4G americas™

Supporting Wireless Video Growth and Trends

April 2013

TABLE OF CONTENTS

Executive Summary 4

1. Introduction 4

1.1 Overview of whitepaper 4

1.2 Market trends in video delivery/market analysis 5

1.2.1 Usecases for video content delivery 6

1.2.2 Factors contributing towards growing demand for mobile video 7

1.2.3 Trends for streamed video consumption 8

1.2.4 Ecosystem support for the growth in mobile video 11

2. Types of video 12

2.1 Elements of video content and summary of video delivery technologies 12

2.2 Short form content, Long form content 13

2.3 Operator video and OTT video 13

2.4 Best Effort video and Guaranteed Bit Rate video 14

2.5 Video for entertainment and video for real time communications 14

2.6 HTTP progressive download 14

2.7 RTP/RTSP delivered video 15

2.8 HTTP Adaptive Streaming (HAS) 16

2.8.1 Comparison between progressive download, streaming and adaptive streaming 18

2.9 Platforms for video delivery 18

3. Challenges for mobile video delivery 19

3.1 Challenges for video delivery due to air interface 19

3.1.1 Spectral efficiency of 3G/4G cellular technologies and impact on mobile video 20

3.1.2 Air interface latency and impact on mobile video 23

3.1.3 Interference and handover issues in 3G/4G cellular technologies 24

3.1.4 Scheduling mechanisms at the air interface 28

3.1.5 Signaling challenges for video 30

3.3 Challenges from the network 32

3.3.1 Backhaul congestion 32

3.3.2 Data plane/control plane encryption 34

3.4 Content based challenges 34

3.4.1 Digital Rights Management (DRM) 34

3.4.2 Entitlement/Content Management Rights 34

4. Video Quality of Experience 35

4.1 Video usage behavior and expectation of experience 35

4.2 Metrics for characterizing video QoE 35

4.3 Behavior of video chunks with TCP/IP 37

4.4 Monitoring of video QoE 40

4.5 Specific attributes that contribute to improved video QoE 42

4.6 New/Upcoming codecs (HEVC) 42

5. Architectures for mobile video delivery 43

5.1 Cellular (HET-NET) and Wi-Fi offload architectures 43

5.1.1 Video users in a heterogeneous network 45

5.2 LTE broadcast architecture 46

5.3 Video delivery architecture 49

5.3.1 Mobile Operators and their migration to IMS 50

5.4 Content management in cloud environment 52

5.4.1 Content management architecture and delivery 52

5.5 Policy management architecture 53

6. Survey of optimization techniques/mitigation strategies 53

6.1 Objectives of wireless video optimizations 53

6.2 End-to-End QoS support in LTE for video delivery 54

6.2.1 LTE QCI and other network QoS techniques 54

6.2.2 LTE guaranteed bit rate bearers 56

6.2.3 Maintaining constant video quality with GBR service 57

6.2.4 Differentiation and fairness with GBR service 57

6.2.5 Limitations of GBR bearer 58

6.2.6 Microbursting 58

6.3 RF specific optimizations for video delivery 58

6.3.1 Guaranteeing minimum video quality on the air interface 59

6.3.2 Broadcast/multicast of video 59

6.4 Non-RF specific optimizations for video delivery 60

6.4.1 DASH (HAS) 60

6.4.2 BIS Scalable Video Codecs (SVC) 60

6.4.3 Transcoding/Transrating/Rate capping 61

6.4.4 Congestion-aware optimization 62

6.4.5 Client optimizations for HAS 62

6.4.6 Content caching in the network 63

6.4.7 Time shifting or sideloading supported by content caching in the UE 64

6.4.8 Video usage behavior 64

7. Recommendations 65

7.1 Recommendations for video application developers for efficient radio resource management 65

7.2 Recommendations on bit rates for various codecs and screen sizes/device types 66

8. Summary and Conclusions 67

List of Acronyms 69

List of References 70

List of Figures 71

List of Tables 72

Acknowledgements 72

EXECUTIVE SUMMARY

Video is increasingly becoming one of the most pervasive technologies in terms of everyday usage, both for entertainment and in the enterprise environments. Mobile video is responsible for a majority of the growth seen in mobile broadband data volume. This white paper presents the expected growth of mobile video based on current trends and user behaviors, describes the various types of video based on content type or delivery strategy, and details the various challenges and solutions to deliver mobile video over a wireless end-to-end network. A multipronged approach to address the various challenges, unique to mobile video, is presented through optimization techniques in the radio and core networks (HTTP Adaptive Streaming, etc.), support of multiple network architectures (HET-NETs, LTE broadcast, content delivery networks), support of client optimizations and through enhanced codecs (HEVC). Metrics for quantifying and attributes that contribute to improved video QoE are described in detail. Recommendations on bit rates for various codecs and screen sizes, and recommendations for video application developers to support the scalable growth of video in mobile networks are made after examining unique aspects specific to mobile video delivery.

1. INTRODUCTION

1.1 OVERVIEW OF WHITEPAPER

New and emerging classes of mobile devices (smartphones, PC data cards, USB modems, consumer devices with embedded wireless, machine-to-machine, etc.) are fostering the explosive growth of wireless data usage and higher-bandwidth data applications and services by public and enterprise users. Users would like to see various kinds of video content delivered to and uploaded from their mobile devices with good quality of experience. With the growth of video traffic on the Internet and end users' expectation of having content delivered with freedom of movement, the reliable delivery of video over wireless networks to mobile end user devices will become essential in the coming years.

Operators are beginning to roll out heterogeneous networks that include base station and core network elements from various 2G, 3G and 4G technologies. Such networks, with macrocells, picocells and femtocells and Wi-Fi offload capabilities help best utilize the limited air interface resources while serving the users with the best Quality of Experience (QoE). By bringing the base station closer to the user, small cells provide valuable radio resources to high bandwidth video application users that are generally low mobility, thus resulting in overall spectral efficiency gains. However, even with the introduction of picocells and femtocells, air interface resource limits will be reached at some point due to increasing demands.

Delivery of wireless video poses particular challenges for the network. A video stream requires a sustained high bandwidth session lasting several minutes. The limited spectrum availability can result in air interface congestion restricting bandwidth to only a few active users in a cell. The congestion problem is compounded by cell handovers because of user mobility and interference from neighboring cell users. These handover and interference issues affect packet throughput, loss and jitter, all of which adversely affect video delivery. Throughput to individual users varies with time and location, based upon cell load and proximity to the cell. Throughput at the cell edge can be significantly lower than near the cell center. The broad range of emerging devices requires flexibility in video delivery to address variations in screen sizes, screen resolution of end user devices, and impact of video usage on battery life. It is therefore critical to minimize wireless congestion due to video traffic, by tailoring the video application to the underlying network resources and device types, while still maximizing end user experience and operating efficiency.

This paper addresses the various aspects of mobile video delivery from operator, vendor and device perspectives. Market trends in video delivery, types of video content, challenges for mobile video delivery and various optimization techniques to mitigate those challenges in the radio and core network are described in detail. Architectures for mobile video delivery including the policy management architecture and content management are discussed in detail. Metrics for measuring video quality are discussed to provide insights into the performance of video in current networks. This paper concludes with recommendations for techniques to deploy technology to support video and for application developers.

The whitepaper is structured as follows:

- Section 1 describes industry trends and drivers that lead to increased use of video over wireless networks. Summary from various industry reports are provided.
- Section 2 addresses different types of video content and use cases which must be examined to help define techniques that ensure robust delivery to the consumer.
- Section 3 focuses the discussion on challenges of video delivery over a wireless access.
- Section 4 addresses video quality of experience including metrics and mechanics of measurements, codec technology and industry direction.
- Section 5 gets into various architectures to support video delivery including HET-NETs, LTE broadcast, and content management and delivery networks.
- Section 6 describes some techniques used in wireless and supporting networks to mitigate the challenge of delivering low latency, high volume video to consumers.
- Section 7 provides recommendations to the reader, including application developers and device manufacturers to help maximize video quality of experience.
- Section 8 summarizes some of the main findings and suggested way forward in the industry to enhance video QoE over wireless networks.

1.2 MARKET TRENDS IN VIDEO DELIVERY/MARKET ANALYSIS

The world of culture including TV, film, video and music is changing. This technology-induced change not only improves people's viewing experience, but it also changes the ways in which consumers conceptualize culture and entertainment. Devices in the wireless market have evolved in the past two years in video capabilities (resolution, rendering abilities), computing power (dual core, quad core) and user interface. These developments coupled with the deployment of advanced wireless networks have led to an increasing adoption of video on wireless devices.

Users like to see various kinds of video content delivered to and uploaded from their mobile devices. Users would also like to participate in bidirectional conversational videos such as mobile-to-mobile video calls, video teleconferences on their mobile devices, etc. New video applications are being spawned at a faster rate that will further cause growth of mobile video. Mobile video traffic is already over 50 percent of mobile data traffic and will account for 66 percent of global network demand by 2017 [1]. The growth in video traffic could soon overwhelm the wireless network resources due to finite and limited spectrum availability.

1.2.1 USECASES FOR VIDEO CONTENT DELIVERY

Video content is delivered for various purposes, as noted below, and will be referenced further in discussing trends in mobile video delivery.

1) Content may be downloaded on the device prior to viewing on the device. Since the viewing of content is separate from the download of content, delays in download of video content may be acceptable to within a limit.

2) Streamed video is driven by availability of content and degree of digital rights management (DRM) in the content. With content being consumed immediately, there are tighter bounds on the delay in delivering the video content.

3) Uplink video delivery can result in sporadic high volume of delivery, in addition to significant downlink video, particularly at sporting events and stadiums when content is shared towards a server.

4) Applications like Facetime are generating traffic with increasing popularity due to device capabilities driving video volume.

The Ericsson Consumer Lab TV and Video Report 2012 [2] shows that smartphone and tablet users spend several hours per week watching video. These will be delivered over cellular and Wi-Fi networks. Several respondents browse the Internet while watching video and TV. The report notes that the use of mobile video is increasing due to increased numbers of social forums and chats. A significant percentage are actually discussing the things they are watching. This kind of multitasking behavior is increasing because of the availability of different easy-to-use mobile devices such as smartphones, laptops and tablets. This trend is expected to grow in the coming years.

Video represents the largest data traffic volume today and continues to increase. As an example, a user on a data plan of around 15 GB consumed on average 2.3 GB per month of online video in Q1 2012, while in Q4 2012 this figure was 3.5 GB per month [3]. According to [3], video traffic has yet to reach its peak. One reason is that improved video quality and increased screen sizes will facilitate looking at longer clips or watching videos for a longer period of time. The increasing availability of content will also result in increased video usage. Generational shifts in viewing habits also contribute to the increased video usage.

Mobile devices, such as the smartphone, are becoming a natural part of the video and TV experience and increasingly important. Many consumers state that their various mobile devices have actually begun replacing their secondary TV screens around the home. While the main TV is still the preferred screen for TV and video consumption, mobile devices are being used more and more, partly due to applications that allow time shifting, e.g. Netflix and Hulu.

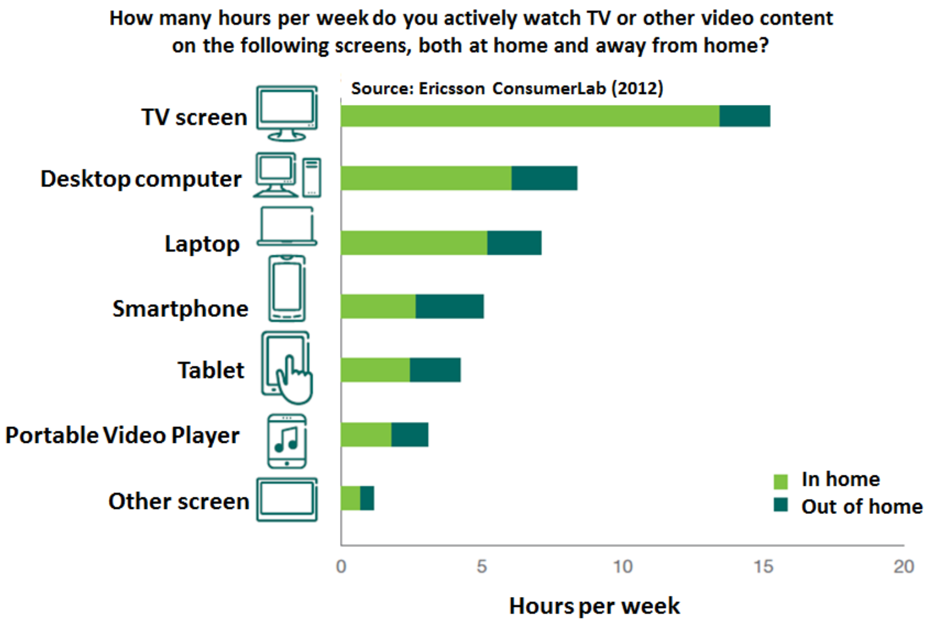


Figure 1: Number of hours per week spent watching TV or Video on various devices [4]

1.2.2 FACTORS CONTRIBUTING TOWARDS GROWING DEMAND FOR MOBILE VIDEO

Factors that have contributed to the growing demand for mobile video include the following:

1) With the growing capacity of the wireless network, through the introduction of HSPA+ and LTE, it is possible to select and immediately view the content on demand. This is particularly popular for short form content such as YouTube clips.
2) Social media makes it easy to share links that result in further increasing the demand. Mobile devices are even becoming a partial replacement for linear TV in homes with catch up TV services such as Hulu and iPlayer as well as TVEverywhere. Some recent services even link the viewing of linear TV on a big screen with apps on tablets to get more information or to enable interaction with the show.
3) Video streaming over the network is growing in popularity. Many mobile tablets and smartphones have front facing cameras making them an ideal video phone terminal. Apple provides Facetime enabling video calls between users of its devices. Skype has been extended to provide video as well as voice. Video streaming over the network not only imposes the sustained bandwidth requirements of on-demand video but also requires low latency.

