Dynamic QoS Management for Scalable Video Flows

Andrew Campbell[*] *and David Hutchison*
Department of Computing,
Lancaster University,
Lancaster LA1 4YR, U.K.
mpg@comp.lancs.ac.uk

Cristina Aurrecoechea
Center for Telecommunications Research
Columbia University
New York, NY 10027-6699.
cris@ctr.columbia.edu

Abstract. We introduce the concept of *Dynamic QoS Management (DQM)* for control and management of hierarchically coded flows operating in heterogeneous multimedia networking environments. The motivation that underpins our scheme is to bridge the heterogenity gap that exists between applications, end-systems and networks. *QoS adaptors, QoS filters* and *QoS groups* are key *scalable objects* used in resolving quality of service capability mismatch. QoS filters manipulate hierarchically coded flows as they progress through the communications system, QoS adaptors scale flows at the end-systems based on the flow's measured performance and user supplied *QoS scaling policy,* and QoS groups provide baseline quality of service for multicast communications. The focus of the work is driven by a) the special features of scalable video flows - in particular MPEG2, b) the needs of both scalable and single-layer video for transmission over multimedia networks such as ATM. A novel *adaptive network service* is proposed for the transmission of multi-layer coded flows that offers "hard" guarantees to the base layer, and "fairness" guarantees to the enhancement layers based on a new bandwidth allocation technique called *Weighted Fair Sharing (WFS).*

1. Introduction

The interplay between hierarchically coded flows [1], end-to-end communication support [2] and receiver-oriented [3] Quality of Service (QoS) requirements is an interesting and active area of research [4] [5] [6]. The basic technique used by coders (e.g., MPEG1, MPEG2 and H261) for compression of audio-visual flows is to remove redundant information in the signal. Using creative design techniques that take into account the perceptual capabilities of the human aural and visual system, it is possible to eliminate substantial parts of the signal with little or no perceived loss of information.

Both end-systems and networks impact continuous media communications through degradation in the delivered QoS (i.e., late frames, momentary loss of bandwidth, or highly variable delay jitter). Many audio and video applications exhibit robustness in adapting to fluctuations in the quality of service offered by the network and end-system. By trading off temporal and spatial quality to available bandwidth [7] [8], or manipulating the playout time of continuous media in response to variation in delay [9] [10], the video signal can be kept meaningful at the playout device with minimal perceptual distortion. The perceptible quality of service presented at the receiver is therefore a complex of quality of service management functions that includes source coding, filtering, communications support and QoS adaptation.

In this paper we present a *Dynamic QoS Management (DQM)* scheme and an *adaptive network service* which have been specifically designed to cater for the needs of both single and multi-layer coded video flows. We first motivate our approach by discussing some of the relevant end-to-end QoS challenges in transporting digital video over multimedia networks. We then present an API for scalable flows, followed by a detailed description of our adaptive service and dynamic QoS management scheme. DQM is in part built on existing techniques in the literature [1] [4] [5] [7] [9] [13] in provinding end-to-end QoS management [2] of digital video flows [15] [7].

[*]Visiting Scholar at the Center for Telecommunications Research, Columbia University.

2. Motivation

Heterogenity issues are present in applications, end-systems and networks. The range of audio and video QoS requirements is likely to be very diverse and the capability of applications, end-systems and networks to handle continuous media is likely to be quite diverse too. For example multimedia conferencing may only require low resolution video but high resolution sound. Considering the end-system only, heterogeneity is present in: CPUs, I/O devices, storage capabilities, compression support, communication support, network interfaces, etc. - all place fundamental limits on the capability of end-systems to consume and generate digital video. End-systems are likely to be connected to a wide variety of networks which not only have differing bandwidth capabilities but also varying access delay, jitter and loss characteristics. A key challenge is resolving this potential quality of service mismatch. We suggest that QoS scalability is central in bridging this heterogenity gap [5]. To meet this challenge we propose a DQM scheme that accommodates differing application, end-system and network QoS needs.

Another challenge is the transmission of variable bit rate video over multimedia networks. In order to derive algorithms for supporting variable bit rate network transmission (e.g., admission control, resource reservation and packet scheduling) of digital video, significant effort has been expended to obtain statistical models for video traffic [11]. The problem is extremely complex, since the rate is intimately related to both the coding scheme and the actual source material through the use of entropy coding: video sequences that deviate from the "norm" will require more bits than the average, and may exhibit highly variable behaviour.

As a result, models tend to be quite complex for variable bit rate video, and there is currently no known universally acceptable one. The drawback of inadequate modeling is that it can result in either loss of information (by underestimating the traffic behaviour) or to wasteful use of network resources (through overestimation) or reliance on constant bit rate "circuits". Some general observations have been made regarding coded digital video [7]: (i) burstiness is heavily dependent on the content of the image being coded and the coding algorithm used; (ii) source generation rates are highly sensitive to scene and background changes; and (iii) highly correlated source traffic is potentially persistent over very long periods. All this makes modeling video in a general sense extremely difficult and complex. Because modeling video sources is so difficult, and constant bit rate channels so inefficient in exploiting statistical multiplexing we focus our attention in this paper on an alternative approach: an adaptive channel, which supports QoS adaptation through the semantics of scalable video flows [5] [12] and resource fairness [12] [14] as part of the DQM scheme.

