A Unified Architecture for Video Delivery Over the Internet

Alistair Blair1, Gerard Parr1, Philip Morrow1, Bryan Scotney1, Steve Appleby2 and Mike Nilsson2
India-UK Centre of Excellence for Next Generation Networks
1Faculty of Computing and Engineering, University of Ulster, Coleraine, Northern Ireland
2Video Delivery Research, BT Innovate & Design, Adastral Park, Ipswich, England

Abstract - Since 1994 when ABC’s “World News Now,” was first broadcast across the Internet, multimedia dissemination has evolved along distinct and unique paths, resulting in two distribution methods that are commonly used across today's Internet, namely Over-the-Top Television (OTT-TV) and Internet Protocol Television (IPTV). With an increasing amount of video content available online, the future of television is increasingly moving towards the Internet. The barriers to widespread deployment of OTT services are now understood and well on the way to being solved. OTT Television has the potential to transform the face of the television industry by creating new business opportunities for existing stakeholders and facilitating many new entrants. With the quality of Over-the-Top services rapidly approaching that of Quality of Service enabled IPTV services it is proposed that the two could be combined into a unified architecture.

This paper documents the two technologies and how they are used in the distribution of multimedia data, namely video. A newly implemented test bed environment is documented along with further work.

Index Terms—Adaptive Streaming, IPTV, Multicast, Over-the-Top Television. Unified Architecture.

I. INTRODUCTION

The uptake of OTT services and IPTV solutions to deploy video to a plethora of multimedia enabled devices is now widespread and their daily use is increasingly commonplace. Video in its various guises (TV, video on demand and Internet) will account for over 91% of worldwide consumer Internet traffic in 2014 [1], with mobile devices generating 6.3 exabytes of traffic every month by 2015, of which around two thirds will be video [2]. Although both technologies employ IP networks to deliver their content to end users, they have both taken independent routes along their paths of evolution and are both architecturally very different.

The fundamental difference between the two techniques is that OTT services are disseminated to any broadband connected Internet Protocol (IP) enabled device using Digital Subscriber Lines (xDSL), fibre, 802.11 a/b/g/n WiFi, Long Term Evolution (LTE), and Third Generation mobile (3G), using Hyper Text Transfer Protocol (HTTP). This distribution is carried out over an un-managed network (sometimes referred to as the Open Internet) by using increasingly intelligent technology at the end points to give the best possible quality using the IP best effort approach, whereas IPTV dissemination occurs over a managed proprietary network using a variety of protocols e.g. Internet Group Management Protocol (IGMP) v2[3]/v3[4], Real Time Streaming Protocol (RTSP) [5], and Real Time Messaging Protocol (RTMP) [6] to a users set top box or computer. By utilising the underlying network, data streams can be prioritised ensuring that delivery can be guaranteed to meet the strict delivery requirements of video.

Typically a managed model, will show a stakeholder, such as a Telecom Operator e.g. BT, play the roles of IPTV Service Provider, Service Platform Provider and Network Provider, thus ensuring that a high quality end-to-end service can be guaranteed to the end user, whereas in the unmanaged model the role of Service Platform provider can be carried out by an Internet portal [7] e.g. Google. Therefore the ability to provide similar quality services to those offered by the managed model cannot be guaranteed due to the lack of quality of service guarantees in Internet delivery.

From a business point of view OTT requires less capital expenditure (CAPEX) and allows content providers to be more flexible and faster when delivering their content to market. As OTT-TV uses the existing network and doesn’t require any special provisioning, content can easily be delivered to any device, anywhere at anytime.

The remainder of this paper discusses how the two technologies distribute content and the advantages that each offer. The section ends with Adaptive streaming being reviewed in detail. Section three presents work related to Quality of Experience assessment (QoE attempts to measure how a user perceives a piece of video), adaptive streaming and IPTV. Section four introduces a unified architecture that combines IPTV with OTT-TV and discusses some of the challenges that will be faced.