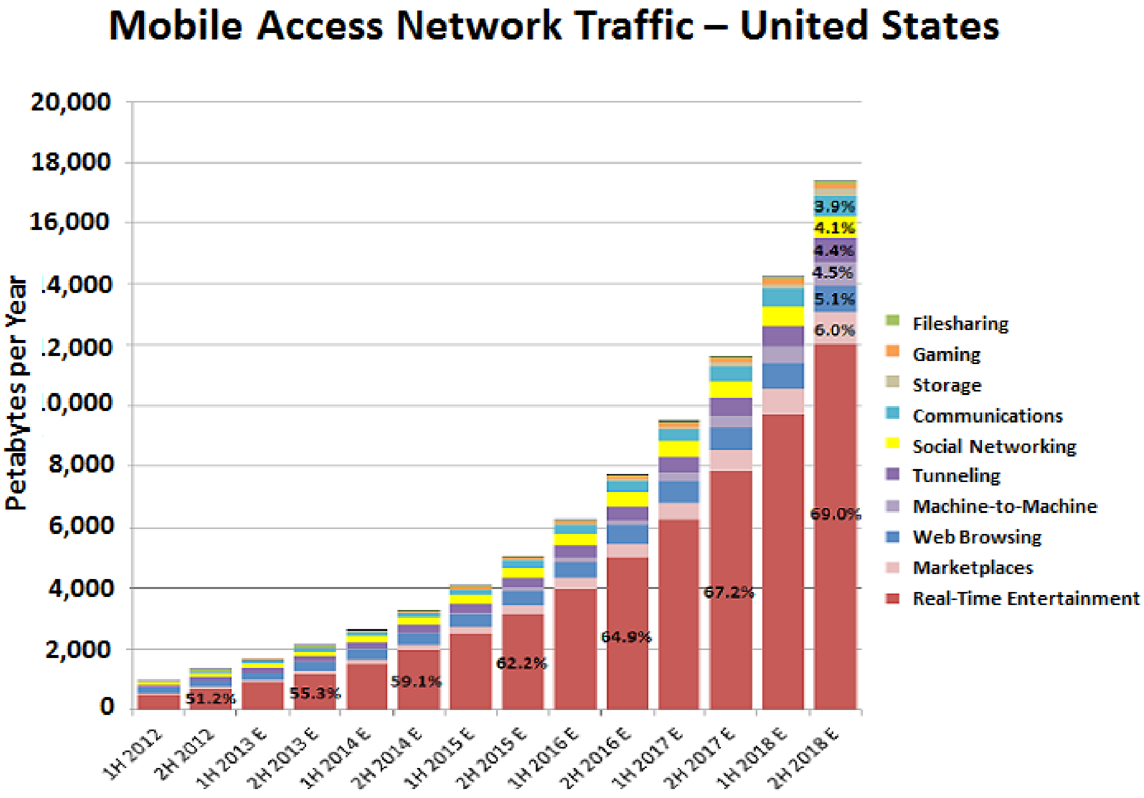


Figure 2: Projection of mobile traffic trends in the United States [5]

Figure 2 shows a projection of mobile traffic trends in the United States from the Sandvine report [5]. It indicates that real-time entertainment comprising of video and audio streaming applications will grow significantly in the coming years and account for close to 70 percent of mobile usage by 2018.

1.2.3 TRENDS FOR STREAMED VIDEO CONSUMPTION

Streamed video consumption is growing significantly as summarized below based on various reports, e.g. Sandvine report [5]. From Figure 3, YouTube video viewing constitutes a significant component of real-time entertainment traffic, expressed as a percentage of downlink mobile traffic in 2012.

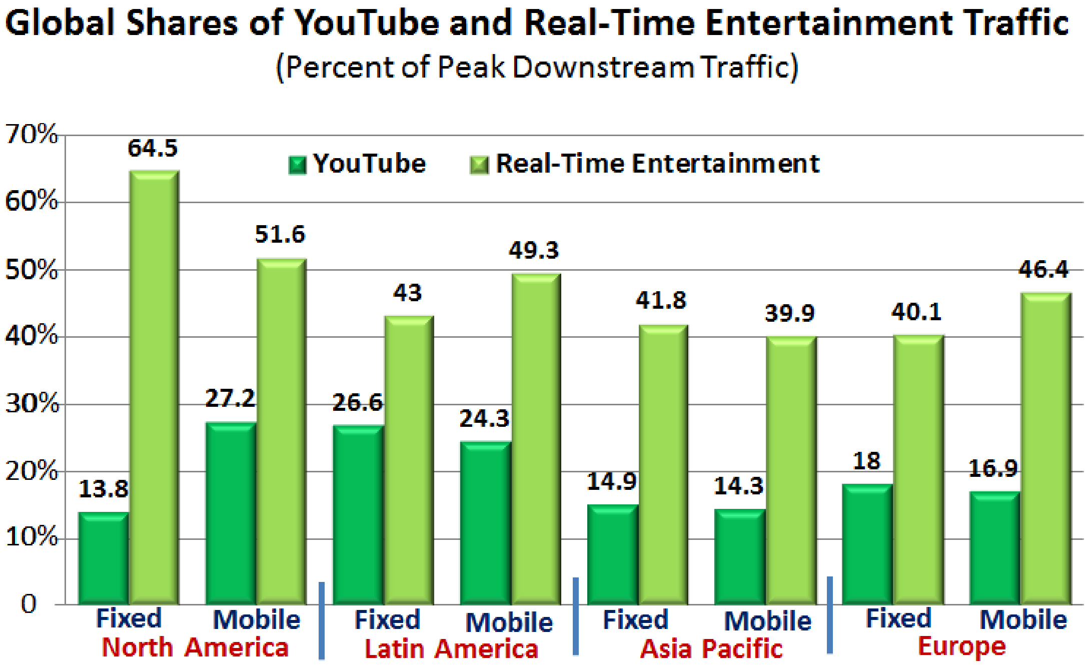


Figure 3: Global Shares of YouTube and Real-Time Entertainment Traffic [5]

During peak traffic periods, streaming video entertainment traffic is by far the most dominant traffic category, accounting for nearly half of all bytes sent and received on the network, continuing a trend that is expected into the foreseeable future. Real-time entertainment is defined as applications and protocols that allow for on-demand entertainment that is consumed (viewed or heard) as it arrives on the device. Examples include streamed or buffered audio and video (RTSP, RTP, RTMP, Flash, MPEG). Both HTTP video and P2P TV are types of online video where the user watches the film while it is being downloaded (streamed). In the case of P2P file sharing the user first downloads the entire file, most often a movie, and watches it offline. Examples of applications that use HTTP include YouTube, Hulu, Netflix, and BBC iPlayer. Examples of P2P TV include PPStream and Octoshape. Octoshape is a service that lets a streaming video site use your Internet connection to stream video to other people.

There is an increasing trend in consumption of streaming video during live events, much of which is on mobile devices. Spikes in video usage are driven by live events and breaking news. Figure 4 shows usage data captured from FLOTV in 2009. It shows that average minutes of daily viewing peaked around live events such as the NCAA Tournament and college football. Breaking news such as inauguration day and Michael Jackson's memorial drove high usage.

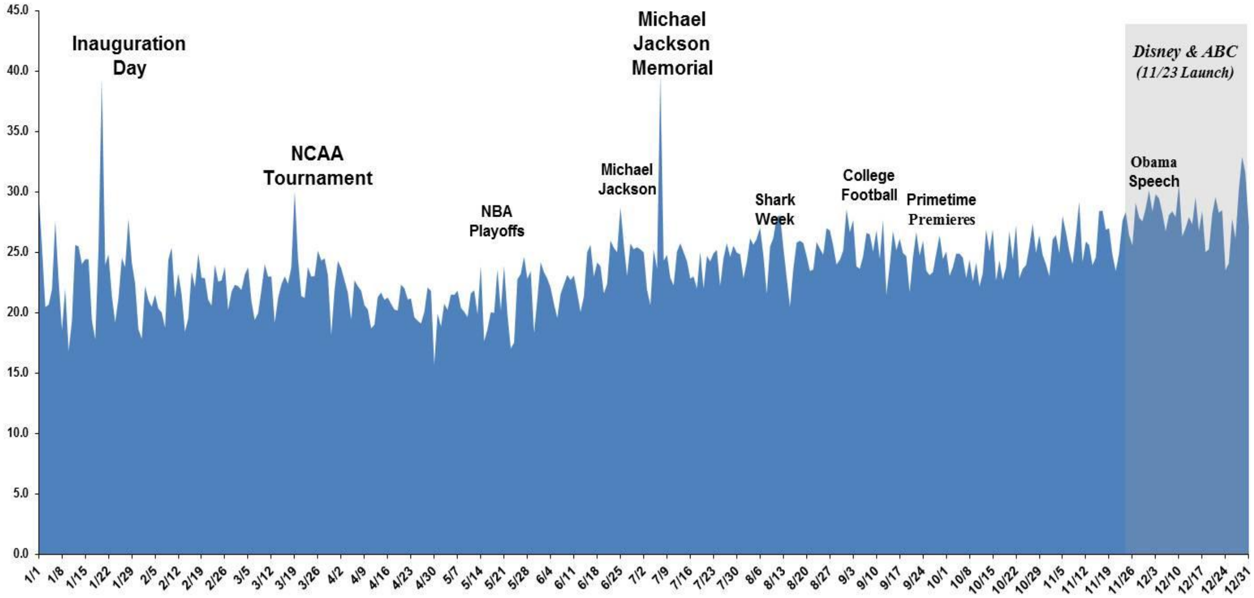


Figure 4: Streaming Video usage by users on different days in 2009 [FLO TV Incorporated]

During the first week of the 2012 Olympic Games, streaming traffic accounted for almost 22 percent of total network traffic and between 8-12 percent of network traffic in the U.S. at its peak daily levels. At the Feb 4, 2013 Super Bowl games, official statistics for live stream of XLVII on CBSSports.com and NFL.com [6] indicated the following statistics:

- Unique viewers: 3 million
- Total minutes streamed: 114.4 million
- Live video streams: 10 million
- Engagement: 38 minutes per viewer

North America continues to lead in adoption of this traffic category, with almost two-thirds of downstream traffic during peak period being streaming audio or video. The dominance of video streaming is due in large part to the continued uptake of video-on-demand, which now accounts for 33 percent of peak period downstream Internet traffic, some of which is mobile video traffic. In July 2012, an important threshold was crossed with 1 billion hours of video streamed in a single month. It is expected that video advertisement will become more and more prevalent and consume an increasingly higher portion of data usage in the future.

In Europe, a major news provider offers regional and international catch-up TV service subscriptions. On demand viewing of most programs broadcast on major TV channels for up to seven days after first airing is allowed. Over 600 devices are supported from PCs, game consoles, connected TVs, smartphones and tablets. The December 2012 statistics [7] show that 31 percent of iPlayer viewings are on smartphones (28 million views per month) and tablets (26 million views per month). Most of this will be streamed via Wi-Fi because of the low penetration of LTE in the UK. This shows the acceptance of people to use these small screens to watch TV. The expectation is that mobile video streaming will increase with more widespread availability of technologies like LTE.

Ooyala recently released the 2012 Video Index [8] that reveals key online video trends. Ooyala measures the viewing habits of nearly 200 million unique viewers in 130 countries every month. The report reflects the anonymous online video metrics of various publishers. Their data shows live streaming is the new norm. Viewers watch live video longer on all devices. The share of tablet video viewing more than doubled in 2012, as mobile, social and video converged on a single device. Smart TVs and gaming consoles continue to change the way people watch TV. Xbox users spend more time watching video than playing games. New findings in this Video Index show how viewing patterns change seasonally.

In summary, data from different sources consistently indicate a significant uptake of mobile video, be it streaming traffic during live events, real time entertainment and You Tube traffic. This trend is only expected to grow with newer video applications contributing to the mix, and with the introduction of newer devices with larger screen sizes and increased capabilities.