3. Scalable Video Flows

MPEG2 [15] provides for the simultaneous representation of a video signal at various different levels of quality, through the use of multiple independent bitstreams or sub-signals. This is achieved through the use of pyramidal, or hierarchical coding: one first constructs a coarse or base representation of the signal, and then produces successive enhancements. The latter assume that the base representation is available, and only encode the incremental changes that have to be performed to improve the quality. There are four different scalability modes: spatial, SNR, temporal, and data partitioning.

Although MPEG's scalability features are useful in resolving heterogeneity problems described above [5], and are useful in numerous applications, their use in continually QoS-varying channels is problematic. This is because they only allow the representation of the signal at a number of discrete quality points (temporal or spatial resolution, or spatial quality). These points are typically significantly apart, and transitions between the two are perceptually significant. Table 1 [16] shows an example of hybrid scalability with spatial

(E1) and SNR (E2) enhancement layers.

layer name	*profile*	*symbol*	*frame size*	*bit rate*	*subjective QoS*
base layer	main	BL	304x112	0.32 Mbps	VHS
enhancement 1 layer	spatial	E1	608x224	0.83 Mbps	super VHS
enhancement 2 layer	SNR	E2	608x224	1.85 Mbps	laser disc

Table 1: MPEG-2 hybrid scalable bitstream using spatial and SNR scalability (24 fps)[16]

Consider for example a channel that temporarily sustains rate variability for a period of a few seconds. Switching to a lower quality point (by discarding the enhancement layer(s)) for such a brief interval will essentially create a "flash" that is very annoying to viewers (we call this *discrete adaptation*); this type of degradation would be rather noticable in the case of the spatial enhancement described in Table 1. An additional issue is that, as soon as compression parameters are established, it is impossible to modify them later on (after compression is completed). Hence scalability modes can be used for well-defined, simple channels that vary slowly. Since the variety of different access mechanism to multimedia information makes it very difficult to select a priori a set of universally interoperable coding parameter, it is necessary to provide mechanisms that allow the representation of the "signal at a continuum of qualities and rates" (we call this *continuous adaptation)* [5], so that scalable channels can be accommodated. This is possible through the use of dynamic rate shaping filters [8] and the provision of adaptive network services - providing a QoS continuum for fully scalable flows. The adaptive service uses explicit feedback from network resource management to dynamically shape the video source based on the available network resources. Some benefits of an adaptive scheme are non-reliance on video modeling techniques and statistical QoS specification and specific support for the semantics of scalable video flows e.g., MPEG scalable profiles. Dynamic rate shaping filters manipulate the rate of MPEG-coded video, matching it to the available bandwidth (indicated by the adaptive service) while minimising the distortion observed by the receiver.

4. Application Programming Interface for Scalable Flows

4.1 Scalable Objects

QoS adaptors, QoS filters and QoS groups are key scalable objects in dynamic QoS management. These objects are used to resolve quality of service capability mismatches in the end-systems and network and provide communication support for single and scalable video flows. QoS adaptation relates to the monitoring (using *flow monitors*) and adjustment (using QoS adaptors) of flows at the edge of the network to ensure that the user and provider quality of service is maintained. In this role QoS adaptors are seen as quality of service arbiters between the user and network. QoS adaptors scale flows at the end-systems based on a user supplied QoS scaling policy (see section 4.2) and the measured performance of on-going flows. In contrast to adapting flows at the end-systems, QoS filters manipulate hierarchically coded flows [1] [6] at the end-systems and as they progress through communications systems. In dynamic QoS management we refer to *scaling* [5] [12] as an "umbrella" term to cover the combination of QoS adaptation in end-systems, and QoS filtering [4] [2] in end-systems and the network. The QoS scaling policy is central to DQM and is the driver of QoS adaptation and QoS filtering mechanisms for end-to-end QoS management.

We describe three styles of QoS filters in dynamic QoS management:(i) *shaping filters*, which manipulate coded video and audio by exploiting the structural composition of flows to match network, end-system or application QoS capability. Shaping filters are generally situated at the edge of the network; (ii) *selection filters*, which are used for sub-signals source selection and media dropping are of low complexity and low computational intensity.

Selection filters are designed to operate in the network and are located at switches; and (iii) *temporal filters*, which manipulate the timing characteristics of media to meet delay bound QoS are also low in complexity and trivial computationally. Temporal filters are generally placed at receivers or sinks of continuous media were jitter compensation or orchestration of multiple related media [2] is required.

The first and third types of QoS filters are predominantly located in the edges of the network. Shaping filters utilise knowledge of the coding details of the flow they are processing, and require non-trivial computational power. Selection filters, on the other hand, perform simple packet filtering and hence can be located in internal network nodes. In some cases, shaping filters can be located in special network nodes, either as a bearer service, or as part of special environments such as mobile communication links. As an example, rate shaping filters placed at base stations can be useful for multi-point communication [1] with both mobile and wired hosts. Through the use of rate shaping, wired users can utilise the full bandwidth at their disposal, without compromising quality to that attainable to the least capable link of the session. A distributed object-based facility (e.g. CORBA [22]) can be particularly effective in providing the foundation for the incorporation of filters throughout the networking infrastructure (see section 6.) [12].

