Video coding and delivery challenges for next generation IPTV

S Appleby, B Crabtree, R Jeffery, P Mulroy and M Nilsson

Current generation, large-scale, Internet protocol television (IPTV) systems borrow heavily from the broadcast industry, which makes a number of delivery assumptions that do not apply to IP networks. Consequently we can perceive major improvements if we better match the delivery of IPTV services with the underlying network transport. We can expect next generation IPTV systems to adapt video streams dynamically to maximise throughput, allow constant quality delivery, and degrade gracefully in congested networks. This paper outlines the challenges in optimising IPTV delivery and the contribution that BT's research has made to overcoming some of these.

1. Introduction

We are at the beginning of a very significant new media industry. From the embryonic Internet protocol television (IPTV) industry will spawn a serious rival to the traditional broadcast entertainment industry.

The technology that delivers IPTV content is very different to that used in broadcast. It is this difference in the underlying delivery mechanism, and particularly the ability to offer a unique delivery stream, and hence potentially unique content, to each customer that will make the mature IPTV industry very different to the current broadcast industry. During IPTV's infancy, we expect that (some of the) technology, services and business models will be borrowed directly from the broadcast television industry. However, the use of IP for media delivery will open up new possibilities, which will materialise as the IPTV industry matures to exploit the strengths of the delivery technology.

In the broadcast industry, business models at the various points in the content production and distribution chain are to a large extent determined by the very limited bandwidth of the radio spectrum. Costs of content production for broadcast are typically very high, but are affordable by virtue of being shared by large numbers of viewers. There is fierce competition for eyeballs within the relatively small number of channels available in the broadcast spectrum.

IPTV is different. By definition, content is delivered in IP packets over a data network. Each receiver will receive their own IP packets. The IP network enables unicast content delivery. It is this feature that has the potential to change the economics of the IPTV media industry drastically, relative to the traditional broadcast industry. It will be technically feasible to make content available which only has a very small audience compared with normal broadcast audiences. We expect IPTV to push well into the tail of the content demand distribution (see Fig 1) [1—3].

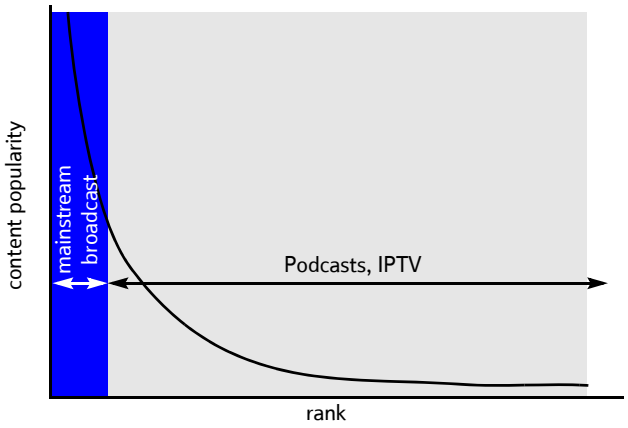

Fig 1 The well-known Zipf curve of consumer demand.

We can see this trend already on the Internet. Podcasts have demonstrated the demand for a very wide variety of content. An interesting feature of the Podcast approach is that content is directly downloaded by consumers — thus much of the broadcaster's role is bypassed.

For the IPTV industry to extend profitably into the tail of the consumer demand curve, dramatic cost

reduction is essential at all stages, from content production, through uploading, ingestion and hosting, to content discovery, and final delivery. In the first generation IPTV systems, much of the video technology will be adopted from the broadcast industry. This results in a higher delivery cost than ultimately necessary or sustainable in the distribution tail [4].

In this paper, we examine some of the techniques than can be employed to optimise the use of network resources to stream the highest quality video streams over IP networks. We will particularly focus on the techniques that could be used for unicast video streaming, as this is will form the essential part of any IPTV architecture, and is the most demanding in terms of network resources.