1.2.4 ECOSYSTEM SUPPORT FOR THE GROWTH IN MOBILE VIDEO

To support the growing need for mobile video, multiple facets of the ecosystem have evolved as below:

1) There is an evolution in video delivery mechanisms with fixed, reliable connections (DSL, fiber, etc.) being complemented by Wi-Fi and small cells cellular access. Initially, video was accessed only from fixed access points using static devices like TVs and tethered devices. As mobile broadband becomes more ubiquitous, video is being increasingly accessed at different types of locations, and in crowded locations in dense urban and urban areas.

2) There is a dramatic increase in overall wireless data consumption as bandwidth intensive video is increasingly being consumed on wireless. With the introduction of new wireless technologies including HSPA+ and LTE, video has become mobile and is being viewed at locations such as transport hubs, stadiums and arenas during games and in enterprise locations where users in buildings are accessing their mobile office network.

   With the ability to watch streamed multimedia applications (on-demand streaming, live streaming, etc.) faster on LTE networks, subscribers are likely to watch longer duration videos more frequently, consuming more data than ever before. In some cases, new LTE customers have experienced as much as a 50 percent increase in their monthly broadband usage simply by moving from a 3G to a LTE network.

3) There is increased complexity in playing of video content across devices, as users move across accesses and expect continuity of sustained video quality. Streaming video to mobile devices is increasing. Previously people would pre-load their portable video player with content. The choice of device for access is now increasingly on highly portable devices such as iPads and other tablets that allow for more visually dynamic, image rich content to be displayed along with audio.

4) There is an evolution in rendering quality (Standard Definition [SD], 720p, High Definition [HD], etc.) on devices and content systems, as end user expectations demand increased quality in video content. While video is currently delivered primarily as SD and below, it is expected that users will, over time, demand that video be delivered as HD. Better viewing experience will translate to increased user demand for more video content on their mobile devices.

5) There is an increased expectation on device clients and content servers to adapt to the changing underlying transport capabilities due to the mobility aspects that are introduced by wireless devices.

6) There is an evolution in business models for video as providers find new ways to monetize video content. Videos can be informative, entertaining, educational, and each provide unique opportunities to content and service providers, tool developers, etc. for new business avenues.

This paper addresses the various issues related to the mobile video delivery in the following sections.

2. TYPES OF VIDEO

The mobile video ecosystem can be broadly classified under two major dimensions as follows:

1) Video content types (encoding, container and delivery methodologies).
2) Video delivery strategy and technologies.

This section gives an overview of these two dimensions to help address the full scope of the mobile video experience, including challenges, metrics, architectures and solutions.

2.1 ELEMENTS OF VIDEO CONTENT AND SUMMARY OF VIDEO DELIVERY TECHNOLOGIES

Videos are specialized content that comprise of three basic elements. They are:

1) Moving video component: This section actually contains the moving images that when played in quick succession gives the feeling of watching a continuous video stream. Some popular video codecs used in the industry today are H.264, VP6, VP8, etc.
2) Synchronized audio component: This section has the audio content corresponding to the video being played back. Some popular audio codecs used in the industry are AAC+, AAC, mp3, etc.
3) Video container: Both the video and audio components are wrapped around a media object called a video container. This has additional information about the audio and video components like, the codecs used and the frame offsets. This information is used by the player to play back the video smoothly and also to handle video seek requests from the user. Some popular video containers used in the industry are MPEG4, Flash Video and WebM.

Both visual quality as well as the smoothness of the video playback influences the end-user's quality of experience. The quick start of the video playback is also often considered to influence the quality of experience. Hence different techniques are used to protect the visual quality as well as ensure the smoothness of the video playback in a dynamic network environment, like in mobile networks.

Video content may be classified in various ways depending on their duration, their type of quality, whether they are delivered by the service provider or not, etc. These are described in detail in Section 2.2 through 2.4. There are four major video delivery technologies that have substantial presence in the mobile network, which are summarized below and described in detail in sections 2.5 through 2.9.

- **HTTP Progressive Download (HTTP-PD):** This is a popular video delivery methodology for short form video content. User generated video content publishers like YouTube and Facebook use this mode for their short form content.
- **HTTP Adaptive Streaming (HAS):** This is an upcoming video streaming technique being popularized by the likes of Netflix, for long form content. This technology empowers the clients to determine the video quality based on the dynamics of the network condition. With HAS, the client requests a higher quality video in case of good network condition and lower quality when the network condition is poor.

- **Non-HTTP, Real Time Streaming Protocols (RTSP, PPStream etc.):** These protocols are typically used by certain video content providers to stream live and long form video content to proprietary clients. While HTTP is stateless, RTSP is a stateful protocol.
- **Video file downloads:** In this scenario, the video content is delivered offline and viewed later by the user using generic or proprietary video players.

The following sections describe the various types of video content and the various video content delivery technologies in detail.

2.2 SHORT FORM CONTENT, LONG FORM CONTENT

Video content can be broadly classified into short form and long form content. Videos shorter than 10 minutes are typically classified as short form, while those that are longer than 10 minutes are classified as long form video content.

Short form video content could be anything like video advertisements, user generated videos, product videos, sports highlights, news extracts, etc. Some of this short form content is produced professionally while the majority of them are not. Hence, there is a wide variation in the encoding quality of short form videos. Due to the short time duration involved, video publishers typically use some kind of file download mechanism (progressive download, for example) to deliver this video content. Players are typically designed to play the video as it receives them from the servers. This progressive nature of the client video playback ensures that users don’t have to wait for the complete download of the video before playback begins. Since this is delivered as a simple file download, any bandwidth variation in the network would adversely impact the smooth playback of these videos.

On demand version of TV shows and live telecast of events are the typical examples of long form content. Long form content involves a more consistent viewing form. Since most of this content is produced professionally and delivered to paying subscribers, the overall quality of video experience becomes important. It is for this reason that most long form content publishers employ some form of real-time streaming technology that modulates the video bit rate according to current network condition dynamically. Due to the premium nature of the content, most content publishers protect the copyrights through some digital rights management (DRM) mechanism. By being DRM protected, neither an in-network entity, like a video optimization server, nor the end user can modify or duplicate the content for shared viewership or later viewership.

2.3 OPERATOR VIDEO AND OTT VIDEO

Many mobile network operators are looking to extend their offers by offering content as a part of their subscriber’s package. They host premium video content inside their network in a form optimized for delivery to the devices they support. To provide a premium viewing experience, they will manage the network in order to optimize the streaming of the video to the user device by caching content closer to the edge of the network and using QoS on the RAN. The operator can implement various business models including zero rating their content as well as introducing additional services and applications such as ad pre-rolls, post-rolls and overlays.

The majority of the videos watched on mobile networks today are from Over-The-Top (OTT) providers. OTT videos are typically delivered over HTTP-Progressive Download channel for short form content and through HTTP-Adaptive Streaming for long form content. Mobile operators typically don’t have much influence in defining the quality of service and quality of experience for the end user.

2.4 BEST EFFORT VIDEO AND GUARANTEED BIT RATE VIDEO

Superior and uninterrupted online video experience relies a lot on the availability of high bandwidth continuously for long durations. Unfortunately, this can never be guaranteed on a mobile network. Even when the user is stationary, the effective bandwidth experienced by his device is likely to vary dramatically within a 5 minute duration (average length of video on YouTube is 5 minutes) due to time-varying channel conditions. Mobile operators have mainly two options to address the dynamic network variations and offer satisfactory service to their subscribers.

1) Best Effort Video Delivery: With the best effort delivery model, operators make it explicitly known to the consumers that actual video experience is dependent on network characteristics. In most cases, mobile operators promote/advertise an “up to” maximum network bandwidth that the specific pricing plan provides. This is by no means a guarantee as this is simply a theoretical maximum. Under lightly loaded network conditions, a good user experience may be realized.
2) Guaranteed Bit Rate Delivery: In this model, mobile operators guarantee a specific minimum bandwidth as part of the data service. 4G networks do provide mobile operators the technical capabilities to provide guaranteed bandwidth to users. It remains to be seen if this can be offered for higher values (like 1-2 Mbps) at all times, at all cell sites, without significantly affecting the throughputs delivered to the remaining users, particularly when there are many users.

2.5 VIDEO FOR ENTERTAINMENT AND VIDEO FOR REAL TIME COMMUNICATIONS

Today, the majority of video on the mobile network is intended for entertainment. Whether short form or long form, this content is made available by video producers to entertain, amuse and educate. Typically, there is a long time (days and even years) between production and consumption. Even for live streaming, there can be a delay of 30-60 seconds or longer between the live event and it being displayed on the screen. The content can be of sporting events, drama, films, cartoons, nature documentaries, etc. Such content is demanding of the highest feasible quality for both the audio and video.

Video telephony and video conferencing are growing in popularity. Such communications need to have a maximum latency of 150 msecs [9] to ensure good human interaction. Hence, they are often termed Real-Time Communications. However, the video image is usually of talking heads so there is no requirement for high fidelity audio and video. The WebRTC initiative from W3C will allow real-time communications to be embedded in and enabled from web browsers. Some browsers like Chrome and FireFox support WebRTC. This allows real-time communications to be added to web pages without the need for additional plug-ins or applications.

2.6 HTTP PROGRESSIVE DOWNLOAD

Basic download is the process by which a content provider will encode video and put it up in a web server used to host HTTP pages. Viewers will click on the URL of the video file and the video file is downloaded into their computer for viewing and the videos can be watched after the download is completed.

Encoding involves two key steps. Before putting the raw file that is the output from the recording device, the video is compressed into a format that the media player could play, based on the right quality chosen. Two decisions need to be made here namely, the choice of video quality and the choice of video format.

- Video quality: higher the bit rates result in higher video quality. For instance, a 100 kbps streaming video is phone-video quality but a 10 Mbps streaming video is DVD quality. Higher bit rates requests consume higher bandwidths.
- Video format: There are different file types supported by different media players, such as H.264 (MP4), FlashVideo (FLV), QuickTime (QT), Windows Media (WMV) etc.

A progressive download is the transfer of digital media files from a server to a client, typically using the HTTP protocol when initiated from a computer. Progressive download allows a user to access a file's contents before the download has been completed. In a progressive download, the video usually does not start playing until the buffer has grown large enough to ensure stall-free playback. Progressive downloading transmits a media file with metadata in the header of the downloaded file. A media player capable of progressive downloading reads this metadata and begins playback of the media file after a specified minimum portion of the data has been downloaded. Media players also require a buffer to hold data for use at a later time. The amount of data held in the buffer before playback is determined by both the producer of the content (in encoder settings) and by additional settings imposed by the media player.

2.7 RTP/RTSP DELIVERED VIDEO

Traditional streaming uses special media protocols, such as RTSP (Real-Time Streaming Protocol), RTMP (Real-Time Messaging Protocol) and RDT (Real Data Protocol). With RTSP, traditional streaming will send the video as a series of small packets in sequence in RTP packets. The RTP packet size is around 1 Mb for 1 Mbps quality video and these packets will get transmitted over UDP transport (UDP RFC 768). The protocol is designed so that it is better to have a momentary glitch in audio or video than for the playback to stop altogether and wait for missing data. The packets are streamed at a specified bit rate. The client can tell the server this rate using RTSP or derive it from the encoding rate of the content.

Forward Error Correction (FEC) can be applied to the UDP stream to help mask packet loss. Alternatively, the client can request the retransmission of the missing packets. This is done using RTP Control Protocol (RTCP RFC3611). RTCP can also be used to indicate the quality of the communication such as jitter and round-trip-time delay (RTT). This can be used to help decide whether to change the codec or bit rate.

RTP/RTSP video is most often used in wireline IPTV deployments. Strict admission control before starting a new stream ensures there will be sufficient bandwidth dedicated for the stream. It also works best over low loss networks so the FEC or packet retransmission overheads are minimized.

In contrast to progressive download described in Section 2.6, during a streaming media session, the file is never completely downloaded to a local storage device as it is with progressive download. This difference makes streaming media a more secure option that reduces the risk of piracy.

Issues with traditional streaming that motivate the need for HTTP adaptive streaming (in Section 2.8) include the following:

- Since dedicated media servers are needed, existing web servers cannot be leveraged. The transmission of media packets is through UDP protocols.

- Once the video is encoded at certain bitrates, the client will receive that quality of video content regardless what kind of CPU condition and bandwidth network the client has. This is similar to progressive download.

- Traditional streaming is harder to scale since, as a stateful protocol, traditional media servers maintain a one-to-one persistent connection with each client.

2.8 HTTP ADAPTIVE STREAMING (HAS)

The available bandwidth over the Internet is highly variable. Wireless technology adds to this variability due to the time and spatially varying radio propagation characteristics of the Radio Access Network (RAN). To handle this variability in bandwidth and to ensure best image quality when bandwidth is good, an adaptive streaming technology is required. For professional content streamed over the Internet, a common approach for streamed content is HTTP Adaptive Streamed (HAS). There are several variants including MPEG DASH, Apple HLS, Microsoft Smooth Streaming but they all work in the same way. Figure 5 shows the operation of HAS and Figure 6 shows the message exchanges to execute HAS.