Before receivers or senders bind [17] media source and sink devices, media protocols and scalable objects to form end-to-end connections they must first join a QoS group [18]. A flow is represented by a QoS group in our adaptive (CORBA-based) environment. Group management announces QoS group capability in terms of its quality of service capability. The concept of a QoS group is used to associate a baseline quality of service capability to a particular flow. All sub-signal of a multi-resolution stream can be mapped into a single flow and multicast to multiple receivers [1]. Each receiver can select to take the complete signal advertised by the QoS group or a partial signal based on resource availability. Alternatively each sub-signal can be associated with a distinct QoS group. In this case receivers "tune" into different QoS groups (using signal selection) to build up the overall signal. Both methods are supported in DQM. Receivers and senders interact with QoS groups to determine what the baseline service is, and tailor their capability to consume the signal by selecting filter styles and specifying the degree of adaptability sustainable (viz. discrete, continuous).

4.2 QoS Specification

In [2] we formalised the end-to-end QoS requirements of the user and the potential degree of service commitment of the provider in a service contract. In this work, we focus on the extensions to the flow specification, QoS commitment and QoS scaling clauses required to accommodate the special needs of adaptive multi-layer flows over multimedia networks. Multi-layered flows are characterised by three sub-signals in the flow specification: a base layer (BL), and up to two enhancement layers (E1 and E2, respectively). Each layer is represented by a frame size, bit rate and subjective or perceptive quality of service as illustrated in Table 1. Based on these characteristics the MPEG2 coder [16] [8] determines approximate bit rate for each sub-layer. In the case of MPEG-2's hybrid scalability, BL would represent the main profile bit rate requirement (e.g., 0.32 Mbps) for basic quality, E1 would represent the spatial scalability mode bit rate requirement (e.g., 0.83 Mbps) for enhancement, and finally E2 would represent the SNR scalability mode bit rate requirement (e.g., 1.85 Mbps) for further enhancement. For full details of deriving these bit rates. The remaining flow specification performance parameters for jitter, delay and loss are assumed to be common across the all sub-signals. We use the term *sub-signal* to represent a single layer of a multi-layer video flow; and the term *flow* as a non-assured simplex uni-media stream, comprising of one source and potentially many receivers; flows always have end-to-end QoS associated with them [2].

The scaling policy characterises the degree of adaptation that the flow can tolerate and still achieve meaningful QoS. The scaling policy has been extended to capture the special needs of multi-layer flows, and includes adaptation modes, QoS filter styles, and user level notifications options for bandwidth, delay, jitter and loss QoS signals. Two types of adaptive mode are supported [5]: continuous mode, for applications that can exploit any availability of bandwidth above the base layer; and discrete mode that is suitable for applications which can only accept discrete improvement in bandwidth based on a full enhancement (viz. E1, E2). The adaptive modes option covers both highly adaptive (e.g., MPEG2 using dynamic rate shaping) and coarsely adaptive (e.g., MPEG2 scalable profiles).

```
typedef struct {                        typedef   struct {
        gid           flow_id;                  adaptMode      adaptation;
        commit        commitment;               filterStyle    filtering;
        mediaType     media;                    events         adaptEvents;
        bitRate       BL;                       actions        newQoS;
        bitRate       E1;                       signal         bandwidth;
        bitRate       E2;                       signal         loss;
        int           delay;                    signal         delay;
        int           loss;                     signal         jitter;
        int           jitter;           } QoSscalingPolicy;
} flowSpec;

typedef enum {MPEG1, MPEG2, H261, JPEG} mediaType;
typedef enum {drs,sbr,sub_signal,hierarch,hybrid,sync,orch} filterStyle;
typedef enum {besteffort, adaptive, deterministic} commit;
typedef enum {continuous, discrete} adaptMode;
```

The QoS scaling policy provides user-selectable QoS adaptation and QoS filtering. QoS filters are broadly divided into source-based (viz., drs-filter and sbr-filter - see section 6.1), network-based (viz., sub_signal-filter, hierarch-filter, hybrid-filter - see section 6.3), and receiver-based (viz., sync-filter, orch-filter - see section 6.2) filters. While receivers select filter styles to match their capability to consume media at the receiver, senders select filter styles to shape flows in response to the availability of network resources such as bandwidth and delay. Receivers and senders can select periodic performance notifications including available bandwidth, measured delay, jitter and losses for an on-going flows. The QoS signal field in the scaling policy allows the user to specify the interval over which a QoS parameter is to be monitored and the user informed. Both single and multiple quality of service signals can be selected depending on the applications needs. For full details on the service contract see [2]. In addition, the QoS commitment clause has been extended to offer an adaptive network service that specifically caters for the needs of scalable audio and video flows.

5. Adaptive Network Service for Multi-layer Flows

The adaptive service provides "hard" guarantees to the base layer (BL) of a multi-layer flow and *Weighted Fair Share (WFS)* to each of the enhancement layers (E1 and E2). To achieve this, the base layer undergoes a full end-to-end admission control test [19]. On the other hand, enhancement layers are admitted without any such test but must compete for *residual bandwidth* among all other adaptive flows. Enhancement layers are rate controlled based on explicit feed back about the current state of the on-going flow and the availability of residual bandwidth.

5.1 Service Goals and Requirements

Both end-system and network communication resources are partitioned between the deterministic and adaptive service commitment classes. This is achieved by creating and maintaining "firewall" capacity regions for each class. Resources reserved for each class, but not currently in use can be borrowed by the best effort service class on condition of pre-emption [19]. The adaptive service capacity region (called the available capacity region and denoted by B_{avail}) is further sub-divided into two regions: (i) guaranteed capacity region (B_{guar}), which is used to guarantee all base rate layer flow requirements; (ii) and residual capacity region (B_{resid}), which is used to accommodate all enhancement rates were competing flows share the residual bandwidth.