II. BACKGROUND

The Internet now has a global reach and so has become a suitable medium to distribute multimedia content, with video usage predicted to grow exponentially over the next few years [1]. Much research has been carried out in an attempt to solve the problems that using an IP based network present (scalability, network heterogeneity and client diversity). Two methods that have led the way are IPTV and OTT-TV.

A) Internet Protocol Television (IPTV)

IPTV can refer to any television service (linear or video on demand) that can reliably deliver multimedia data that is commonly encoded at a constant rate, over an assured bandwidth channel using a network infrastructure that is run by either the Internet Service Provider (ISP) or network provider. This reservation is straightforward to provision but can be costly to implement, as it requires a combination of bandwidth reservation and admission control, making it necessary to create both technical and commercial relationships between the content and underlying service providers. Available bandwidth is divided among the
accepted user sessions. When a user attempts to initiate a
new session, a request for bandwidth reservation is made, if
enough exists the session is created. This process continues
until this bandwidth is exhausted and sessions begin to be
rejected. The success of multimedia websites such as "YouTube" and "MySpace" has shown that in theory OTT
content can be reliably delivered without the need for
resource reservation or quality of service (QoS) provisions.

B) Over-The-Top Television (OTT-TV)

As OTT systems have no way to utilise layer 3
management to implement bandwidth reservation, packet
prioritisation and overload protection they must use other
methods to deliver content. This lack of QoS provisioning
and the last mile problem has led to the use of adaptive
streaming protocols (e.g. Microsoft Smooth Streaming [8],
Apple HTTP Adaptive Bitrate Streaming [9], Adobe HTTP
Dynamic Streaming [10] and Live Adaptive Streaming by
WideVine [11]) that dynamically monitor end user
bandwidth and video performance and subsequently
optimises video quality by switching between lower or
higher quality streams [12].

Figure 1 gives a representation of how content is
currently distributed. A content asset library stores pre-
recorded media that can be combined with a live source if
required to create a broadcast stream, that is then broadcast
by traditional methods, terrestrial, satellite and cable. The
stream can also be multicast via the IPTV architecture to
consumers’ set top box for viewing using a variety of
protocols. IPTV is able to implement bandwidth reservation
and admission control, however this means that the
implementation of the architecture requires both a large
initial capital expenditure and technical expertise from the
network provider.

The OTT architecture takes existing content from the
asset library and after compression transmits it individually
to IP enabled devices that have adaptive clients on board to
view the content or in the case of Quavlive [13] the adaptive
logic resides server side. The HTTP protocol is used,
however specific content is only available on specific
devices, for example OTT is only available on the IP
enabled devices. The OTT architecture is not reliant on
specific network infrastructure or technical expertise, allowing a broader range of services to create a streaming
media node.

C) Adaptive Streaming

Adaptive Streaming or stream switching is a technique
whereby a piece of content is encoded into multiple copies
of varying quality levels each with different bit rates. These
different files are stored as ‘segments’ or ‘chunks’, normally
ranging between two and five seconds in duration. Chunks
can either be stored physically as a separate file per chunk
or logically as a single file that has an addressing structure
that allows any chunk to be accessed individually. As each
copy is encoded at varying levels of quality, its component
chunks will vary in quality as well. When a streaming
session is initiated the client receives a manifest file that
documents the number of different quality streams available.

During transmission the streaming protocol can assess
network conditions and attempt to optimize quality by
requesting a chunk that is suitable to the current conditions
from the manifest file, the selected chunk is then
downloaded using progressive file download.