.

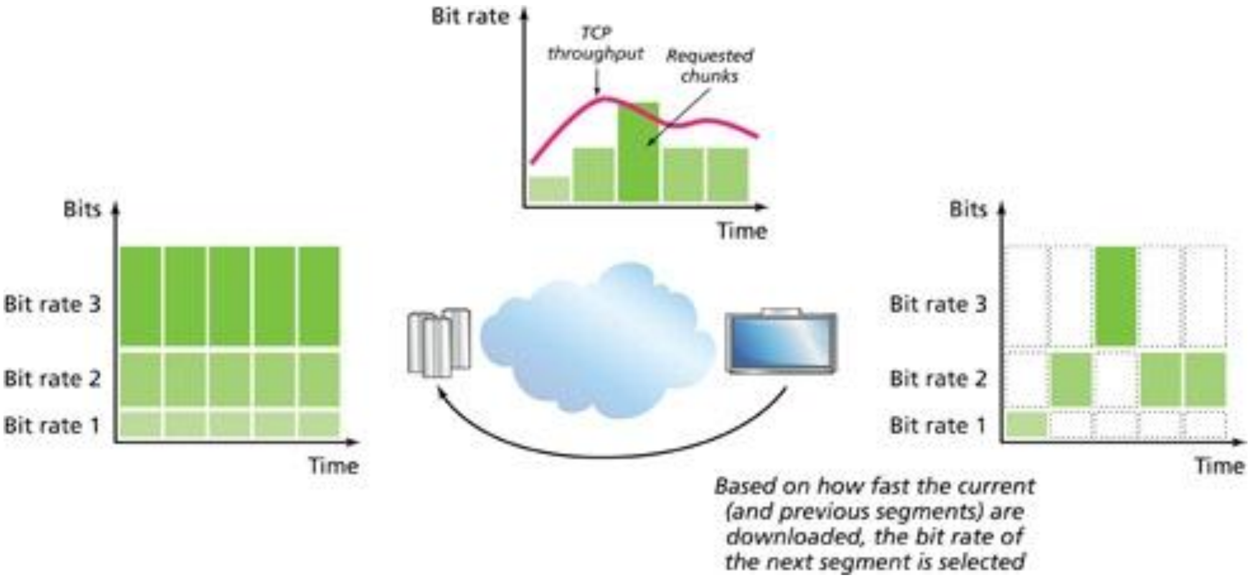
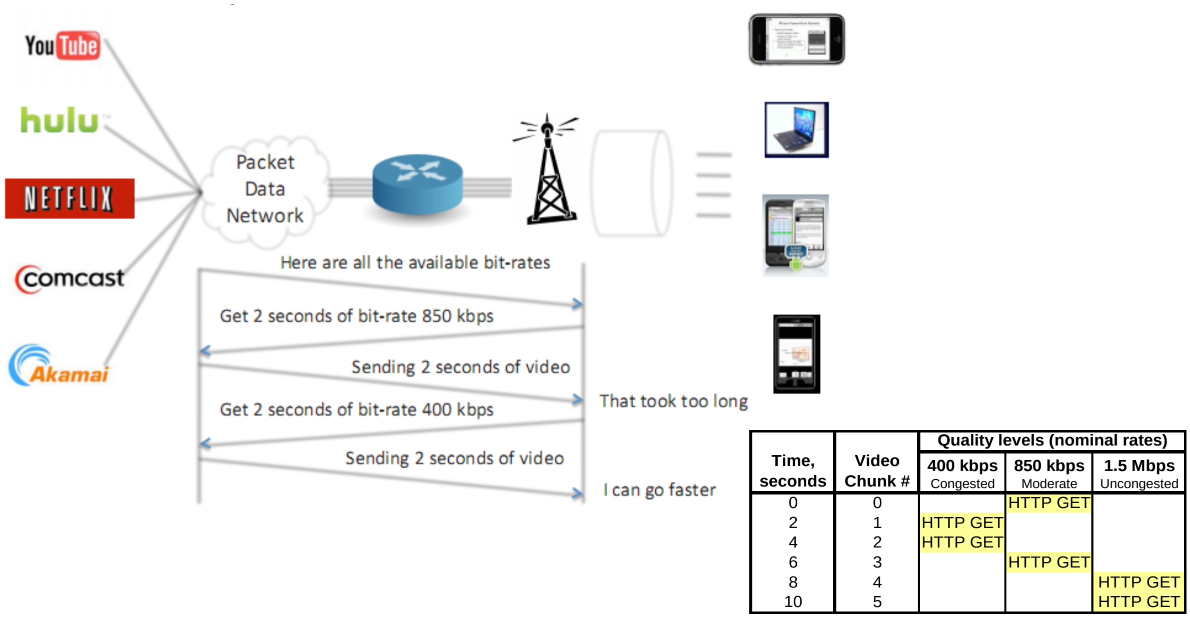


Figure 5: HTTP Adaptive Streaming


Packet
Data
Network
Here are all the available bit-rates
Get 2 seconds of bit-rate 850 kbps
Sending 2 seconds of video
Get 2 seconds of bit-rate 400 kbps
That took too long
Sending 2 seconds of video
I can go faster

Time, seconds	Video Chunk #	Quality levels (nominal rates) 400 kbps Congested	 850 kbps Moderate	 1.5 Mbps Uncongested
0	0		HTTP GET	
2	1	HTTP GET		
4	2	HTTP GET		
6	3		HTTP GET	
8	4			HTTP GET
10	5			HTTP GET

Figure 6: Message exchanges during HTTP Adaptive Streaming

Video gets encoded into different formats and video qualities. HTTP Adaptive Streaming uses normal web servers and transmits data via HTTP protocol. Video files are divided into small non-overlapping video fragments (or intervals) in sequence, between 1 and 10 seconds long. Each interval is then encoded in multiple bit rates. The bit strings associated with these encodings are referred to as chunks (also called segments in MPEG DASH). The HAS client uses HTTP to request chunks – the first one usually at a low bit rate. If chunks are consistently delivered in a time shorter than the interval length, the Rate Determining Algorithm (RDA) in the client selects a higher bit rate for the next chunk (see Figure 5). In that way, the RDA continually senses the available throughput and adapts the video rate accordingly. In order to absorb mismatches between video rate and throughput, the video client maintains a play-out buffer. A larger playout buffer means a longer period of poor radio conditions can be tolerated. But it also increases the startup time. Assuming that the RDA is working properly, there should be no stalls while it automatically adjusts to the available bandwidth.

Using HTTP has several advantages. It allows reuse of standard web technologies in the distribution and delivery of video. This helps minimize costs and avoids the need for special network technologies or configurations. Furthermore, HTTP operates over TCP, which automatically provides flow control and recovery from packet loss. Another advantage of HTTP-based video delivery is its ability to traverse firewalls through port 80. Finally, HTTP uses caching mechanisms on the web. Cached data is generally closer to the viewer and hence more easily retrievable. This capability should both decrease the total bandwidth costs associated with delivering the video, since more data can be served from web-based caches rather than the origin server, and improve quality of service.

Adaptive streaming enables viewers to start video play-back while the content is being downloaded. Consider two common sources for video streaming traffic in the Internet namely, Netflix and YouTube. Users can view Netflix and YouTube videos either on PCs, using a Web browser, or on mobile devices, using a Web browser or a mobile application. A mobile application is the native Netflix or YouTube application running on mobile devices.

Netflix and YouTube use TCP to stream videos. A number of professional video content providers, such as Netflix, Hulu and BBC iPlayer, currently use HAS to deliver video content over unmanaged networks. HAS allows video delivery to adapt dynamically to the available bandwidth of a network while providing adequate quality of experience (QoE) for subscribers.

Below are some common implementations in the industry:

- Microsoft Smooth Streaming: http://www.iis.net/download/SmoothStreaming
- Adobe Dynamic Streaming for Flash
- Apple HTTP Adaptive Streaming

Apple's HTTP Live Streaming (HLS), Microsoft's Smooth Streaming, and Adobe's HTTP-based Dynamic Streaming (HDS) use Hypertext Transfer Protocol (HTTP) delivery protocol and standard HTTP web servers to deliver streaming content, thus obviating the need for a streaming server. HAS technology is standardized in MPEG DASH (Dynamic Adaptive Streaming over HTTP). The International Standard was published in April 2012 as ISO/IEC 23009-1. DASH is discussed further in Section 2.9.

2.8.1 COMPARISON BETWEEN PROGRESSIVE DOWNLOAD, STREAMING AND ADAPTIVE STREAMING

With HTTP- Progressive Download and traditional streaming protocols (RTP/RTSP), once the client starts streaming a video, the bit rates of that video clip will stay unchanged. Even if the client has a very bad network condition later on, it still needs to load the video with high bit rates. This is in contrast to HTTP Adaptive Streaming (HAS), where a client could enjoy a smooth viewing experience because of video bit rates that get switched seamlessly according to the network bandwidth availability and possibly the client's CPU condition.

In addition, HAS utilizes HTTP protocol which leverages the scalability of the whole Internet. There are a lot of intermediary proxies or gateways between client and the server. Since video files get chunked into small file sizes, it is easier to cache in all these intermediaries. Therefore, HAS works well with content delivery networks (CDNs) to enable media content providers scale in a very cost-effective manner.

2.9 PLATFORMS FOR VIDEO DELIVERY

Adaptive streaming involves producing several instances of a live or on-demand source files and making them available to various clients depending upon their delivery bandwidth and CPU processing power. By monitoring CPU utilization and/or buffer status, adaptive streaming technologies can change streams when necessary to ensure continuous playback or to improve the experience.

One key difference between the various adaptive streaming technologies is the streaming protocol utilized. For example, Adobe's RTMP-based Dynamic Streaming uses Adobe's proprietary Real-Time Messaging Protocol (RTMP), which requires a streaming server and a near-continuous connection between the server and player.

There are multiple video delivery platforms that are being used. The important point to note is that most video delivery platforms are building in some form of bandwidth and device awareness into them so as to enable them to provide superior video experience. Proprietary adaptive streaming platforms rely on cooperative video players that implement their proprietary transport protocol. The popular platforms belonging to this category are Apple HLS, Adobe Dynamic Streaming, Microsoft Smooth Streaming, etc.

Recently, major video delivery platforms have gotten together and are formulating a dynamic streaming standard called DASH. MPEG DASH (Dynamic Adaptive Streaming over HTTP) is a standard for adaptive streaming over HTTP that has the potential to replace existing proprietary technologies like Microsoft Smooth Streaming, Adobe Dynamic Streaming, and Apple HTTP Live Streaming (HLS). A unified standard could be helpful to content publishers, who could produce one set of files that play on all DASH-compatible devices. DASH is an attempt to combine the best features of all HTTP-based adaptive streaming technologies into a standard that can be utilized from mobile to OTT devices.

All HTTP-based adaptive streaming technologies have two components: the encoded audio/video streams themselves and manifest files that identify the streams for the player and contain their URL addresses. For DASH, the actual audio/video streams are called the Media Presentation, while the manifest file is called the Media Presentation Description. The Media Presentation is a collection of structured audio/video content that incorporates periods, adaptation sets, representations and segments. The Media Presentation defines the video sequence with one or more consecutive periods that break up the video from start to finish. The DASH manifest file, called the Media Presentation Description, is an XML file that identifies the various content components and the location of all alternative streams. This enables the DASH player to identify and start playback of the initial segments, switch between representations as necessary to adapt to changing CPU and buffer status, and change adaptation sets to respond to user input, like enabling/disabling subtitles or changing languages. With several big players like Microsoft, Adobe, and Google participating actively, it is possible that DASH will be widely adopted in the marketplace.

From the client standpoint, most browsers today start a separate app/plug-in to play the selected video. With the advent of HTML5, browsers would be capable of playing videos natively. This vertical integration of video rendering into the browsers provides web authors the added ability to control the look and feel of the video area.

3. CHALLENGES FOR MOBILE VIDEO DELIVERY

Video delivery is characterized by the need to send large amounts of data over long time durations. The sections below describe the challenges of delivering video in a mobile environment over the air interface and within the core network.

3.1 CHALLENGES FOR VIDEO DELIVERY DUE TO AIR INTERFACE

Mobile video can create a multitude of problems on mobile networks due to their highly demanding requirements on bit rate, latency, delay and jitter. Although actual mobile networks do comprise both wired and wireless interfaces, it is the air interface and possibly the backhaul that suffer the most from highly strained operational conditions for video delivery. Air interface actually represents the weakest link in any network architecture due to several reasons:

- The high error rate potentially present in mobile environments must be overcome through the use of multiple techniques, such as adaptive modulation and coding, retransmissions at different levels, and limitation of the block sizes to be transferred to the users. Some of these techniques increase the effective delay incurred in the radio access part to transfer a given application-level information block. An increased delay results in a higher Bandwidth-Delay product (BxD), and this in turn may impair the net throughput of TCP connections if receive window sizes are not appropriately dimensioned.

