers to placement of the scalable video stream on a disk or disks such that it can be read with high disk I/O bandwidth utilization. The amount of post-read processing required at the server is also a function of the data layout.

The conflict exists because a scalable stream can be fundamentally organized in two ways: in time, or in rate. A stream *organized in time* (or *time organized*) is an ordered sequence of bits as required by the decoder in order to decode the highest-rate video. When lower-rate video is required, bits must be discarded within the stream whenever higher-rate information is reached in the sequence. This is illustrated in Figure 3. For example, in each frame, all bits are used to a certain point, and then the remaining bits are skipped until the next frame begins. As such, a time organized stream is maximally efficient at the decoder, because the stream can be continuously decoded with only the memory required to perform the decoding. The latency at the decoder from reception of the data to display is simply the decode time of the data.

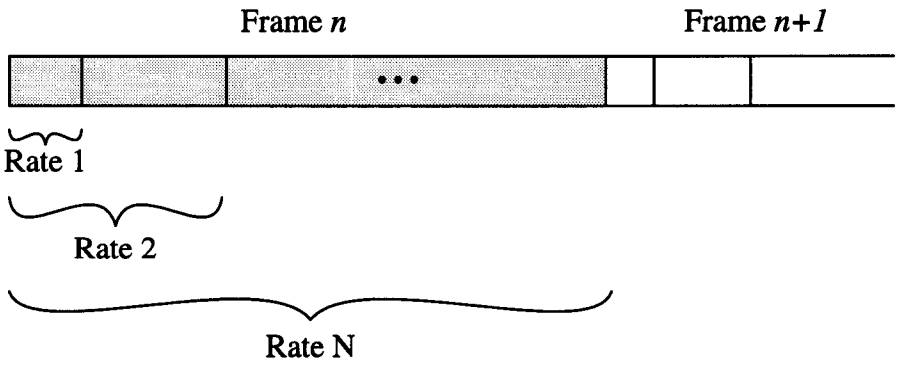

Figure 3 A scalable single frame segment of a video stream, organized in time.

The required time-organized stream can be generated at the server in one of two ways:

1. Read the full-rate stream from disk, and perform post-read processing to strip unnecessary bits. This technique requires real-time stream processing capabilities at the video server. Furthermore, disk bandwidth utilization is low, because excess information is read and discarded, unless a full-rate request is being satisfied.

   Alternatively, the post-read processing can be performed in the network or in the decoder. While this alleviates the processing demands on the server, unless a full-rate stream has been requested, both server and network bandwidth are wasted.

2. Read only the bits that are required from disk. This technique does not efficiently utilize the bandwidth either, because of the excessive number of seeks required. A constant number of seeks is required, regardless of the rate being read, so at lower rates, the disk bandwidth utilization decreases.

A stream *organized in rate* (or *rate organized*) is a stream of bits as required by the decoder to progressively increase the rate of a video sequence. First the lowest-rate information is stored, followed by the second-lowest rate information, and continuing to the highest-rate information. A rate organized stream is efficient with respect to server layout, because large blocks of data are read, starting with the lowest-rate and continuing until the desired rate is read. Obviously, a rate organized stream requires some segmentation in time; otherwise each rate block consists of video data for the entire program. A rate organized stream for $k+1$ frames is illustrated in Figure 4. Once the rate organized video data has been read, it must be reordered for decoding or for playback. Reordering can be performed at one of two locations:

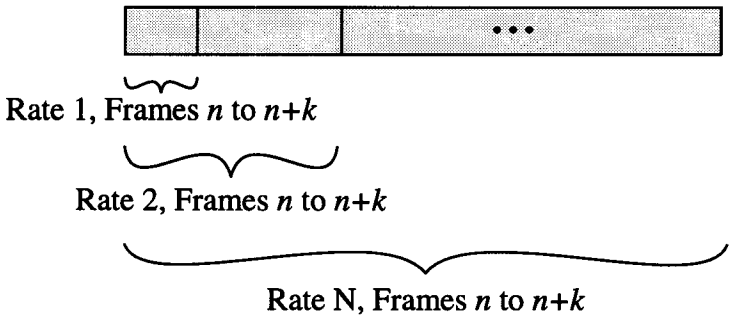

Figure 4 A scalable $k+1$ frame segment of a video stream, organized in rate.

1. Rate organized streams are remultiplexed into time organized streams at the server before being transmitted to the decoder. This requires real-time stream processing at the server. However, the video coding algorithm can be designed to facilitate such processing. If the rate data for each frame is restricted to be an integer number of byte groups that are packetized and placed on the network, then the remultiplexing simply consists of reading the video data into the packetizer not from contiguous buffer locations in the server, but from buffer locations in a predetermined order. The read memory locations could be stored with the video data at each rate. In this case, the latency at the decoder is identical for a time organized stream.
2. Rate organized streams are simply transmitted to the set-top box and decoded. As each rate is decoded, it is stored in memory. When all rates for a time segment have been decoded, they are combined and displayed. The length of the time segments in which the data is rate organized affects the memory requirement at the decoder: longer time segments provide for more efficient disk reads, but require more memory and produce more latency at the decoder. The latency includes the time to transmit the rate-organized time segment, plus any overhead required to recursively decode the stream at increasing rates rather than combining all the information into one rate and then decoding simultaneously. To a first order approximation, the latency is linear in the number of frames in the rate-organized time segment.