We will continue by explaining some of the concepts behind video coding and streaming (or broadcast) before describing some of the more advanced techniques that can be adopted in a unicast IPTV scenario.

2. Video coding and streaming

Typically, in the digital broadcast industry, content is received by a broadcaster in electronic form (e.g. digital Betacam tape) [5]. This format is inappropriate for delivery to the end user as it will need further compression to match the capabilities of the channel and the decoding abilities of the receiver.

Once the audio, video, subtitles, etc, have been individually compressed, they are combined (multiplexed) together into a single bitstream. In a broadcast system, there is normally a two-level multiplex which contains multiple programmes [5, 6].

Typically, the bit rate of a multiplexed bitstream is constant. This means that the sum of the individual elements plus the multiplexing overhead should equal the bit rate of the channel. It is, however, not necessary for each component of the multiplex to have a constant bit rate.

Statistical multiplexing [6] works by taking a number of video sources and allocating a total bandwidth to the group. The statistical independence of each video source means that those parts which are difficult to compress are unlikely to overlap in time, so some variability from the nominal bit rate can be allowed.

2.1 Constant and variable bit-rate encoding

When a raw video sequence is to be compressed, the encoder can typically operate in one of two modes — 'constant quality' or 'constant bit rate'.

Constant quality is the simpler mode. In this case, the encoding parameters (primarily the quantiser step size) are kept constant during encoding. This results in an approximately constant quality of video.

Since, at a constant quality, some pictures will compress much better than others, constant quality means that there will be a tremendous variation in the number of bits required to represent each picture. It is difficult to manage the transmission of such a variable bit-rate stream, so encoders will typically be operated in a constant bit-rate mode instead.

In constant bit-rate mode, the encoder will attempt to stay close to some target bit rate by adjusting the quality of the encoding — more difficult sequences being encoded at lower quality. The client buffer will be capable of smoothing the bit rate to some extent (at the expense of increased latency) and therefore some variation in the number of bits per picture can be tolerated while still maintaining a constant channel bit rate.

If the encoder does not maintain a sufficiently constant bit rate such that the client buffer can smooth the variations, then the average bit rate of the video will have to be set lower than the (video part of the) channel bit rate in order to prevent the client buffer from underflowing.

The variation in quality caused by constant bit-rate encoding can be quite a disturbing artefact. The extreme case of trying to assign a constant number of bits to each picture is never used, since this produces a completely unacceptable variation in quality from one picture to the next. From the user's perspective, constant quality encoding is optimal. The challenge, then, is to manage the distribution of content of widely varying bit rate [7].

By way of an example of just how variable the bit-rate requirements of constant quality video content are, Fig 2 shows the change in bit rate over a standard definition video sequence approximately 1.5 min long. The solid line shows the short-term bit-rate variation and the dotted line shows the variation when averaged over a window that might approximate the length of the client buffer.

3. Optimised video compression

In this section we will focus on the video compression itself, and techniques that we are investigating for optimising the trade-off between delivery rate and the user's 'quality of experience' (QoE).

For practical reasons, a very simple quality metric is normally used called peak signal-to-noise ratio (PSNR).

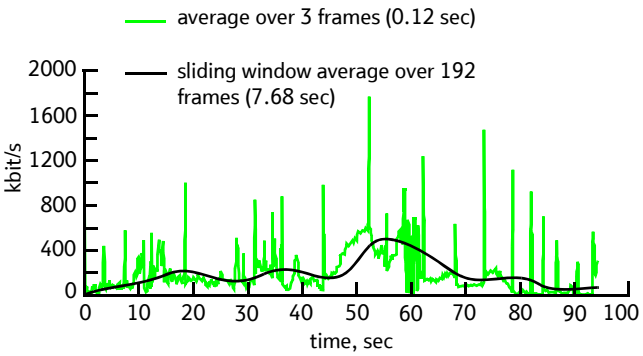

Fig 2 Bit rate as a function of time, averaged over a 3-frame group of pictures and a 192-frame sliding window.

This is based on the mean of the squares of the error introduced by the video being encoded, then decoded.