- The scarce auctioned spectrum must be shared among the mobile users at the radio interface, and fair network operation must ensure that users have at least a minimum access to resources (at least for normal tariff plans).

- Typical mobile channels involve several degradation phenomena such as multipath, large-scale and small-scale signal fading, Doppler spreading, and increased noise, to name a few. In addition to these, there appear other interference issues between cells caused by single-frequency operation in technologies such as UMTS, WiMAX or LTE.

- Limitations in devices usually incur some penalty on processing power capabilities, resulting in a lower spectral efficiency for a given radio condition. Limitations in the RF circuitry of devices introduce additional degradations such as poorer transmission mask, lower sensitivity, increased noise figure or reduced antenna efficiency, to name a few. All these limitations directly translate into poorer air interface capacity.

- Many connection-oriented protocols are not optimized for operation in mobile networks because they usually assume that a packet loss should probably be caused by a network congestion issue instead of a link error. As an example, TCP congestion control reacts to a packet loss by reducing the size of the transmission window, thereby decreasing the peak data rate. High delays experienced by packets when the user suffers from poor signal conditions should be matched by appropriate TCP receive window sizes, which are not usually optimized for the actual radio interface performance [11].

3.1.1 SPECTRAL EFFICIENCY OF 3G/4G CELLULAR TECHNOLOGIES AND IMPACT ON MOBILE VIDEO

Limitations in the spectral efficiency of radio access technologies, which account for typical aggregate throughput values in a given amount of spectrum, have a clear impact on video performance. The higher the spectral efficiency, the higher the air interface throughput and hence the improved user experience for mobile video. Table 1 shows the data consumed by different streaming applications [10]. This table indicates that different devices need different bit rates for acceptable Quality of Experience. Moreover, even a modest use (e.g. with content encoded in AVC/H.264, 500kbps on smartphone for 30 minutes per day 20 days a month) will add up to a high monthly data usage totals for subscribers. It is apparent that medium- and high-quality video applications demand net throughput values of several Mbps, which directly results in minimum required values for the average spectral efficiency of the weakest network link (the air interface). Unique ways to minimize congestion to the network due to video application download, utilization of radio resources for video download during non-peak hours, mobile video delivery through broadcast solutions, and influencing subscriber behavior strategies will be needed in order to make mobile video delivery over cellular networks efficient and scalable for large number of users. These will be discussed in the detail in the following sections.

Table 1: Data Consumed by Different Streaming Applications [10]

Application	Throughput (Mbps)	MByte/Hour	Hrs./day	GB/month
Stereo Music	0.1	58	0.5	0.9
			1	1.7
			2	3.5
			4	6.9
Small Screen Video (e.g. Feature Phone)	0.2	90	0.5	1.4
			1	2.7
			2	5.4
			4	10.8
Medium Screen Video (e.g. Smartphone)	0.5	225	0.5	3.4
			1	6.8
			2	13.5
			4	27
Medium Screen Video (e.g. Tablet)	1	450	0.5	6.8
			1	13.5
			2	27
			4	54
Larger Screen Video (e.g. High Quality on Tablet)	2	900	0.5	13.5
			1	27
			2	54
			4	108

In order to satisfy the requirements for these typical video services, typical air interface performance values are given in Table 2. Table 2 shows a comparison of various technologies in terms of downlink and uplink speeds.

Instead of considering peak data rates (which can serve for benchmark comparison between technologies), typical average aggregate throughput values can give a better idea of realistic performance in average cell and user conditions. This is particularly true for video users, since they require high sustained rates for long periods of time (tens of minutes to hours), and are likely to be moving across multiple cells with varying radio and hence interference conditions during that time.

Growth of video is likely to be facilitated with recently introduced technologies such as LTE and LTE-Advanced. A study of various 3GPP technologies [10] indicates that introduction of OFDMA in LTE allows very high throughput rates in uplink and downlink with all communications handled in IP domain. LTE-Advanced introduces carrier aggregation that recognizes the asymmetry of uplink/downlink traffic patterns and thus increases the downlink bandwidth by aggregating multiple carriers while keeping uplink operation unchanged.

Table 2: Performance comparison of 3GPP mobile technologies [10]

Technology Name	Type	Characteristics	Typical Downlink Speed	Typical Uplink Speed
UMTS	CDMA	3G technology providing voice and data capabilities. Current deployments implement HSPA for data service.	200 to 300 kbps	200 to 300 kbps
HSPA	CDMA	Data service for UMTS networks. An enhancement to original UMTS data service.	1 Mbps to 4 Mbps	500 kbps to 2 Mbps
HSPA+	CDMA	Evolution of HSPA in various stages to increase throughput and capacity and to lower latency.	1.9 Mbps to 8.8 Mbps in 5/5 MHz 3.8 Mbps to 17.6 Mbps with dual carrier in 10/5 MHz	1 Mbps to 4 Mbps in 5/5 MHz or in 10/5 MHz
LTE	OFDMA	New radio interface that can use wide radio channels and deliver extremely high throughput rates. All communications handled in IP domain.	6.5 to 26.3 Mbps in 10/10 MHz	6.0 to 13.0 Mbps in 10/10 MHz
LTE-Advanced	OFDMA	Advanced version of LTE designed to meet IMT-Advanced requirements.		

HSPA and HSPA+ throughput rates are for a 5/5 MHz deployment. N/M MHz means 5 MHz is used for the downlink and M MHz is used for the uplink. From the above table it is clear that only HSPA, HSPA+, LTE and LTE-Advanced technologies may cope with the huge bit rate requirements of present wireless video applications in smartphones, tablets and laptops. Legacy devices with limited features will definitely suffer from reduced performance in the same network conditions.

Air interface performance can conveniently expressed in terms of the spectral efficiency, which is defined as the throughput value per unit of occupied bandwidth (bps/Hz). LTE is more spectrally efficient with wider bandwidths (10, 15 and 20 MHz), due to the significant overhead incurred by layer-1 control signaling in narrow bandwidths. Figure 7 shows the LTE spectral efficiency as a function of system bandwidth, thus highlighting the ability of LTE to support mobile video delivery with good Quality of Experience for the user.

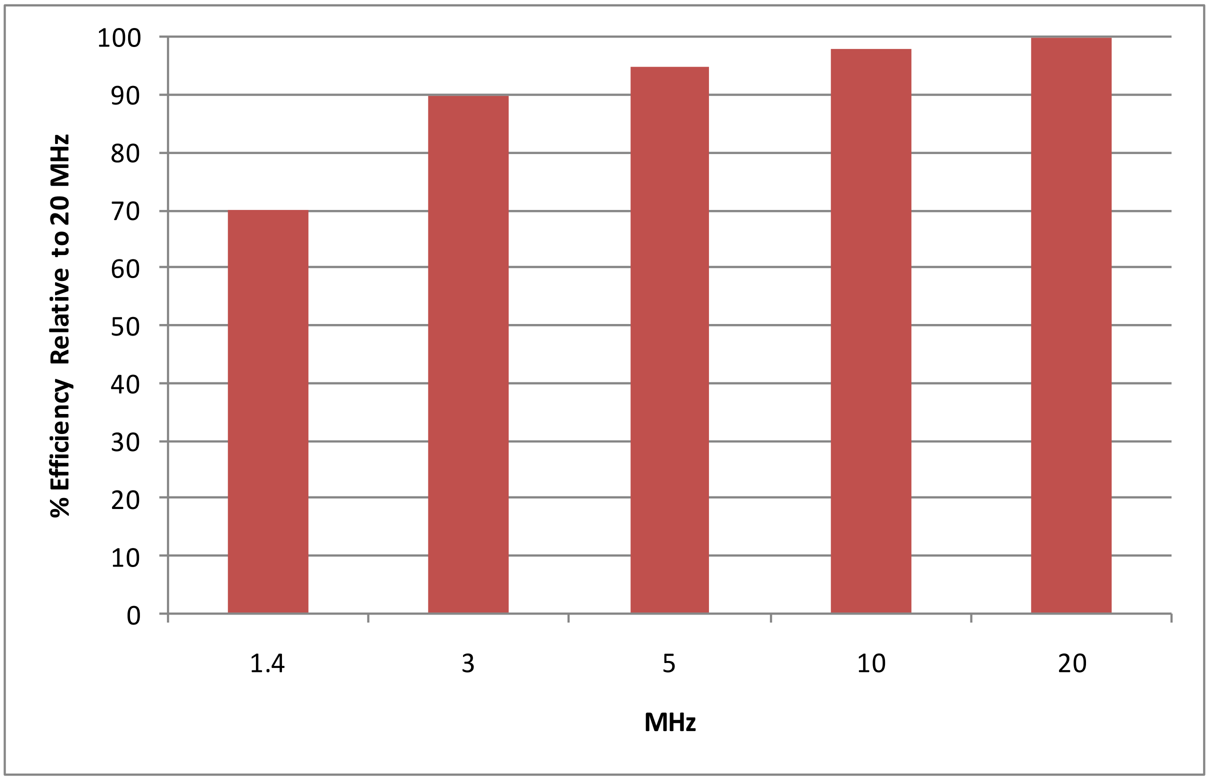


Figure 7: LTE Spectral Efficiency as a Function of the System Bandwidth [10]

Typically, the need for downlink capacity is more important particularly for video. A study of various 3GPP technologies indicates that spectral efficiency is usually lower in uplink, which is aligned with typical usage patterns of mobile users. However, with the advent of WebRTC and apps such as Skype and Facetime, it is expected that the uplink capacity will also come under pressure as people increasingly use video telephony.

Higher peak rates are usually obtained only when the mobile is in good stationary conditions, with high and stable signal to noise ratios and little Doppler spread. In these conditions, video delivery can be more efficient due to the feedback mechanisms that usually exist between the mobile and the base station. Proper link adaptation can only be achieved through use of some sort of signal quality feedback from the users, such as the Channel Quality Indicator (CQI) reports employed in HSPA and LTE.

3.1.2 AIR INTERFACE LATENCY AND IMPACT ON MOBILE VIDEO

Closely related to the spectral efficiency is air-interface latency. Video quality of experience is facilitated with the introduction of LTE and LTE-Advanced due to their high bandwidth and low latency characteristics. Higher bandwidth allows for users to receive video with higher quality without resulting in packet loss due to congestion in the radio interface. Lower latency allows for problems with TCP delivery to be minimized, which is critical for the delivery of video through HTTP Adaptive Streaming (HAS).

Air interface introduces latencies due to mainly three mechanisms:

- The link adaptation function in the base station must choose the preferred modulation and coding scheme for each particular user, based on the amount of information to be transferred and the signal quality reported by the users. Information blocks of different sizes are sent over the radio interface thus requiring different amounts of time to transmit a given higher-level packet (such as an IP packet). This results in different air interface latencies that depend heavily on radio conditions.

- The air interface usually introduces a hybrid retransmission mechanism to overcome channel errors, sometimes known as Hybrid Automatic Repeat Request (HARQ). These retransmissions involve additional latencies depending on the HARQ round-trip-time (HARQ-RTT), as e.g. 8 ms in LTE and 12 ms in HSPA (for the downlink).

- The scheduler in the base station ensures proper access to the available shared resources for coexistence of multiple users in a given cell. If the cell is in low load conditions, the scheduling process is easy and the users are eventually given full cell resources (if allowed by the applicable Quality of Service policy). In high load conditions, the scheduler must maximize cell throughput while ensuring fair use of resources among users. This results in higher queue sizes and increased packet delay.

A study of the latency performance in various 3GPP technologies [10] indicates that LTE deployments are likely to result in the lowest possible latency values according to the physical layer capabilities. Round trip time refer to the measured time delay value between the subscriber unit and a node immediately external to the wireless network and does not include Internet latency. As an example, trial results reported by LTE/SAE Trial Initiative (LSTI) measured LTE round-trip times ranging from 18 to 28 msec [12].

Air interface latency and jitter significantly impact the performance of mobile video. An increased latency represents a drawback for the deployment of wireless interactive services, such as high-quality videoconferencing or online gaming based on streaming protocols like RTP, RTSP or RTCP. Other video services like progressive download video (as opposed to streaming video) are not so sensitive to latency due to their limited interactivity with the user. With progressive download video, the contents are transferred via standard HTTP protocol and appropriate receive buffers at the client minimize the impact of throughput and latency impairments. Note that the air interface latency is in addition to the expected latency incurred by the chain of network nodes that the information must traverse.

3.1.3 INTERFERENCE AND HANDOVER ISSUES IN 3G/4G CELLULAR TECHNOLOGIES

Interference and handover issues can affect the video quality of experience, depending on the user mobility, type of video being accessed etc. Metrics for video QoE are discussed in Section 4.2.