Three goals motivate our adaptive service design: First, to admit as many base layer (BL) sub-signals as possible. As more base layers are admitted the guaranteed capacity region B_{guar} grows to meet the hard guarantees for all base signals. In contrast, the residual capacity region B_{resid} shrinks as enhancement layers compete for diminishing residual bandwidth resources. The following invariants must be maintained at each end system and switch:

$$B_{avail} = B_{guar} + B_{resid} \text{ , and } \sum_{i=1}^{N} BL_{(i)} \leq B_{avail} \qquad (1)$$

Second, to share [13][14] the residual capacity B_{resid} among competing enhancement sub-signals based on a flow specific *weighted factor (W_{fact})*, which allocates residual bandwidth in proportion to the range in bandwidth requested that in turn is related to the range of perceptual QoS acceptable to the user. In DQM, residual resources are allocated based on the range of bandwidth requirements specified by the users (i.e., BL.. BL+E1+E2 is the range of bandwidth required e.g., from 0.32 Mbps to 3 Mbps for the hybrid scalable MPEG2 flow in Table 1). As a result, as resources become available each flow experiences the same "percentage increase" in the perceptible QoS, we call this weighted fair share. W_{fact} characterises the notion of WFS (see (2)) and is calculated for each flow as the ratio of a flow's perceptual QoS range to the sum of all perceptual QoS ranges.

$$W_{fact(i)} = (BL_i + E1_i + E2_i) \,/\, \sum_{j=1}^{N} (BL_j + E1_j + E2_j) \qquad (2)$$

All residual resources B_{resid} are allocated in proportion to $W_{fact(i)}$ metric. Using this factor we calculate the proportion of residual bandwidth allocated to a flow to be $B_{wfs(i)}=W_{fact(i)}B_{resid}$ and the proportion of the available bandwidth allocated to be $B_{flow(i)}= B_{wfs(i)}+ BL_{(i)}$. We describe this aggregate as the flow bandwidth (B_{flow}).

Third, to adapt flows both discretely and continuously based the adaptation mode. In the discrete mode no residual bandwidth is allocated by the WFS mechanism unless a complete enhancement can be accommodated (i.e., $B_{wfs(i)}$ = E1(i) | E1(i)+E2(i) e.g., 0.83 Mbps or 2.68 Mbps from Table 1). While in continuous mode any increment of residual bandwidth $B_{wfs(i)}$ can be utilised (i.e., $0 < Bwfs(i) \leq E1(i) + E2(i)$ e.g., from 0 to 2.68 Mbps from Table 1). Adaptive applications can be considered to be either coarsely (e.g., MPEG-2 scalable profiles) or highly adaptive (e.g., scalable MPEG2 using a dynamic rate shaping). By selecting continuous adaptation highly adaptive applications can take advantage of any availability in bandwidth to enhance QoS. While coarsely adaptive applications are more suited to the discrete mode were only E1 and E2 sub-signals can be accommodated, nothing more nothing less. In addition, WFS always accommodates the full E1 signal before attempting to deliver the E2.

5.2 Rate Base Flow Control

We build on the rate-based flow control mechanism described in [2] were the transport protocol at the receiver measures the bandwidth, delay, jitter and loss over an interval which we call an "era". An era is currently defined as simply the reciprocal of the frame rate in the flow specification (e.g., for a frame rate of 24 frames per second as shown in Table 1 the interval era is approx. 42 ms.). The receiver-side transport protocol periodically inform the sender-side about the currently available bandwidth, and measured delay, loss and jitter. This rate control information is used by the source or virtual source (see later) as the rate over the next interval. The reported rate is temporally correlated with the on-going flow. An important result in [7] shows that variable rate encoders can track quality of service variations as long as the QoS feed back is within four frame times or less. This feed-back is used by the dynamic rate shaping filter and network based filters to control the data generation of the video or the selection of the signal respectively. In the case of dynamic rate shaping the rate is adjusted while keeping the perceptual quality of the video flow meaningful to the user.

Based on the concept of eras, control messages (see [2] for format and semantics) are forwarded from the receiver-side transport protocol to either virtual source or the source-side transport using reverse path forwarding. We use the term virtual source to represent a network switch that modifies the source flow via filtering. A core-switch [20] [18] were flows are filtered is always considered to be a virtual source for one or more receivers. The WFS mechanism updates the advertised rate as the control messages traverse the switch on the reserve path to the source or virtual source. Therefore any switch can adjust the flow's advertised rate before the source or virtual source receives the rate based control message. The source-side transport hands the measured delay and aggregate bandwidth off (B_{flow}) to the dynamic rate shaping filter.