PSNR is good as a practical metric, but it only provides a very approximate indication of the quality that would be perceived by a person watching a video sequence that had been encoded and decoded. Indeed, there are various straightforward distortions that can be added to a video sequence that produce poor values of PSNR, yet are not perceived as having any significant effect on the video quality. For example adding a small amount of uncorrelated noise to a whole sequence will tend to give a poorer PSNR value, and yet not reduce the perceived quality significantly, whereas adding specific correlated artefacts (e.g. extra edges) will have a big effect perceptually.

The perceptual effect of distortion is not only dependent on the nature of the distortion, but also on the ability of the underlying video material to mask the distortion. Generally, very 'busy' areas will mask distortions better than smooth areas.

BT has developed a model of the human perception of the quality of video sequences which matches the actual perceived qualities extremely well. This model has been standardised by the ITU-T [8]. It takes many factors into account, such as the introduction or removal of edges, the effect of distorting the spatial frequency spectrum, etc.

As part of our research, we are seeking to combine this perceptual model with a model of visual attention in order to ensure that the limited bandwidth is used in the most effective manner. We have already shown that it is possible to achieve significant overall bit-rate reductions with virtually no effect on perceived quality [9]. As an illustration of this, Fig 3 shows an example of a picture encoded such that the parameters of the encoder vary across the picture, such that the quality is adjusted to be lower in regions of low visual attention. The overall result is a saving of around 30% in bandwidth with virtually no effect on perceived quality.

Fig 3 Use of eyegaze information to drive the encoding process. The lines show regions of constant encoding quality and the circles show regions of visual attention. The result of reducing coding quality in regions of low visual attention can be around 30% with no loss of perceived quality.

One way of applying visual attention techniques could be through the latest developments in the H.264 video standard. The joint video team of ITU-T SG16 and ISO-MPEG [10], which developed the H.264 Advanced Video Coding standard, is currently developing scalable extensions of the technology, where the encoded video bitstream consists of a hierarchy of layers, each layer building on the lower layers, to enhance resolution, either spatially or temporally, or reduce distortion. This extension to the H.264 standard is expected to support region of interest coding in scalable layers, allowing, for example, an enhancement layer to provide spatial enhancement just for a particular region of picture, which could be selected by consideration of human perception. Scalable coding also naturally supports adaptive video encoding, the need for which is discussed later.

4. Video over IP

4.1 Best-effort and bit-rate guaranteed delivery

In a broadcast scenario, where several programmes are multiplexed into a single bitstream channel, some trade-off between the ideal, constant quality encoding and more manageable constant bit-rate encoding can be achieved by allowing one programme to 'borrow' bandwidth from others, using, for example, statistical multiplexing.

When delivering video over packet networks, the situation is somewhat different. The bit rate that can be sustained by the physical network is normally constant. However, this is shared by so many services, that, when

left to best-effort delivery, each service receives a varying amount of the physically available bit rate.

So, if streamed video traffic is left to compete with other traffic for bandwidth, the available bit rate for any programme stream will be unknown and time-dependent. For this reason it may be desirable to add bandwidth assurance to the network infrastructure, so that streamed media can be certain of getting the (minimum) bandwidth it needs.

This is probably the most viable approach for first generation IPTV, but will result in very inefficient use of network resources since either some of each video sequence will be encoded at much higher quality than necessary (based on constant bit-rate encoding), or the bandwidth that has been reserved will not be fully utilised. Since there will only be one programme per bandwidth-assured channel (in current proposals), there is no possibility of using statistical multiplexing to ameliorate the situation.

4.2 Adaptive bit-rate streaming

As an alternative to either constant quality or constant bit-rate encoding, suppose we could allow the bit rate to adapt to network conditions. This eliminates the need to choose the encoding policy at the time of encoding.