3.1.3.1 IMPACT OF RADIO INTERFERENCE ON VIDEO DELIVERY

Mobile video users are likely to experience significantly varying interference levels, both as they consume video over long periods of time and as they move across cells. Higher interference translates to lower signal to noise ratio and hence lower bandwidth available for video, in addition to resulting in various video impairments described in Section 4. It is therefore important to consider the impact of interference over the quality of video as viewed by the users, and on the allocation and optimization of network resource to address the challenges of spatial and temporal variations in interference as described below.

Interferences in wireless networks appear when the spectrum in use is shared among users and cells. By sharing the same frequency in all cells, effective mechanisms must be designed to cope with increased interference between users, cells and services (especially at the cell edges).

Soft handover in WCDMA allows the user to be connected to more than one base station simultaneously because the mobile can coherently add downlink signals in order to improve coverage. This feature enables a more continuous coverage when moving between cells, especially for control channels where correct demodulation is critical to maintaining synchronization. In a stationary situation, the user in soft handoff can typically be connected to two or three cells. Soft handover is not present in HSDPA channel for the downlink, due to the fast allocation mechanisms that would be involved and the need for tight phase synchronization of the cells. For the rest of the downlink channels, as well as all uplink channels, soft handover is employed.

LTE does not allow such soft combining (except for multicast/broadcast data transmissions). However, the flexibility of assigning time and frequency resources with such low granularity enables a highly-efficient utilization of radio resources in each cell, at the cost of increasing terminal complexity through advanced interference-cancellation techniques.

Interference is an impairment that ultimately limits performance in cellular networks. It also prevents base stations from delivering the highest data rates to the users if their perceived signal-to-noise ratios are not high enough. In the context of traditional homogeneous networks (comprising a collection of macrocells), interference represents the dominant issue at the boundaries between cells and forces handovers and cell reselections to keep the best radio conditions at every location. Interference and handover could result in reduced data throughput for video delivery. Therefore, the TCP and the lower layer protocols must cope with increased interference when crossing the boundaries between cells.

Heterogeneous networks comprising a mixture of cells with different powers, frequencies and (possibly) radio access technologies, make interference issues more prevalent due to increased percentage of locations being in the boundaries between cells.

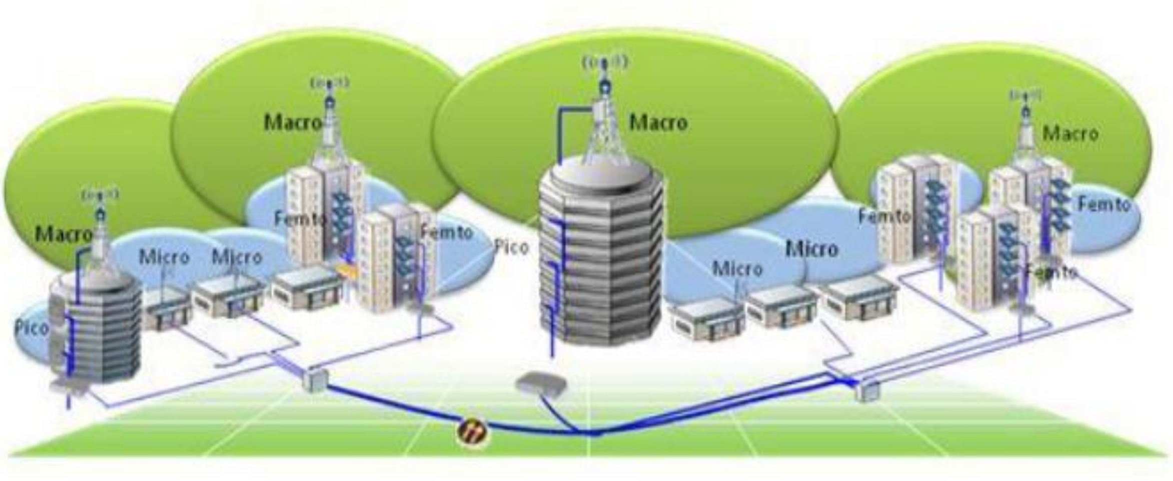


Figure 8: HET-NET architecture [13]

HET-NET deployments pose their own challenges as compared to homogeneous networks. A number of techniques have been devised in order to cope with increased interference in the boundaries between macrocells and small cells. These issues, as well as mitigation techniques, can significantly impact mobility mechanisms in these networks and hence the performance of mobile video users.

3.1.3.2 IMPACT OF MOBILITY IN WIRELESS CELLULAR NETWORKS ON VIDEO DELIVERY

Mobility presents unique challenges in delivery of video. Mobility allows a user to get the impression of being seamlessly connected to the Internet and/or the circuit-switched operator's network, irrespective of his/her actual location. Mobility can be accomplished through a set of radio access network and core network features dealing with the maintenance of data session over the radio access.

Mobility management at the air interface must cope with eventual variations in the best server due to user movement. The best server represents the most suitable base station to which the user must be connected to in order to get the best radio conditions. Soft handover operation extends this best server concept to multiple cells, but a single serving cell always exists that handles control signalling to/from the user. Radio resource management (RRM) algorithms are defined by the different radio access technologies to enable mobility through handovers (changes in connected-mode best server during the course of an ongoing connection) and cell reselections (cell changes in idle-mode during inactivity periods). Although the basic framework for performing radio resource management is standardized by each technology (in terms of the reporting quantities to be considered by both devices and base stations), actual RRM algorithms are implementation-dependent thus introducing an additional level of uncertainty in performance. Depending on the network configuration and equipment in use, significant performance differences can be observed as a consequence of RRM and mobility issues. Even highly-optimized radio access networks must cope with increased signaling traffic coming from mobility management due to handovers, information exchange among radio access network nodes (such as RNCs), cell updates, location area updates, etc.

3.1.3.2.1 HANDOVERS AND IMPACT ON MOBILE VIDEO USERS

Handover mechanisms allow users to move from one cell to another while maintaining continuity of service. Understanding of the handover mechanisms allows for identification of appropriate techniques to address the challenges of mobile video delivery.

In hard handovers (like in HSDPA data channel), connection with the serving cell is lost when switching from one cell to another. When delivering mobile video, this interruption should be absorbed by appropriate receive buffers at the client's player irrespective of the type of video service in use. Furthermore, video users tend to be generally low mobility users which could result in fewer handover failures by permitting sufficient time for the handover signaling messages to reach the UE successfully as the user is riding through the strong interference region. Soft handovers (present in UMTS Rel-99 channels) are characterized by an active set where a number of cell connections are simultaneously kept for the user, hence allowing for softer cell transitions and improved reception (at the cost of increased radio resources). With soft handover, there is no interruption in the video service.

If the player's buffer is not sufficiently large, handovers can cause excessive delays for some packets due to retransmissions at the MAC or TCP levels, therefore increasing jitter and significantly impairing video performance. If retransmissions are not present (as in video over UDP), the effect of handovers would result in a reduced available bit rate during the handover time, thus resulting in significant video artifacts. Video over UDP (i.e. non HTTP based protocols) may benefit from solutions like "bicasting" that allows duplication of RLC unacknowledged mode Packet Data Units from the RNC to both the old and the new serving cells when the user makes a serving cell change to minimize the impact of handover interruption times. Through bicasting, dropped packets in the source cell can be received from the target cell after handover.

Mobile video places demands on the air interface. RTP/UDP streaming connections are usually constrained by tight values of latency and jitter, therefore retransmissions are not usually allowed at the radio interface to limit delay. Streaming video (employing real-time streaming protocols like RTP, RTCP or RTSP) must address the challenges of user mobility, since the radio bearer does not usually provide any retransmission mechanisms for error correction. Users at the cell edges may therefore experience appreciable quality degradation if bit error rates increase due to extra interference and/or signaling overhead when performing handovers. However, in all cases, the context is maintained together with the IP address and quality of service attributes, as a result of which applications should not be aware of any user mobility. HTTP Adaptive Streaming will not likely experience the issues related to jitter and delay, since typically the client has a buffer of media to be played out and fluctuations in latency and in jitter can be absorbed into this buffer without any degradation of the end user experience.

Progressive download video has the advantage of allowing more flexibility in traffic patterns due to the existence of a receive buffer that compensates for fluctuations in user data due to interference or mobility. Moreover, TCP connections guarantee error-free reception, and the radio access is allowed to perform retransmissions in order to reduce the number of end-to-end TCP retransmissions to a minimum. Retransmissions however come at the expense of increasing the overall delay, as a consequence of which the player’s buffer may empty and the video playback may stalls.

HTTP Adaptive Streaming (HAS) can cope with mobility issues in a more efficient way compared to other forms of video delivery, as the bandwidth is adaptively changed according to the instantaneous network conditions, therefore reducing the need for retransmissions. Issues might only appear in handovers if the instantaneous bit rate after cell change is significantly lower than the screen resolution, giving rise to appreciable visual artifacts.

3.1.3.2.2 RETRANSMISSIONS AT THE AIR INTERFACE AND IMPACT ON MOBILE VIDEO USERS

Retransmission mechanisms for packet data exist at various layers including the MAC layer and at the IP and application levels. End-to-end retransmissions should be avoided whenever possible due to the extra delay incurred that will impact video quality of experience. Furthermore, TCP errors are considered as congestion events that trigger a reduction in transmission window, thus potentially reducing throughput. Therefore it is important to recognize the importance of MAC layer retransmissions on video delivery as described below. Retransmissions at the radio access usually comprise two levels:

- Hybrid retransmissions at the Medium Access Control (MAC) layer, usually under a mechanism called HARQ that corrects most channel errors within a few milliseconds. This is the fastest retransmission scheme available at present in cellular networks.

- Layer-2 retransmissions above the MAC layer (when the Radio Link Control (RLC) protocol Acknowledged Mode (AM) is used) are in the form of ARQ mechanisms and managed by radio access nodes. ARQ is performed at the Radio Network Controller (RNC) in UMTS and HSPA, at the eNodeBs in LTE, and avoids expensive end-to-end retransmissions. HARQ is faster than ARQ in correcting errors, and both of them are in turn much faster than TCP retransmissions.

For these reasons, it is advisable to activate retransmissions at the radio access network when error-free video delivery services are deployed. However, appropriate receive buffers are needed to avoid video stalls because of the extra delay incurred. In situations with poor coverage, or increased interference levels, the number of retransmissions can be so high that the video player stalls, and this may also happen when the cell is so congested that the scheduling process in the base station is not able to serve as much throughput as needed by the client to maintain an acceptable video quality.

3.1.3.2.3 PHYSICALLY FAST-MOVING USERS USING MOBILE VIDEO

Physical speeds of users can significantly impair RRM mechanisms devised between users and the base stations. Handover processes, link adaptation processes, closed loop MIMO operation and transmission of LTE reference signals may get compromised when the user speed is above a certain limit.

Delivery of consistently high throughput for high mobility for wireless users, including LTE and LTE-advanced users, is a challenge due to various reasons that have been described in recent publications [14, 15]. It is therefore unlikely that wireless cellular networks are able to deliver most demanding video services above a certain user speed. The exact value of the speed limit is dependent of multiple factors including propagation environment, actual network configuration, etc. In these cases, it is more advisable to switch to a lower resolution video service (if available), instead of trying to cope with increased errors through a higher number of retransmissions.

3.1.4 SCHEDULING MECHANISMS AT THE AIR INTERFACE

Packet-based radio technologies like HSDPA or LTE bring the capability to share air interface resources in a more efficient way among the active users, therefore posing the problem of how to distribute them across the available resource dimensions (time and codes in the case of HSDPA, time and frequency in LTE). This is addressed through scheduling of air interface resources.

Scheduling in both LTE and HSPA radio access networks is performed by the base stations so as to take advantage of reduced latency and higher level of interaction with the users. The choice of scheduling strategy affects the performance of the mobile video application that co-exists with other applications that are competing for the same radio resources. Scheduling strategies are designed to satisfy a given number of conflicting requirements:

- Cell capacity should be maximized.
- Users should not suffer from starvation due to other data-hungry users.
- The scheduling process should grant a given fairness level by which all users can access resources with the same opportunities.
- Users with different radio conditions will, in principle, experience different qualities, in spite of the intended fairness level.

Scheduling strategies pursue a given fairness level which depends on the instantaneous radio conditions and the history of the user, with variations that lead to different scheduling strategies. Three important dimensions can be highlighted in any scheduling strategy:

- The choice of suitable scheduling strategies for different radio conditions according to the desired fairness level, such as whether users in good radio conditions should be boosted against cell-edge users, or whether cell-edge users should be favored instead. This gives rise to different scheduling strategies among which Round-Robin, Proportional Fair and Maximum Rate are typical examples that serve as bases for more elaborate strategies.

- The management of different user subscriptions (if available), by which premium users would be given special treatment compared to normal or low-grade users.
- The management of different expectations for different services, each with particular target requirements on maximum packet delay, maximum error rate, and guaranteed throughput (for the case of GBR bearers), all of which configure a given Quality of Service (QoS).

While the first dimension can be dealt with by the particular scheduling strategy, the other two should be appropriately tackled by the Policy and Charging architecture of the wireless network.