DQM maintains *flow state* at each end-system and switch that a flow traverses. Flow state is updated by the WFS algorithm and the rate-based flow control mechanism. Flow state maintained in the network constitutes: (i) capacity (viz. B_{avail}, B_{guar}, B_{resid}); (ii) policy (viz. filterStyle, adaptMode); (iii) flowSpec (viz. BL, E1, E2) ; and (iv) wfsShare (B_{flow}, B_{wfs}, W_{fact}). The end-systems holds an expanded share turple for measured delay, loss and jitter metrics. An admission control test is conducted at each end-system and switch on route to the core for base layer signal. This test simply determines whether there is sufficient bandwidth available to guarantee the base layer BL given the current network load:

$$\sum_{j=1}^{N} BL_{(j)} \leq B_{avail} \qquad (3)$$

If the admission control test is successful, WFS determines the additional percentage of the residual bandwidth made available B_{wfs} to meet any enhancement requirements in the flowSpec:

$$B_{wfs(i)} = W_{fact(i)}.(B_{avail} - \sum_{j=1}^{N} BL_{(j)}) \cdot \qquad (4)$$

The WFS rate computation mechanism can causes new B_{wfs} rates to be computed for all adaptive enhancement signals that traverse the output link of a switch; switches are typically non-blocking which means the critical resource are the output links, however, our scheme can be generalised to other switch architecture [19].

6. Dynamic QoS Management

Dynamic QoS management, illustrated in Figure 1, is broadly divided into three "middleware" domains (which are represented as "slices" in the diagram) for end-to-end dynamic QoS management: (i) *sender-oriented DQM,* senders select source filters (i.e. drs, sbr filters) and adaptation modes, and setup flow specifications for video and audio communications. The sender-side transport protocol provides periodic bandwidth and delay assessments to the dynamic rate shaping filter which regulates the source flow. Senders creates QoS groups which announce the quality of service of the flow to receivers via QoS group management [18]; (ii) *network-oriented DQM,* provides the adaptive service to receivers and senders by guaranteeing the base signal and provides weighted fair share using a novel rate base flow control mechanism to switch in discrete or partial enhancements. Network level QoS filters (i.e., sub-signal, hierarch and hybrid-filters) are instantiated based on the user selection, and propagated in the network under the control of filter management[21]; and (iii) *receiver-oriented DQM,* receivers join QoS groups and select the portion of the signal which matches their QoS capability. Receiver selected filters propagate in the network for source and signal selection. In addition, receiver-based QoS filters (i.e., sync-filter and orch-filter) are instantiated by default unless otherwise directed. These filters are used to smooth and synchronise multiple media. The receiver-side transport provides essential bandwidth management for enhancement announcements; delay-jitter and late packet management trading off timeliness against loss.

Based on the receiver supplied scaling policy, QoS adaptors can take remedial actions to scale flows, inform the user of a QoS indication and degradation, fine tune resources and initiate complete end-to-end QoS renegotiation based on a new flowSpec [2]. The QoS scaling policy also allows the user to modify existing QoS filters; and based on this policy, filter management [21] installs new filters at optimal points in the media path.

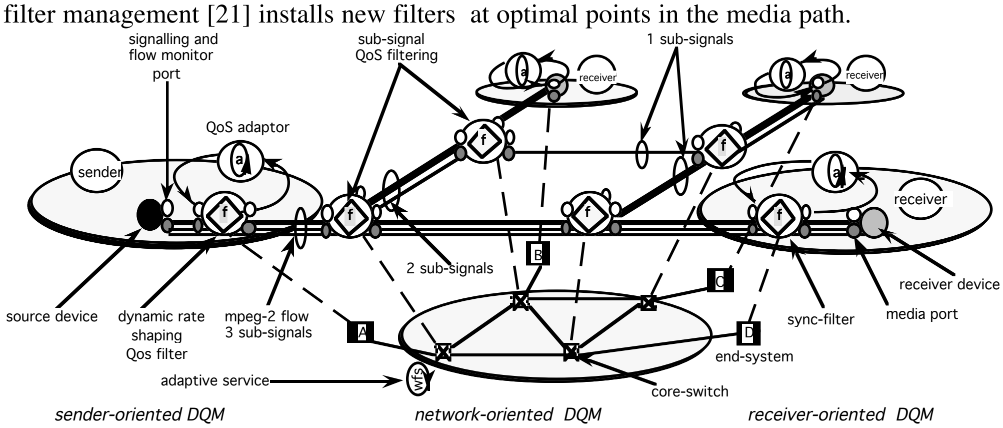

Fig.1: Dynamic QoS Management

Components of a distributed adaptive environment consist of a number of algorithms for QoS group management, flow management, filter management (we neglect to illustrate the interaction of these components with senders and receivers in Figure 1 for reasons clarify) and dynamic QoS management. Each element provides a set of interfaces and methods to manage flows in multimedia networks. Communication between interfaces is based on CORBA running over AAL5/ATM in our distributed environment at Columbia [17] [18] Figure 1 illustrates how adaptive receivers (at nodes B, C and D) and senders (node A) are built on top of multimedia networks which consist of Fore Systems ASX200 ATM switches. The middleware includes the sender, network and receiver-oriented DQM infrastructure and represents the system software components lying in the region between the switching and transmission firmware and specific multimedia applications. CORBA [22] runs on the end-systems and in the ATM switches, providing a seamless object oriented environment where filters, adaptors, QoS group manager, flow monitors and flow managers can propagate [4] [21] throughout the communication system base. In DQM each sub-signal (i.e. base and enhancement layers) can be carried as sub-signal multiplexed onto a single flow, or independently by distinct flows. DQM can handle either case, and leaves it up to the receivers and senders to determine which approach is more suitable.