BT has developed a system called Fastnets [11], which exploits the normal congestion feedback implicit in TCP transport to control the media bit rate delivered by the server. The streaming server contains a model of the client buffer, so that the server knows to adjust the delivery rate according to the client buffer fullness. Fastnets was designed to stream over mobile networks, where throughput is highly variable. However, the Fastnets approach is equally applicable to IPTV streaming over fixed networks.

Adaptive rate streaming requires the ability to send a sequence of encoded pictures to the client, based on real-time decisions made by the streaming server, i.e. each original picture will be encoded in advance at a number of qualities. The server will have to choose which quality version of a picture can be sent at each point in time based on its estimate of the fullness of the client buffer and its assessment of network conditions.

This creates a number of challenges both for video encoding and for the server. For instance, a video encoder uses previously decoded pictures as references for predictions of subsequent pictures. If, at the time of encoding, we do not know which decoded pictures will be available to the decoder, the server decides this in real time. Therefore, we cannot know which decoded pictures can be used as references when encoding.

There are various options to overcome this problem. Switching from one quality stream to another would be limited to special switching points. At these switching points, the dependency of the video stream on the history of decoded pictures would be restricted.

The technical difficulties of stream switching are similar to those in random access, for playing a video stream from some prescribed point. This would be the case when switching from say, a fast-forward stream back into the normal play stream, or entering a live stream at an arbitrary point.

In a broadcast system, a restricted dependency on history is achieved by periodically encoding pictures as so-called I-pictures (intra-pictures). I-pictures are not predicted from any previously decoded pictures — essentially they are compressed as if they were stand-alone images. Regular I-picture insertion could be used in a unicast IPTV scenario. However, I-pictures are generally encoded much less efficiently than pictures encoded using temporal prediction, and therefore this method of allowing random access and stream switching is best avoided if possible.

In the Extended Profile of H.264, a special encoding of pictures (called switching pictures) is used to allow temporal prediction between streams, i.e. using a decoded picture from one stream as the reference to predict a picture in another stream. This is more efficient than using regular I-pictures. Such an approach is only possible in unicast streams, since the server can send the appropriate type of picture for the client's state.

Figure 4 illustrates this approach to stream switching. In Fig 4, the boxes labelled AP and LP represent different types of switching picture. At each point where the switching pictures occur, the server has the option to send either an AP switching picture, and remain within the same stream, or send an LP switching picture to move to a higher or lower bandwidth, depending on the server's model of the client buffer and its assessment of network conditions.

4.3 Equitable quality streaming

It is always the desire to maximise the user's quality of experience for a given channel capacity. In the case of using rate-adaptation in conjunction with the packet network as a means of multiplexing different programmes, the optimisation problem is a complex and dynamic one. The situation is made worse by the fact that in a true video on demand (VoD) system, it would be practically difficult to allow the different VoD servers to optimise collectively. This means that each server, and indeed each port on each server, should have an independent optimisation algorithm that