Decision-making takes place at the client during run time
and it is this process that ultimately forms an optimal
streaming solution. The client is responsible for monitoring
the current local network condition and reacting to any
changes that occur that could affect transmission time and
ultimately content quality. Depending on these conditions
the client will estimate the transmission time for the next
segment of video. Transmission time will vary depending on
the quality of the video segment and the localised network
conditions therefore the end client must choose a suitable
segment that minimises the probability of late arrival while
aiming to keep the quality level as high as possible [12]. The
previously mentioned techniques systems analyse network
conditions at the client side, an exception is Quavlive
Adaptive Streaming [13], where the transmission logic is
implemented server side providing a solution whereby only
the server transmission software has to be implemented.

The previous section has documented how in the case of
IPTV, bandwidth reservation and admission control need to
be available to guarantee the Quality of Service
requirements of video whereas OTT-TV, is able to deliver
video by adapting the video to meet the available bandwidth
and the best effort nature of the IP network.

This video adaptation may lead to a case where the
traditional relationship and Service Level Agreements
(SLA’s) between a network operator and content service
providers will change as costs associated with content
distribution will no longer be accurately represented by bit
rate or byte volume. Content providers may have to move
towards service level agreements based on perceived
quality, where the network provider may agree to deliver
content to meet specified constraints on perceptual quality.
Such constraints could be defined either by parameters
output by automatic quality measuring equipment or by
some reference encoded video content, which could, in the
case of disputes, be compared by human observers with the
quality of video actually delivered, to determine whether the
service level agreement had been met.

This kind of SLA leads to problems, as with best effort
Internet as the service isn’t guaranteed, however on a QoS
enabled service then the network operator is exposed to risk.
This has lead to a great deal of research taking place into the
perceived quality of video and ways that it can be measured
[14].

The remainder of the document discusses some of the
research that has taken place into some of the topics that are
key to video distribution, bandwidth management, perceived
quality of experience and video encoding.

III. RELATED WORK

Initial attempts to distribute multimedia content were
based around IP multicast, which is implemented at the IP
layer with the router taking responsibility for duplicating
packets and forwarding them to the relevant destinations. To
show interest in a particular piece of content an end-users
client simply sends an IGMP join message upstream to
initiate transmission of the new source [15]. Despite the fact that IP multicast maximises utilisation of the network resources, deployment by ISP networks across the Internet has been slow [16]-[18].

The drawbacks of IP multicast and the strict requirements imposed by multimedia data (jitter of 50ms, packet loss of $10^{-6}$, one way delay of 200ms and security) have caused researchers to look at different technologies for distributing multimedia data. The majority of these solutions aim to provide multicast functionality at higher levels of the network stack. This approach is commonly referred to as Application Layer Multicast (ALM) [19], since most of the operations take place at the application layer [20].

The ALM service is implemented by network endpoints. Each end point replicates data when necessary and forwards it to the required destination. Various problems exist, for example a link may transmit duplicate data packets, the path along which the data is transmitted may not be optimal and data lose may occur as hosts leave the network. This requires that “an elaborate failure detection and recovery scheme must be provided alongside the main method to deal with this problem” [21]. ALM has been further refined with different approaches being suggested that optimise the routing of packets within the overlay network and also by combining IP multicast with overlay multicast in a hybrid approach [7].

Currently the bulk of implemented content delivery systems are based around Content Delivery Networks (CDNs) [22] peer-to-peer (P2P) [23] and hybrid CDN [24] systems that can switch between P2P and CDN based delivery methods [12]. The use of these technologies gives an almost guaranteed distribution of content from a content provider’s server to an edge-server that in turn relays content to an end user. However these end-to-end connections are not provisioned for QoS and are commonly referred to as the ‘last mile’ problem.

Much research has taken place in the areas of content distribution, video encoding techniques, video adaptation and enhancing Quality of Experience measurements, both subjectively and objectively. The main aim of the research has been to fully utilise the resources that are available while offering the end user the best quality experience.