Achieving maximum efficiency in both the decoder and the server is impossible, so a compromise is required. The data-layout problem has been examined by [9], in which a rate-organized technique is proposed, with a time segment length set to maximize efficiency, but the resulting disk bandwidth utilization only reaches approximately 25%. A technique that facilitates utilization of over 80% is needed for a VOD system [10].

4.1.2 Data Block Sizing for Disk Read Efficiency and Ease of Scheduling

When only single-rate constant-bit-rate (CBR) video is provided by a server (such as MPEG-coded video at a single rate), a constant number of bits corresponds to approximately a constant display time, if the video segment is long enough to integrate temporal bit variations caused by the periodicity of MPEG coding. Therefore, reading video data based on *constant data length (CDL)* blocks or *constant time length (CTL)* blocks is equivalent. Furthermore, since all users require data at the same rate, the streams can be scheduled for reading in a round-robin fashion, reading a data block from each stream once in each service round, where a service round is defined as the amount of time to service all currently playing and newly requested videos.

When a server stores multi-rate video, CDL and CTL are not equivalent, and scheduling block reads for each video stream can become difficult. If data is read from disks in CTL blocks, then scheduling remains as round-robin servicing, because each stream consumes its data over the same time period. However, in this case disk bandwidth utilization can suffer, as blocks of varying sizes are being read. (say: same as in server section but at a different level of granularity) CTL readout has been used to develop a statistical admissions control analysis in [11].

If data is read from the disks in blocks of constant length, then streams of differing rates must be read at different intervals. These reads must be scheduled so that each stream is read in a time window such that the data is not too early, causing a buffer to overflow (the buffer could be at the set-top box, or at the cable head-end, for example), or not too late, causing a buffer to underflow. In both cases, the smooth decoding of the video stream is disrupted. However, reading CDL blocks allows for design of a maximum disk bandwidth efficiency which can be achieved asymptotically if the scheduling algorithm works well.

CDL block reads also require excess memory in the system to permit work-ahead when reading blocks. If no excess memory is present, no latitude exists with respect to read times for a given stream. Once the first block is read, the rate completely determines the subsequent times at which blocks for the stream must be read. Such a deterministic schedule can easily produce overloads, in which too many streams require service at the same time. However, if each stream has the buffer available to hold one extra CDL block, the system can add users until the maximum users/service round theoretical limit is reached. This solution requires a doubling of system buffer space. Increasing the buffer space by a factor less than two is expected to produce similar results, but requires more complex scheduling.

In both cases of CTL and CDL block reads, a video coding algorithm that produces Pseudo-CBR (PCBR) is desirable. PCBR streams provide regularly spaced rates, such as $p \times 64$ kb/s coded video, or at least discrete rates within the scalable stream. PCBR CTL blocks then correspond to a approximately fixed numbers of bits, and PCBR CDL blocks then correspond to approximately fixed display times. In both cases, scheduling is much easier to analyze than if a continuum of rates is available.

4.2 Rate Delivery in the Network

If the network is a public network, then it requires the capability of delivering streams of varying bit rates, but beyond that the network issues are not interesting. However, if the network is a private network, then network distribution of the video becomes feasible. A common example of a private network is the CATV network, in which the content provider own the distribution network from the head-end to the set-top box.

If the network monitors its content, that is, if it is aware that multiple users downstream from a particular location are viewing the same video but at multiple rates, then it can reduce its load by only carrying the video once, at the highest-rate required. As the rates reduce downstream, the network processes the stream, stripping out the unneeded higher rate information, and only passing the highest required rate. For such processing, a compressed video stream organized in rate, as described in Section 4.1.1, and transmitted as such is desirable. Transmitting rate organized streams minimizes the amount of network processing required to remove unneeded rate information when compared to transmitting time-organized streams. However, as mentioned, transmission of rate organized streams comes at the expense of decoder latency.

Obviously, such rate delivery in the network requires processing power in the network, as well as network knowledge of and input to the applications layers being transmitted. This is typically not desirable. Furthermore, if encryption is included at the server, then many security issues quickly arise, and the network nodes themselves may have to act as users in requesting programs from the server, receiving their own encryption key, and then re-encrypting the videos for users downstream. In this case, the network node is acting in the role of a "mini-server" that performs a bulk "fetch" of a single stream for multiple clients. Such a concept was proposed in [12], in which the network provides storage for video streams to provide decoupling of the number of users from the number of server accesses.