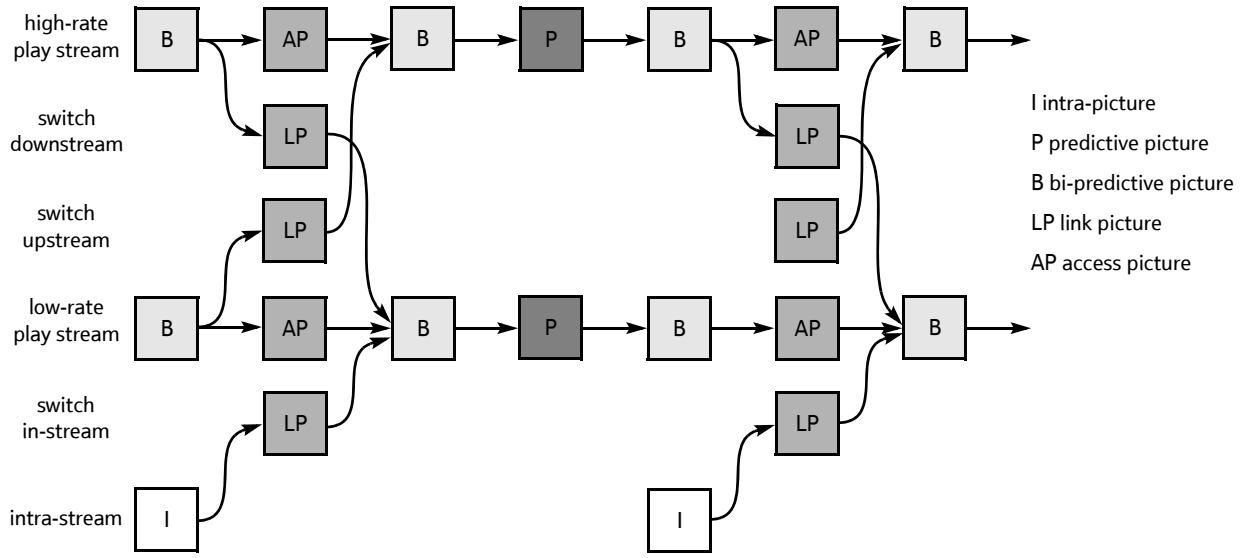

Fig 4 An example of video stream switching.

somehow produces an acceptable result overall. As mentioned before, we would ideally like to deliver each programme stream at constant quality. If we attempt to send a number of programme streams at constant quality, then the total bit-rate requirement would vary. Further, the bandwidth available to the video streams is likely to vary over all time-scales, and crucially, on time-scales longer than the duration of the client buffer.

Therefore, strictly constant quality is not likely to be a practical possibility. The issue then is to use the channel capacity that is available to maximise the viewing experience of the users who share the channel collectively.

Since the bandwidth demands of content vary over time, and is very different from one genre to another, what would the nature of the optimal scheme for allocating bandwidth be?

A transport protocol such as TCP will provide an equitable bandwidth allocation. That is, for a constant total bit rate for the contended network, TCP will stabilise on an approximately equal bit rate for each stream. This is not the optimal solution, since we require an approximately equal quality (of user experience) for each stream. Since the bit rate required for delivery of content at fixed quality varies over time and from one stream to another, the priority of any individual video stream must correspondingly be allowed to vary both over time and from one stream to another.

For instance, if one customer wishes to watch fast-moving sport content, they will require a much higher bit rate than another customer watching slow-moving, low-detail content. In order to ensure that both customers get the same quality of experience, the bandwidth allocation in the distribution network will be very different for these two customers.

This is still an open area of research and the solution is likely to involve dynamic prioritisation at the IP level.

5. Conclusions

We have argued that the nascent IPTV industry will be qualitatively quite different in nature to the existing broadcast industry. This opens many new questions and research challenges, from the roles of the different participants in the industry, through the kind of applications and content that it will be based on, down to the mechanics of delivering the media and services. Many aspects of the IPTV industry will be different in some very significant ways to the current broadcast industry.

Among all the possible areas of research, this paper has focused on the mechanics of compressing and streaming video, which will be at the core of any IPTV service. We have discussed the various optimisation problems in video delivery and shown that the solutions are very different to their over-air broadcast counterparts. The technology for mass IPTV systems is only just becoming a reality, and there is much research left to do.

References

1 Zipf G K: 'Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology', Addison-Wesley, Cambridge, MA (1949); 2nd edition Hafner, New York (1965).

2 Ovum: 'Telcos Need to Keep an Eye on the 'Long Tail'', Ovum Euro View Telecoms (29 March 2006).

3 Anderson C: 'The Long Tail', Wired Magazine, 12, No 10 (October 2004) — http://www.wired.com/wired/archive/12.10/tail.html

4 van Tassel J: 'Digital TV over broadband — Harvesting Bandwidth', Focal Press (2001).