In Figure 1 a sender at end-system A creates a flow by instantiating a QoS group which announce the characteristics of the flow and its adaptation mode i.e., for MPEG-2 in Table 1 (viz. layer, frame size, bit rate, subjective quality) for BL, E1 and E2 respectively. Receivers join the QoS group. In the example scenario shown in Figure 1 three end-systems join the QoS group created by sender A and "tune" into different parts of the multi-layer signal. The example shows B taking BL the main profile (which constitutes a bandwidth of 0.32 Mbps for VHS perceptual QoS), C taking BL and E1 (which constitutes an aggregate bandwidth of 1.15 Mbps for super VHS perceptual QoS), and D taking the complete signal BL+E1+E2 (which constitutes an aggregate bandwidth of 3 Mbps for laser disc perceptual QoS). In this example the complete signal is multiplexed onto a single flow, therefore, sub-signal selection filters are propagated by filter management [21]. Receive, senders, or any third party or filter management can select, instantiated and modify source, network and receiver-based QoS filters.

6.1 Sender-oriented DQM

A sender-oriented end system architecture illustrated in Figure 2 shows the functions of the sender-side transport that support dynamic QoS management and the interface to a dynamic rate sharing filters. Currently senders can select from two types of shaping filter at the source: drs and sbr. Both of these QoS filters adapt the signal to meet the available bandwidth by keeping the signal meaningful at the receiver or core. The sender-side transport mechanisms includes a QoS adaptor, flow monitor and media scheduler. Bandwidth updates are synchronously received by the flow monitor mechanism from the network as part of the adaptive service (described in section 5). The QoS adaptor is responsible for synchronously informing the drs-filter of the current bandwidth availability (B_{flow}) and measured delay (D_{flow}), and calculating new schedule and deadliness for transport service data units [19]. Media progresses from drs-filter at the TSAP, and is scheduled by the media scheduler to the network at the NSAP based on the calculated deadlines.

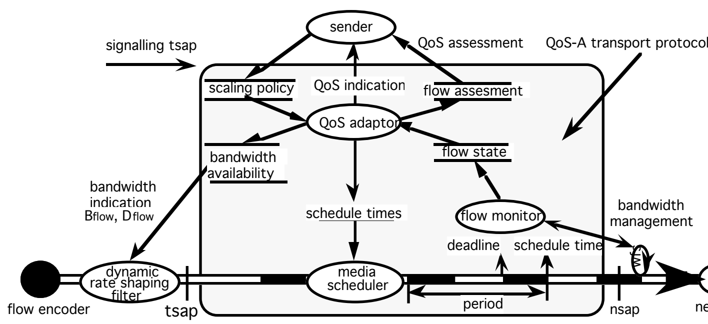

Fig. 2.: Sender-side Transport QoS Mechanisms

The QoS adaptor is responsible for informing the sender of the state of the on-going QoS based on options selected in the QoS scaling policy. As in the receiver side, senders can request periodic updates of bandwidth, delay, loss and jitter. Alternatively, senders can select the QoS monitoring selection [2] to receive periodic updates of all QoS parameters as part of the QoS maintenance operation. Senders, therefore, have the option to discretely select specific QoS parameters, or a set, or all QoS parameters as in the case of QoS assessments [2]. Informing the application of the current state of the resources associated with a specific flow is key in implementing adaptive application in end-systems. In this case the sender is simply used to managed the flow by receiving updates and interacting with the QoS adaptor to adjust the flow e.g., change adaptation mode from continuous to discrete, request more bandwidth for BL, E1 and E2, or change the characteristics of the source filter, etc..

Dynamic rate shaping of compressed digital video [8] (as shown in Figure 2) is a technique to adapt the rate of compressed video bitstreams (MPEG1, MPEG2, H261, as well as JPEG) to dynamically varying rate (and delay) constraints. The approach provides an interface (or filter) between the encoder and the network, with which the encoder's output can be perfectly matched to the network's quality of service characteristics. Although a number of techniques have been developed for the control of live source (which call source bit rate filters) [7] [8] they cannot be used for the transmission of pre-compressed material (e.g., in on-demand video systems). Dynamic rate shaping filters do not require interaction with the encoder and hence are fully applicable to both live and stored video applications. The drs-filter operate directly in the compressed domain of the video signal, manipulating the bitstream so that rate reduction can be effected.

6.2 Receiver-oriented DQM

Receiver-oriented adaptation can be broken down into a number of receiver-side transport functions: (i) delay-jitter management, which calculates flow playout points based on the actual measured delay-jitter from the network; (ii) late-frame management, which monitors late arrivals in relation to the loss metric and the current playout times and takes appropriate action to trade of timeliness and loss; and (iii) bandwidth management, which receives bandwidth indications in the control message portion of the TSDU and adapts the receiver appropriately. In essence the transport protocol "controls" the progress of the media while the receiver "monitors and adapts" to the flow based on the flow specification and the scaling policy

QoS adaptors, which are resident at senders and receivers, are transport-based QoS adaptation managers that arbitrate between the receiver specified QoS and the monitored QoS of the on-going flow. When the transport is in monitoring mode [2] the flow monitor uses an absolute timing method to determine frame receptions times based on timestamps [9] [10]. The flow monitor, as shown in Figure 3, updates the flow state to include these measured reception times statistics. Based on these flow statistics, the sync-filter derives new playout times which are used by the media schedule to adjust the playout point of the flows to the decoding delivery device.