4.3 Rate Delivery at the Decoder

The decoder presents only one requirement in rate delivery. It must have the capability to decode streams of varying resolution. As the resolution will be coded into the steam, this is not a problem. As previously mentioned in the Server subsection, extra buffer space may be required.

5.0 Conclusions and Research Topics & Design Issues

Inclusion of multi-rate video in a VOD system clearly adds flexibility for the service provider, network provider, and user. However, as this report has described, providing such a feature affects the design of each component of the system, and the overall system will be most effective if each component is designed based on a common multi-rate model. For example, the use of rate-organized video streams at the server, which provides maximum disk bandwidth utilization, can be coupled with network rate delivery, which requires additional network processing. The combination of these two options requires the set-top box to have additional memory to cope with latency.

The use of a downstream flow of rate information places additional requirements on all components, each of which must be adequately addressed in order for the rate negotiation process to function reliably.

The technical problems that arise from inclusion of multi-rate in a VOD system can be addressed. However, because the entire system must be coordinated, non-technical concerns may impose the most constraints. Economic factors, such as the cost of server, network, and user equipment, may strongly influence what is determined to be feasible. Furthermore, a system may require coordination across multiple owners: the service provider and the network provider may be separate entities with differing views of what the system should provide.

This report has provided an overview of system-level issues for incorporation of multi-rate video into a video-on-demand system. Specific research topics and design issues for further investigation follow.

- Development of a server admission control algorithm, considering frequency of update of rate availability information in the L2GW. The update schedule and admission control algorithm should consider user request models in their design.
- Issues related to service provider driven multi-rate, in which the server modifies rates of current videos being delivered to increase users and throughput:
  - Development and optimization of an algorithm to reduce rates such that all viewers continue to receive video of the highest perceptual quality possible.
  - A study of perceptual coding in conjunction with the video coding algorithm to determine the sizes of the rate ranges for various quality levels.
  - Determination of the minimum required rate granularity for the desired system performance.
- Development of a multi-rate video coding algorithm in conjunction with a data layout strategy for high disk I/O bandwidth throughput.
  - Analysis of CDL vs. CTL layout, with an appropriate scheduling algorithm for CDL data blocks.
- Analysis of how public networks can participate in the downstream flow of rate information.
- Systems analysis of the rate granularity breakpoint below which it is easier and less costly to simply store multiple copies of the same video.
- In the case of a cable head-end with 6 MHz channels, development of multiplexing algorithms for streams of unequal rates to maximize bandwidth usage: the streams must be packed efficiently into the available channels so that at any time, the rate of any new stream that can be accommodated is maximized.

6.0 References

[1] Y.-H. Chang, et. al., "An Open-Systems Approach to Video on Demand," *IEEE Communications Magazine*, Vol. 32, No. 5, pp. 68-80, May 1994.

[2] Y.-H. Chang, "Video Dial Tone (Level 1) Gateway", Draft Document, June 14, 1994.

[3] "Broadband Services Gateways," TPO Draft Document, Rev. 2.2, August 5, 1994.

[4] ISO/IEC JTC1/SC29 WG11/602, Information Technology —Generic Coding of Moving Pictures and Associated Audio, Recommendation H.262, ISO/IEC 13818-2 Committee Draft, November 4, 1993.

[5] K. M. Uz, M. Vetterli, D. LeGall, "Interpolative Multiresolution Coding of Advanced Television with Compatible Subchannels," *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 1, No. 1, March 1991, pp. 86-99.

[6] J. W. Woods and T. Naveen, "Motion Compensated Multiresolution Transmission of HD Video Using Multistage Quantizers," *Proc. ICASSP '93*, April 1993, Vol. 5, pp. 582-85.

[7] E. Chang and A. Zakhor, "Scalable Video Using 3-D Subband Velocity Coding and Multirate Quantization," *Proc. ICASSP '93*, April 1993, Vol. 5, pp. 574-77.

[8] D. Taubman and A. Zakhor, "Multi-rate 3-D Subband Coding of Video," *IEEE Trans. on Image Processing*, Vol. 3, No. 5, Sept. 1994, pp. 572-88.

[9] E. Chang and A. Zakhor, "Scalable Video Data Placement on Parallel Disk Arrays," *IS&T/SPIE International Symposium on Electronic Imaging, Volume 2185: Image and Video Databases II*, San Jose, Feb. 1994.

[10] Manu Thapar, Yitzak Birk, private communication.

[11] E. Chang and A. Zakhor, "Admissions Control and Data Placement for VBR Video Servers," *to be presented at The International Conference on Image Processing, November 1994.*

[12] D. Deloddere, W. Verbiest, and H. Verhille, "Interactive Video On Demand," *IEEE Communications Magazine*, Vol. 32, No. 5, pp. 82-8, May 1994.