- UMTS allows for some QoS control by defining so-called QoS traffic classes, namely Conversational, Streaming, Interactive and Background classes.
- LTE with its Evolved Packet Core (EPC) provides an evolved policy and charging architecture and introduces the concept of QoS Class Identifier (QCI), which offers full control of QoS parameters on a per-service and per-subscription basis. The various QoS attributes foreseen by the network, are inputs to the scheduler. Section 6.2.1 describes the various QCIs and the priorities and delay budgets associated with each that can be used to select the appropriate QCIs for a given mobile video application.

From the above, it may seem that cellular technologies provide sufficient control over QoS attributes so as to ensure fair and efficient resource cell utilization. However, reality proves that actual applications may create multiple coexistence problems even in low load situations, thus making the schedulers very sensitive to “greedy” applications. For example, consider that a YouTube mobile user is being served with a high bit rate through progressive download in an LTE low load situation. Upon connection, the scheduler will grant a high number of resources for that user in order to serve the high amount of expected video data within the delay constraints imposed by the bearer characteristics. For the QoS delay restrictions to be fulfilled, for a given QCI, high bit rate video services will require a high number of LTE resource blocks and a high sub-frame occupancy, thus preventing other low-bandwidth users from exploiting cell resources even in low load conditions. This may effectively overload the air interface even with very few users, and other low-bandwidth services will suffer from lack of performance in the same priority conditions as the greedy video user. In this case, the dynamic modification of bearer attributes (and hence QCIs) or modification of attributes within a given QCI may be needed, or some traffic shaping may be required at the core network gateway to limit the bit rate for a given mobile video application.

An additional scheduling dimension appears in LTE when so-called frequency selective scheduling (FSS) is exploited. Due to the ability to grant specific frequency bands depending on the channel state information as reported by the users, the scheduler can select a given subset of frequencies in which the best possible channel conditions are met according to the radio environment. Mobile radio impairments such as fading and shadowing introduce deep variations in the signal quality even in static conditions, due to movement of the surrounding objects and obstruction of the radio link. These impairments create a frequency-selective channel response full of dips and fades, which can be exploited by selecting only those frequencies that present an appreciable peak response. The main implication for mobile video delivery is that the granted bandwidth will not be constant throughout the connection, and this may introduce impairments in constant-rate video services if the video stalls due to the player’s buffer emptying. HTTP Adaptive Streaming can help circumvent this problem by lowering the quality of video at the user device to an acceptable level without resulting in freezing of the video frame.

3.1.5 SIGNALING CHALLENGES FOR VIDEO

The “signaling storm” problem caused by increased smartphone penetration in 3G and 4G wireless data networks refers to the disproportionate amount signaling resources consumed relative to amount of data traffic carried by the network. In cellular networks, substantial number of signaling messages is generated prior to data transfers to activate the radio bearers that are in idle state. When most data transfer sessions are short-lived, a substantial amount of signaling is generated in relative proportion.

Smartphone signaling traffic stems from a plethora of applications that rely on short data transfers for a variety of reasons. Some of the common causes for short data transfers are application and TCP keep alive messages sent to maintain long-lived application or TCP sessions, application notifications sent from servers to the smartphone apps, status, presence or location updates, short instant messages or Twitter type posts. However, video has distinct traffic characteristics with significant amount of data being downloaded nearly continuously and generally does not cause disproportionately large amount of signaling traffic.

A poor choice of the inactivity or dormancy timer that prematurely drives the radio bearer from active state to idle state can cause unnecessary signaling in video streaming sessions. This is because highly bursty video sources can create significant inactivity periods between chunks of data. As an example, the YouTube video service delivers chunks of video according to a bursty pattern, characterized by significant peaks followed by relatively large inactivity periods [16]. If the state transition timer is not larger than the gaps in video transmission, the network can move the device to a dormant or idle state. Subsequently, when new video chunks are to be delivered to the device, a new radio connection must be established, with the corresponding signaling traffic and switching of states and this situation may repeat during the whole video session thus creating many signaling events.

Furthermore, every transition between idle and connected states involves a large number of RRC messages exchanged through the air interface. This takes an appreciable amount of time as well as significant RNC processing. These delays increase the effective delay of the end-to-end connection, and if accumulated over time they can force the player’s buffer to empty, thus resulting in video playback stall. Streaming video services with limited buffer capabilities will additionally experience increased jitter and therefore impaired user experience due to signaling storms. However, this particular issue can be easily mitigated by setting the inactivity timer value in the range of 8 to 10 seconds since with HTTP Adaptive Streaming, the gap between requests of chunks is a video transmission is unlikely to be larger.

In summary, signaling traffic generation is not expected to be a major challenge presented by video applications provided the right timer value settings are configured.

3.1.5.1 IMPACT OF SIGNALING ON BATTERY CONSUMPTION

Figure 9 shows the effect that various video playout buffer algorithms have on network load and battery consumption. This is shown through four examples.

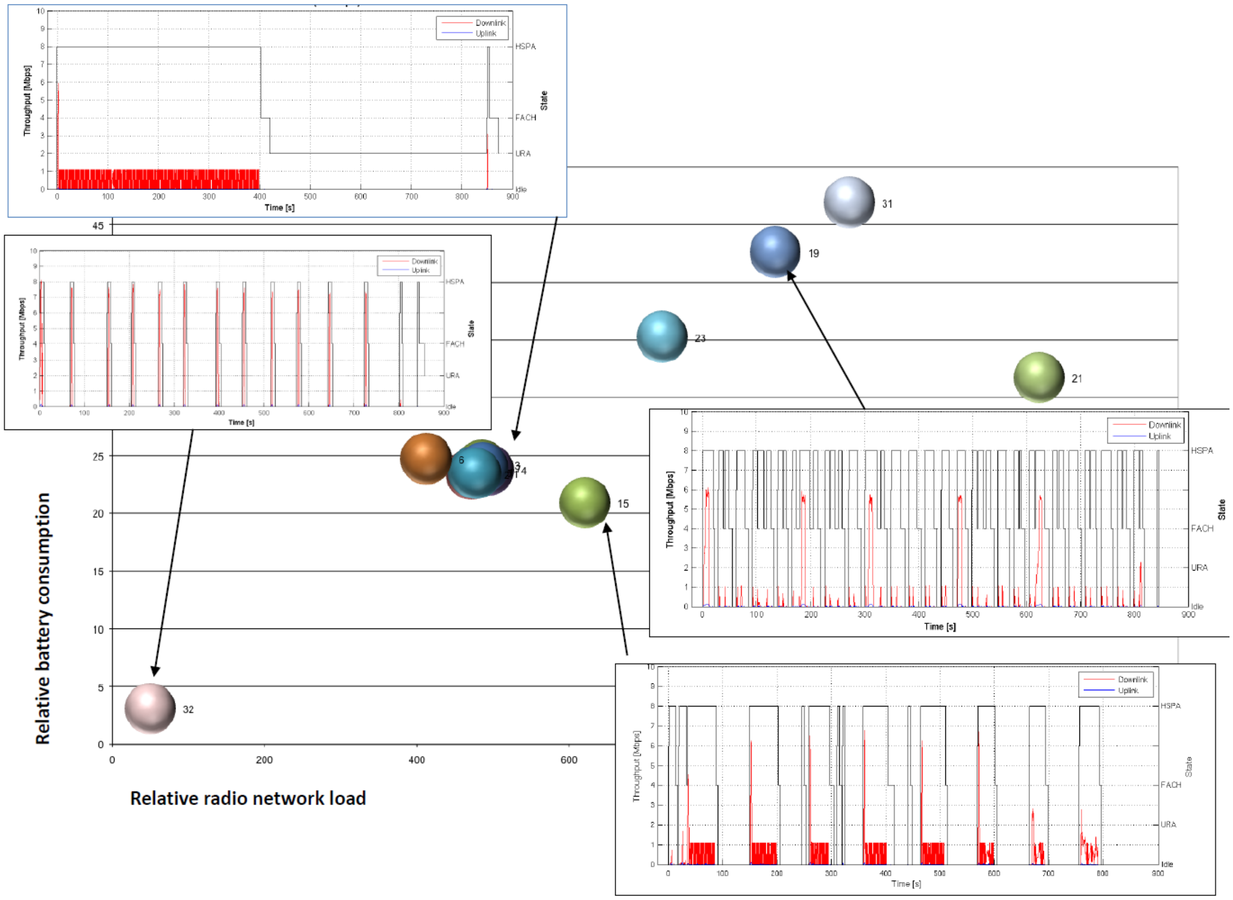


Figure 9: Effect of video playout buffers on Relative network load (X-axis) and Relative battery consumption (Y-axis) [17]

In all four examples, the following legend applies:

- Red lines correspond to download of video data.
- Black lines correspond to signaling state transitions for the device between different WCDMA activity states e.g. IDLE, URA/CELL-PCH, FACH and DCH.

On the main graph, the x axis is “Relative radio network load” and the y axis is “Relative battery consumption”. The x axis in each of the four examples is “Time in seconds” and the y axis is “Throughput in Mbps”.

Consider the various use cases shown in Figure 9. The arrows in the figure point to the respective graphs referred to below.

1) **Short Bursts (upper right corner):** Here the playout client algorithm plays pulls video from the network with multiple short bursts using multiple network accesses with small video downloads at each network access. This “nibbling” at network resources results in multiple short transitions between states, and small data downloads. This results in high network load and high battery consumption due to the high signaling traffic.
2) **Trickle Download (bottom right corner):** Here the playout algorithm accesses the network for longer periods of time, with a low amount of downloads during each access. Both the network

load and battery consumption is reduced compared to the upper right example due to the more optimized network accesses.

3) **Download / Progressive Download (upper left corner):** In this example, the signaling transitions are minimized, and video download is sustained. However, this algorithm is inefficient for the majority of cases where the end user does not watch the entire video since unnecessary video data is downloaded, thus resulting in waste of battery resources.
4) **Adaptive Streaming (bottom left):** This is the most efficient of all algorithms from a relative battery consumption and network load perspective. Here the algorithm downloads an amount of video data that is statistically adequate for the average end user, and minimizes radio bearer state transitions while doing so.

It is therefore important that the design of video download algorithms be done prudently based on efficient use of network and device resources.

3.3 CHALLENGES FROM THE NETWORK

Video delivery can pose considerable challenges outside the RAN, in the backhaul and core network.

3.3.1 BACKHAUL CONGESTION

The term "backhaul" refers to the transport network that connects the core network nodes with the radio access network and represents a crucial part in any wireless architecture, as shown in Figure 10. In most advanced wireless networks, the backhaul may become the bottleneck given the huge bandwidth requirements that result from the aggregation of multiple cells, unless appropriate dimensioning is made in order to limit congestion. Multiple links are usually multiplexed in the aggregation part of the transport network prior to connecting to the core.

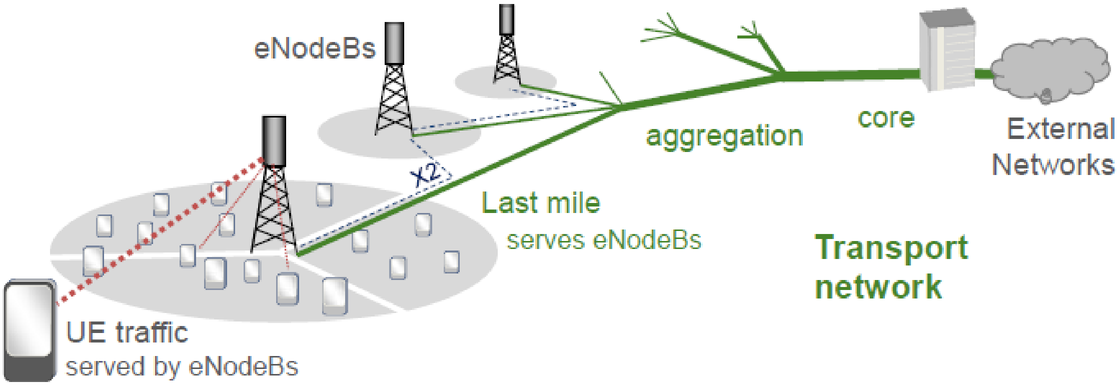


Figure 10: Typical backhaul architecture [18]

Backhaul capacity should be able to absorb the peak cell throughput in low load situations, where one or few users in very good radio conditions can be granted full cell resources. Since mobile video applications require significant data bit rates, appropriate backhaul dimensioning is critical to the quality of experience for mobile video users. In the aggregation network, a number of backhaul links are multiplexed prior to connection with the core network and the resulting backhaul costs would clearly be not affordable if the peak cell throughput is to be guaranteed for all the cells. Instead, when the number of sites grows, it is unlikely that all resources are simultaneously served at a given instant, and appropriate dimensioning based on statistical multiplexing of the cells' traffic is therefore applied. Backhaul congestion may however appear whenever traffic exceeds the backhaul capacity at some point.