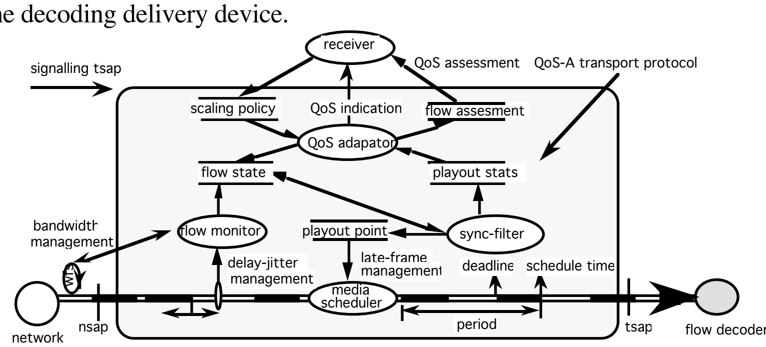

Fig. 3: Receiver-side Transport QoS Mechanisms

Packets that arrive after their expected playout points are discarded by the media scheduler and the late-packet metrics in the playout statistics are updated. The media scheduler simply discards late packets which have missed their schedule time. The media scheduler is based on a split-level scheduler architecture [19] which provides hard deadline guarantees to base layer flows via admission control, and best effort deadlines to the enhancements. Some remedial action may be taken by the QoS adaptor should the loss metric exceed the loss parameter in the flow specification. If the QoS adaptor determines that too many packet losses have occurred over an era, it pushes out the playout time to counter act the late state of packets from the network. Similarly, if loss remains well within the prescribed ranges then the QoS adaptor will automatically and incrementally "pull in" the playout time until loss is detected; see [2] for full details.

Another important receiver-side transport function is bandwidth management. The adaptive service built on the notion of WFS, periodically informs the receiver that more bandwidth is available or announces that the flow is being throttled back. Bandwidth management only covers the enhancement signals of multi-resolution flows. The baseline is not included since resources are guaranteed to the baseline and shared only amongst the enhancements. The announcement of available bandwidth on a flow allows the receiver to take either a full or partial enhancement layer. This depends on whether the flow is in continuous or discrete adaptation mode.

In scalable MPEG2 coding, sources (or virtual sources) produce a base layer signal and up to two enhancement signals, packetize them, and then transmit multiple flows over the network towards the core-switch [18]. The network inevitably introduces some variation in delay of the delivered flows. The receiver depacketizes the flow and then attempts to faithfully play back the complete or partial layers to the decoder. This is accomplished in part by buffering the incoming flow to remove the network-induced jitter and then playing the signal back at some fixed delay offset from the original departure; the term playback point [9] [10] refers to this point in time, which is offset from the original departure time by this fixed delay. The transport protocol utilises sync-filters object for delay-jitter management by calculating playout times of on-going flows based on the user supplied jitter parameter in the flow specification. Sync-filters can also operate on multiple related audio and video flows to provide low-level orchestration management (via the orch-filter).

6.3 Network-oriented DQM

Filter management [4] [21] includes filter placement algorithms which selects the optimum position on a core-based tree [20] to locate filters. Filter placement criteria is based on the current flow topology and the QoS specified by the receiver. As illustrated in Figure 1, QoS filters can be placed at any switch on the media path to meet the needs of receivers. DQM supports a QoS-based multicast scheme where senders and receivers can independently join QoS groups and connect to a core (which is addressed as a core-switch and core-id) [18]. Receivers and senders are said to be "decoupled" at cores. Currently DQM supports three selection filters at the network. These are low complexity and computationally simple filters for selecting sub-signals. Selection filters do not transform the structure of the internal stream i.e. have no knowledge of the format of the encoded flow above differentiating between BL, E1 and E2 sub-signals: (i) *sub-signal filters,* manipulate base and enhancement layers of multi-layer video which are multiplexed on a single flow. The definition of sub-signals is kept general here. Since a flow may be comprised of an anchor and scalable extensions or the I and P pictures of MPEG2's simple profile, or the individual hybrid scalable profile. Sub-signal filters are installed in switches when a receiver joins an on-going session; (ii) *hierarchical-filters,* manipulate base and enhancement layers which are transmitted and received on independent flows in a non multiplexed fashion. In functional terms sub-signal and hierarchical filters can be considered to be equivalent in some cases. In sub-signal filtering one flow characterises the complete signal and in hierarchical-filtering a set of flows characterise the complete signal; and (iii) *hybrid-filters,* combine the benefits of sub-signal and hierarchical filtering techniques to meet the needs for complex sub-signal selection. For example hierarchical filter allows the BL, E1 and E2 to be carried over distinct flows, and the user can accordingly tune into each sub-signal. As an example, the base and enhancement layers of the hybrid scalable MPEG2 flow are each in turn made up of I,P and B pictures at each layer i.e., BL(I,P,B), E1(I,P,B) and E2(I,P,B). Using hybrid-filters the receiver can join the BL QoS group for the main profile and E1 QoS group for the spatial enhancement and then select sub-signals within each profile as needs be (e.g., the I and P pictures of the BL or E1 flows).

7. Conclusion

We have introduced a scheme for the dynamic management of multi-layer flows in heterogeneous multimedia networking environments. Dynamic QoS management (DQM) manipulates and adapts hierarchically coded flows at the end-systems and in the network using a set of scalable objects. The approach is based on three basic concepts: weighted fair share (WFS) service for adaptive flows, the scalable profiles of the MPEG2 standard that can provide discrete adaptation, and dynamic rate shaping algorithms for compressed digital video that provide continuous adaptation. At the present time DQM his being implemented at Lancaster University. The experimental infrastructure at Lancaster is based on 80486 machines running a multimedia enhanced Chorus mirco-kernel [19] using programmable Olivetti Research Limited 4x4 ATM switches. At Columbia we currently using CORBA [17] to propagate selection filters in the network using ASX200 switches. In addition to the implementation we are conducting an extensive simulation study into the feasibility of the adaptive network service for large scale use, and investigating the feasibility of extending the adaptive service concept into the micro-kernel itself [19] for end-to-end QoS support. The results of this phase of the work will be the subject of a forthcoming paper.