A new subjective quality assessment methodology based on full-length movies is proposed by the authors in [25] and shows that every hour of content playback, only one visible impairment is acceptable to maintain QoE for standard definition video, while users viewing high definition video will accept two visual impairments during a movie without questioning the QoE. In [14] the researchers propose a method to build QoE prediction models using limited subjective data and Machine Learning techniques. The results can then be used to predict an achievable remedy to reach a satisfactory level of QoE.

In an attempt to boost QoE of end users in an IPTV system the authors in [26] propose the use of a transcoder in the presence of congestion to maintain QoE and in [27], a server side modification of the standard Transmission Control Protocol (TCP) rate control algorithm called MulTCP is introduced. Greater control over the aggressiveness of the rate control allows MulTCP to reliably deliver video at differential rates aligned to the demands of the content.

Adaptive streaming is still an important research area and in [28], the authors design and implement a system that can monitor and control quality of service depending on the networks ability to provide enough bandwidth to transmit a piece of media. Implementing scalable video encoding allowed the authors to improve the service quality provided. While the authors in [12] created an intelligent client that uses a set of optimal strategies for chunk selection. However the strategies can only be used in the same network that provided the characteristics for the Markov Decision Process (MDP). In an attempt to offer a high level of quality of service to end-users and reduce stress on the network by server load balancing the authors in [19] propose a new streaming architecture that can automatically react to network changes along with client arrival and departure patterns.

A model for the estimation of perceived quality in an IPTV transmission that uses easily measurable video coding and network quality of service parameters is presented in [29]. It is hoped that the model could be combined with
other proposals to monitor and assess channel change time, lip sync and audio quality. The authors in [30] present objective and subjective results of high definition IPTV service and show that Peak Signal to Noise Ratio (PSNR) can be used to track the quality change inflicted by transmission errors.

This research has lead to a belief that IPTV systems and adaptive streaming can be further refined to increase utilisation. Both techniques have their limitations both technically and financially. IPTV providers have spent large sums of money to provision the required infrastructure and expertise while OTT-TV has relatively less financial and technical expertise requirements. The adaptation of HTTP to transmit OTT-TV has led to a reduced need for multiple protocols and so makes distribution much simpler than IPTV. Although IPTV offers a single platform by which users can access media, while the Internet offers a plethora of locations that offer the ability to view an immense variety of video clips, these clips are regularly disjointed and spread across many different sites, resulting in complex navigation for novice Internet users, who would require navigation to be much more simplistic.

This has lead to a belief that Adaptive streaming could be implemented within the IPTV ecosystem in a unified architecture. This would allow clips to be catalogued under a single umbrella offering a simple system that can offer multimedia from many different locations. IPTV providers would be able to offer their service across a wider range of devices while reducing the complexity of the delivery infrastructure. OTT-TV content providers would be able to offer a more reliable higher quality service that would incorporate set top boxes and would allow content to be displayed easily and efficiently on television sets.

The next section describes a new unified architecture that will allow the two techniques to be combined in an attempt to create an IPTV architecture that will allow video to be distributed to a greater selection of IP enabled devices and not be limited to only set top boxes or personal computers.

IV. UNIFIED ARCHITECTURE

The combination and integration of OTT-TV with IPTV will create an offering that allows IPTV providers to expand their service to include a much wider and more diverse audience. This may open up the system to the possibility of increasing revenue from streams such as advertising and customer profiling.

By combining both architectures, television broadcasters will be able to provide many new services, rapidly expanding their existing geographic footprint without the need for significant capital expenditure. The key to combining the two technologies is that end users only see a single sign on service. An OTT platform integrated with subscribers who use hybrid set top boxes will allow content providers to combine existing linear television services with new on demand services, and advertise all this content using a single Electronic Program Guide (EPG). The on demand services can be a mixture of streaming media and existing stored media. Television is mostly made up of a sequential play out of pre-existing assets with very little genuinely live content. There are a number of TV stations that are nothing more than a scheduled play out of pre-existing assets.