5 Cameron P: 'VTR Technology', in Tozer E P J (Ed): 'Broadcast Engineer's Reference Book', pp 457—474, Elsevier (2004).

6 Tozer E P J: 'MPEG-2', in Tozer E P J (Ed): 'Broadcast Engineer's Reference Book', pp 299—316, Elsevier (2004).

7 Nilsson M, Key P and MacFadyen R: 'VBR Video, Policing and Dimensioning', International Workshop on Audio-Visual Services over Packet Networks (September 1997).

8 ITU-T Recommendation H.144: 'Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full-Reference', (2004).

9 Crabtree B: 'Video Compression Using Focus of Attention', to be published in 'Picture Coding Symposium', (April 2006).

10 ISO/IEC 13818-1: 'Information Technology — Generic Coding of Moving Pictures and Associated Audio: Systems', ITU Recommendation H.222.0 (2000).

11 Nilsson M, Turnbull R, Jebb T and Walker M: 'Multimedia Streaming Over IP', IEE Visual Engineering Conference (2003) — http://technology-standards.intra.btexact.com/dms/pages/itu-t/rec/j/T-REC-J.144.html

Steve Appleby currently heads the Video Coding and Streaming Research team in BT at Adastral Park, focusing on the topics covered in this paper.

Since joining BT in 1983, he has worked on a wide variety of research topics, including language translation technology, mobile agents, statistical population modelling, signal processing, symbolic coding and acoustic imaging for underground plant detection.

Barry Crabtree joined BT in 1980 after completing a Physics degree at Bath University. After a number of years in software development he moved into the field of artificial intelligence (AI), gaining a PhD, and helped begin BT's systems in automated diagnosis and workforce management. He moved on to work in distributed AI, optimisation problems and adaptive systems, setting up the first international conference on Intelligent Agents (PAAM), and was seconded to BT's North America office to help build their agents research team. He spent some years working on developing personalised systems and worked closely with MIT's software agents group. After a spell as the CTO of a start-up company in BT's incubator, focusing on delivering personalised information on mobile devices, he moved back into BT. There he concentrated on developing 3-D applications before moving to work on digital rights management.

Richard Jeffery joined BT in 1991 after studying spatial and 3D design at KIAD.

For the past five years he has been working within the Advanced Technology Centre at Adastral Park, in the video/audio encoding, transmission, storage and management fields.

His current project work includes intranet.tv product development, and future user interface design for streaming-media applications.

Patrick Mulroy joined BT in 1992 working on video coding algorithms and standards. He was active in both H.263 and MPEG-4 development, implementing part of the H.263 reference software and representing BT at ISO/MPEG until 1997. He also researched object segmentation and tracking and holds a PhD in Content Based Video Coding from Dublin City University. In 1997 he transferred to a group researching network and mobile computing using Java technology. Here he led BT's participation in a Eurescom project researching network computing services on mobile devices. In 2000 he left BT to work in the wireless technology practice of PA Consulting. Here he developed and patented ideas relating to the control of a real-time video codec over UMTS. He also implemented application and TCP/IP driver code for a DSP-based software defined radio interference test system which implemented multiple 3G wireless standards. In 2004 he returned to BT, this time to the Broadband Applications Research Centre, and has continued research into video coding and systems. He is a Chartered Engineer and a Member of the IET.

Mike Nilsson joined BT's Visual Telecommunications Division in 1988 working on H.320 videoconferencing systems. He has represented BT in ITU-T and MPEG video compression standardisation meetings, contributing to the development of H.263 and H.264, as well as MPEG 1, 2 and 4. He has been the editor of ITU-T H.245, the multimedia control protocol used in H.323 and H.324, since its beginning in 1994. In 1996 he represented the ATM Forum as an ambassador on the topic of video transmission over ATM.

He was a major contributor to BT's Fastnets world leading streaming technology for mobile networks, which has been recognised by winning BT's Sir Alan Rudge medal for innovation in 2002, and the telecommunications category of the IEEE Innovation in Engineering Awards in 2005.

He is currently working in the Broadband Applications Research Centre in BT Research and Venturing at Adastral Park on video compression and streaming, and the provision of video services over heterogeneous networks. He is a Chartered Engineer and Member of the IET.