8. References

1. Shacham, N, "Multipoint Communication by Hierarchically Encoded Data", *Proc. IEEE INFOCOM'92,* Florence, Italy, Vol.3, pp. 2107-2114, May 1992.
2. Campbell, A., Coulson, G. and Hutchison, D., "A Quality of Service Architecture", *ACM Computer Communications Review,* April 1994.
3. Zhang, L., et. al. "RSVP Functional Specification", Working Draft, draft-ietf-rsvp-spec-03.ps, 1995.
4. Pasquale, G., Polyzos, E., Anderson, E., and V. Kompella, "Fitter Propagation in Dissemination Trees: Trading Off Bandwidth and Processing in Continuos Media Networks", *Proc. Forth International Workshop on Network and Operating System Support for Digital Audio and Video*, Lancaster, November, 1993.
5. Delgrossi, L., Halstrinck, C., Henhmann, D.B, Herrtwich R.G, Krone, J., Sandvoss, C., and C. Vogt, "Media Scaling for Audio-visual Communication with the Heidelberg Transport System", *Proc ACM Multimedia'93* Anaheim, USA, August 1993.
6. Hoffman, D., Speer, M. and G. Fernando, "Network Support for Dynamically Scaled Multimedia Data Streams", *Fourth International Workshop on Network and Operating System Support for Digital Audio and Video*, Lancaster, November, 1993.
7. Kanakia, H., Mishra, P., and A. Reibman, "An Adaptive Congestion Control Scheme for Real Time Packet Video Transport", *Proc. ACM SIGCOMM '93,* San Francisco, USA, October 1993.
8. Eleftheriadis, A., and D. Anastassiou, "Meeting Arbitrary QoS Constraints Using Dynamic Rate Shaping of Code Digital Video", *Fifth International Workshop on Network and Operating System Support for Digital Audio and Video*, April 1995.
9. Jeffay K., Stone, D.L., Talley T. and F.D. Smith, "Adaptive, Best Effort Delivery of Digital Audio and Video Across Packet-Switched Networks", *Proc. Third International Workshop on Network and Operating System Support for Digital Audio and Video*, San Diego, November 1992.
10. Shenker, S., Clark, D., and L. Zhang, "A Scheduling Service Model and a Scheduling Architecture for an Integrated Service Packet Network", Working Draft available via anonymous ftp from parcftp.xerox.com: /transient/service-model.ps.Z, September 1993.
11. Lazar, A. A., Pacifici, G. and D. Pendarakis, "Modeling Video Source for Real Time Scheduling", Multimedia Systems, 1994.
12. Kappner, T. and L. Wolf, "Media Scaling in Distributed Multimedia Object Service, *2nd International Workshop on Advanced Teleservices and High Speed Communication Architectures,* Heidelberg, Germany, September 1994.
13. Steenstrup, M., "Fair Share for Resource Allocation", pre-print, December 1992.
14. Tokuda, H., Tobe, Y., Chou, S.T.C. and Moura, J.M.F., "Continuous Media Communication with Dynamic QOS Control Using ARTS with an FDDI Network", *Proc. ACM SIGCOMM '92,* Baltimore, August 1992.
15. H.262 "Information Technology - Generic Coding of Moving Pictures and Associated Audio", Committee Draft, ISO/IEC 13818-2, International Standards Organisation, UK, March 1994.
16. Paek, S., Bocheck, P., and Chang S.-F., "Scalable MPEG-2 Video Servers with Heterogeneous QoS on Parallel Disk Arrays", *Fifth International Workshop on Network and Operating System Support for Digital Audio and Video*, April, 1995.
17. Lazar, A. A., Bhonsle S., Lim, K.S., "A Binding Architecture for Multimedia Networks", Proceedings of COST-237 Conference on Multimedia Transport and Teleservices, Vienna, Austria, November, 1994
18. Aurrecoechea, C., Campbell, A., Hauw, L. and Hisaya Hadama, "A Model for Multicast for the Binding Architecture", Technical Report, Center for Telecommunications Research, Columbia University, 1995.
19. Coulson, G., Campbell, A and P. Robin, "Design of a QoS Controlled ATM Based Communication System in Chorus", *IEEE Journal of Selected Areas in Communications (JSAC),* Special Issue on ATM LANs: Implementation and Experiences with Emerging Technology, 1995 (to appear)
20. Ballardie, T., Francis, P. and Jon Crowcroft, "Core Based Tree (CBT) An Architecture for Scalable Inter-Domain Multicast Routing", Proc. ACM SIGCOMM '93, San Francisco, October 1993.
21. Yeadon, N., Garcia, F., Campbell, A and D. Hutchison, "QoS Adaptation and Flow Filtering in ATM Networks", *2nd International Workshop on Advanced Teleservices and High Speed Communication Architectures,* Heidelberg, Germany, 28th September 1994.
22. OMG, "The Common Object Request Broker: Architecture & Specification, Rev 1.3., December 1993.