Within the context of this unification, the major challenge is the proposed decoupling of the delivery architecture into three sub layers. These layers take responsibility for different aspects of the customer viewing experience, how the consumer imagines the service is being provided, how the content gets to the consumer and finally how the content is displayed to the consumer. The three layers are detailed below:

1) Service Presentation
This layer is responsible for portraying how the service that the consumer is viewing is being provided. For example is the content being viewed as a traditional linear broadcast, requested on-demand or catch up?

2) Delivery
Delivery applies to the methods used to distribute the content from source to the consumer (unicast, multicast, caching). One important aspect of this layer is that a program may not necessarily be delivered by the method presented. For example a linear program or service could be delivery by a combination of on-demand or caching. Also a service offered as on-demand could in actual fact be cached on the consumers set top box.

3) Playout
Playout is characterised by encoding ahead of time regardless of the delivery or service presentation.

Figure. 2 gives a representation of how this proposed unified architecture might look. Assets from the compressed library can be combined with a live source and advertisements to create a single stream that is transmitted using either unicast or multicast and can be presented as either Video on Demand, Linear TV or catch up television.

Bandwidth utilisation should be increased, as streams are suited for the display devices. Multicast facilities will be available allowing for a reduction in core network traffic while improving upon existing Over-the-Top services.

The unified architecture is not without its problems, some of the issues that will have to be faced are:

1. Non-adaptive streams suffer from bandwidth under-utilisation, leading to congestion collapse (eventually degrading video playback quality) and it can also lead to unfair bandwidth allocation as TCP traffic is adaptive

2. The IP multicast model only provides a basic framework for group communications. This may mean having to implement application layer multicast to over come these limitations.

3. Over-the-top multimedia content is sparsely distributed across the Internet and requires a method that allows this media to be incorporated into the system.

4. The Internet is a heterogeneous network comprising of many different devices, all with different compute, storage and display characteristics. A single session on the unified architecture could encompass many of these devices and so must be able to adjust accordingly to fully utilise resources.
5. Traditional service level agreements are no longer suitable, perceived quality will become more apparent as total byte volume will no longer be suitable.

VI. CONCLUSION & FUTURE WORK

IPTV is carrier led and therefore only available to certain content providers. Its closed wall architecture is expensive and geographically limits its coverage. OTT on the other hand, needs no special provisioning and exists on the same principles as the Internet, anyone who holds the rights to a piece of content can create an endpoint and distribute it to a global audience.

The lack of layer 3 management functionality within the technology used to distribute OTT-TV has seen it evolve to utilise the best effort IP network effectively. The use of adaptive streaming has created a distribution method that is highly effective and easily implemented. HTTP protocol traffic can offer effortless video streaming by avoiding NAT and firewall issues while in some instances allowing the client control over the stream. This has created the need for IPTV providers to embrace the technology or risk losing out on a valuable revenue stream.

In future work we shall attempt to combine the two techniques into a single unified architecture, allowing content providers to concentrate on creating new and innovative propositions, rather than being constrained by architecture. The unified architecture will require less technical knowledge and dependence on the protocols that are used by IPTV systems allowing content providers to concentrate on producing new and innovative media that can be displayed across multiple platforms.

As part of on-going work we have designed and implemented an initial Testbed (Figure 3) that is a representation of the proposed unified architecture from Figure 2. The Testbed contains various streaming platforms and features a virtualised machine infrastructure to facilitate rapid deployment.

Initial benchmarking experiments have been conducted to verify the behaviour of Microsoft Adaptive Streaming, Flash Streaming and progressive streaming. Further experiments will be conducted on the use of multicast, bandwidth reservation and admission control to a variety of fixed and mobile clients.

REFERENCES

Figure 3 Initial Test bed Architecture Design.


[22] Hao Yin; Xuening Liu; Geyong Min; Chuang Lin; , "Content delivery networks: a bridge between emerging applications and future IP networks," Network, IEEE , vol.24, no.4, pp.52-56, July-August 2010