NGMN provides some practical guidelines for provisioning of the backhaul network. Given N number of cells, one common approach for the minimum required backhaul is:

Backhaul provisioning for N cells = max (N x busy time mean, peak)

The busy time mean corresponds to the average throughput served by each cell in the busy hour, where multiple users with different radio conditions are simultaneously active. The peak cell throughput is usually the 95-percentile user throughput under light load conditions. These values should be affected by appropriate overhead from the control plane, OAM, synchronization, transport protocol and IPSEC (if applicable).

The above formula states that, for large values of N, the average traffic should dominate the backhaul capacity instead of the peak traffic. The difference between them is illustrated in Figure 11. Simulations conducted by NGMN members for typical LTE scenarios show very low throughput values in the busy time: as an example, a typical 150 Mbps LTE cell with 2X2 MIMO in downlink would reach a mean throughput of less than 25 Mbps per sector. This high peak-to-mean ratio suggests that significant aggregation gains are available at the transport network.

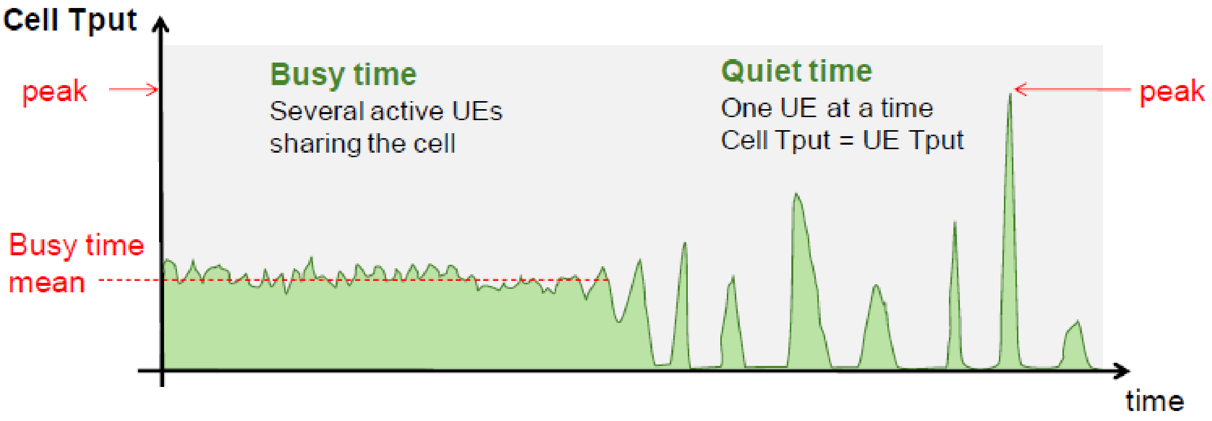


Figure 11: Cell throughput during busy time and quiet time [18]

Backhaul congestion may have different consequences depending on the type of video application. In streaming video services, congestion can result in instantaneous reduction of the available bit rate, thus appearing as pixilation and other visible artifacts. In HTTP Progressive Download over TCP, the TCP protocol would react to congestion by reducing the server bit rate, but this could be absorbed by the client's buffer with no appreciable quality degradation if the congestion episode is isolated. In HTTP adaptive streaming, the instantaneous bit rate would be reduced accordingly, thus avoiding significant degradation unless congestion forces the bit rate to decrease much below the screen resolution.

The majority of the end-to-end delay is expected at the radio access part (due to retransmissions, scheduling, fading etc.) and much reduced delay budget values would be advisable for the backhaul network. Based on select 4G Americas vendor implementations, the recommended delay budget values for the backhaul are well below 20 ms for conversational and streaming video, less than 20 ms for medium-quality videos and less than 5 ms for high-quality videos.

Digital video decoders used in streaming video receivers need to receive a synchronous stream, with tight jitter tolerances of approximately only +/- 500ns [19], in order to decode without visible impairments. Such jitter tolerances may not be achievable natively in packet data networks. Hence, broadcast video services use de-jitter buffers (also called playout buffers) to remove delay variations caused by the network. Furthermore, the jitter experienced by mobile users, if not compensated by appropriate buffers, can be alleviated at the network side by the introduction of traffic shapers at the gateway nodes located at the boundaries of the mobile network. These traffic shapers can transform undesired bursty traffic into more regular patterns which can be more easily managed by base station schedulers. Limitation of the maximum bit rate according to the operator's policies and user subscriptions can also be performed, in order to prevent starvation of non-video users due to hungry video sources. Policy and charging rules can be defined by the operator and stored in a core network node called PCRF (Policy and Charging Rules Function), which interacts with other core network nodes for enforcement of the operator's policies for the various video users. Encryption can be performed at different layers and on different links. For instance, the link between the RAN and the UE may be ciphered. There may also be encryption applied at the IP layer, such as IPSEC.

3.3.2 DATA PLANE/CONTROL PLANE ENCRYPTION

Mobile video enables users to consume a variety of video content both on the go and while stationary. For the streaming use cases, the content may be commercially produced or user generated. Because of usage policies on commercial content as well as privacy concerns for personal content, it may be necessary to encrypt the data and control plane traffic. Here the control plane traffic refers to both the logical signaling channels, and also the signaling over IP used by the video service providers. In some cases, it may be necessary to encrypt the video content to enforce usage policies and user privacy.

3.4 CONTENT BASED CHALLENGES

3.4.1 DIGITAL RIGHTS MANAGEMENT (DRM)

All video content travels through the data plane in order to arrive to the user UE from the source. The source may be another UE or a video server. The DRM is the use restriction based on policy. In some cases, it uses encryption, but not in all. In majority of DRM technologies, embedded watermarking is used to prevent copyright infringements and multiple unrestricted reuses of contents. In order to allow legitimate use of encrypted information, the Security Key Management and Distribution must be deployed. Without it, the confidentially protected video data and control information is reduced.

3.4.2 ENTITLEMENT/CONTENT MANAGEMENT RIGHTS

For video playback control to be implemented, a control protocol must be used. In this case, this protocol may need to be encrypted as well for privacy concerns. This encryption can protect against rogue actions such as: reverse engineering what programs a particular user may be watching.

Control protocols are often proprietary to each video service provider. It is generally up to the video service provider to provide this protection to the control protocols.

4. VIDEO QUALITY OF EXPERIENCE

4.1 VIDEO USAGE BEHAVIOR AND EXPECTATION OF EXPERIENCE

An important consideration while devising the best video optimization strategy is to understand the typical mobile user's video usage behavior and their expectations. At home or at work, most users connect to Wi-Fi connection and hence their video traffic is rarely seen on the mobile network. But when the person is waiting for a train in the subway, or when the person is waiting in the queue on a local supermarket he/she is likely to quickly take a peek at what's happening around the world. Thus, most mobile video sessions are of short duration (5 – 10 minutes long), and often the video is not watched in its entirety.

Similarly, from an expectation standpoint, the key metric of relevance to the user is a smooth and uninterrupted video delivery. This alone, more than anything else, determines the subscriber's satisfaction rating [20].

4.2 METRICS FOR CHARACTERIZING VIDEO QOE

Quality of Experience (QoE) is concerned with the overall experience of the video consumer. It goes beyond the Quality of Service (QoS) used to deliver the video. To end users, quality is a subjective term – "they know it when they see it." In the multimedia world, the measure Mean Opinion Score (MOS) is often used to quantify quality of experience.

MOS is a numeric value between 1 (one) and 5 (five). It was originally developed for bench marking audio transmission and reproduction. Five meant the listener could not detect any impairment. One meant the impairments were very annoying. Clearly, this is a very subjective measure. By taking the mean of a large enough sample of opinions, a reasonably accurate measure of QoE can be achieved. Before considering the application of MOS to video, it is important to identify the issues that can affect the end users QoE.

Quality of the Encoding – The original video source can be in very high quality. However, in order to deliver through mobile networks, it needs to be compressed. Even a standard definition uncompressed video signal can require 270 Mbps, while high definition video can require 1.2 Gbps without compression. This process is lossy and will introduce artefacts into the video image. Modern video compression algorithms and associated pre and post processing steps are very effective in encoding the video. However, typical problems include:

- *Blurring* – This can be caused by unfocused camera lens when the material was originally captured. It can also be introduced by the encoder. The effect is for a part of the image to have a loss of detail that is perceptible.
- *Blocking* – This is caused when a part, or all, of the image is made up of blocks of the same colour. The effect is the appearance of very large pixels. It is often most noticeable on lines.
- *Quantisation errors* – Where the original image has an expanse of similar colour but gradual change in luminosity, the decoded and rendered image consists of jumps in colours.

Some scenes are very difficult to encode. When compression is high, scenes with a lot of information including hard lines at an angle often show blocking errors. The sky and other areas of similar colour but different luminosity show up quantisation errors. High motion is another challenge as it requires more bits to describe how parts of the image have moved frame by frame. However, some motion is relatively easy to encode because large parts of the scene move together – panning across a landscape for example.

The very hardest are scenes with a lot of information and chaotic motion (waves on water and grass/trees moving in the wind).

Device – The screen size has a significant impact on QoE. The same encoding viewed on a smartphone can look adequate, poor on the larger screen of a tablet and totally unsatisfactory on even a modest TV.

Viewing Distance – The distance the display device is held from the eye makes a big difference on the ability to resolve the image. The closer the display device is to the eye, the higher the quality of the image has to be. Note, watching a video on a tablet at conventional distance has a very similar viewing angle as watching a large TV from the couch.

End User Expectations -- If people pay for the content (PPV, subscription, etc.), then their expectations of video and audio quality are high. Also, expectations will be high for serious entertainment content (TV show, film, and sports events) compared with short YouTube clips. User expectations vary depending on the type of content and over time, they become more demanding and expect improved quality.

Playout Issues – Even if the content is prepared to meet the quality expectations for users on a particular display device, network issues can cause problems which manifest themselves in visual and / or audio artefacts. Buffering results in a streaming media being "preloaded" to allow for smooth playback. Dropped IP packets can result in picture breakup unless they are masked. Network issues can cause jerky playout (low frame rate), skipping (especially in live streams) or even picture freezing (due to rebuffering which is the reloading of a stream to prevent skipping). The latter must be avoided if at all possible [20]. Usually less annoying, but still important to the overall QoE, is lip sync. If the audio is out of synchronization with the video, the experience is of watching a badly dubbed film.

System Responsiveness – There is a certain maximum acceptable time beyond which if there is a significant delay between requesting a video and it is starting to play out, then users will experience dissatisfaction. Likewise, the system needs to be adequately responsive to channel change in order to meet user expectations. Hence a streamed solution is generally preferred over (progressive) download.

Ease of Use – The complete experience involves discovering the content to be viewed and optionally paying for the content. Providing a user interface making it quick and easy to find the desired content is essential to a good QoE.

Varying Image Quality – To handle differences in available bandwidth, many video solutions use some form of adaptive streaming. The consequence is that when the available bandwidth is low, the image quality will become worse. Conversely, improvement in bandwidth will result in improvement of the image quality. Both the uncertainty and change in image quality will impact the QoE. It is still an area of research on how much temporary poor image quality impacts QoE.

Finally, IEEE reference [21] demonstrates the relationship between viewer QoE (quality of experience) and MOS (Mean Opinion Score) and the HTTP Adaptive Streaming Video Quality (VQ) rate. Rather than using mean MOS scores to judge the Video Quality (VQ) rate, it is shown that MOS scores are impacted by the variance (or standard deviation-SD) of the VQ rate more so than the mean VQ level. In other words, users care more about maintaining a consistent VQ level than they do in a higher bit rate. It is shown that one can accurately predict MOS scores incorporating both the mean VQ level and the standard deviation of the VQ level.

4.3 BEHAVIOR OF VIDEO CHUNKS WITH TCP/IP

HTTP adaptive steaming (HAS) is becoming the ubiquitous and reliable method of streaming video content over service provider networks to a sizable number of classes of devices, including smartphones, wireless tablets, laptops and desktop personal computers. The HAS client continually senses the available bandwidth and adjusts the bit rate of the requested chunks to ensure the playout buffer does not run dry. While HAS provides a better implementation over the traditional HTTP progressive download mechanism by dynamically adjusting the video quality to match the available bandwidth (see section 2.6), the nature of the protocol has several impacts on the network.

The performance of HAS depends on the client implementation, which is not at all defined by the standards. A well-designed HAS client will try to get the highest quality possible, but it will also back off when there is competing traffic, including other video streaming clients. A well-designed HAS client should provide the best experience possible given current network conditions, but should not try to provide better playback than is possible based on current network conditions. The HAS client is based on HTTP, and thus cannot be any more greedy than any other HTTP-based application. If the HAS client tries to be greedy and download a bit rate that is too high to be supported by the current download rate available using HTTP over the network, then the end user will have a bad experience. This is because the content will not be downloaded fast enough to playback in full time, and thus the playback will stall or skip. Thus, a well-designed HAS client will adjust which bit rate it is downloading to match the current download rate of HTTP and provide a good user playback experience. A badly designed HAS client will provide a poor end user experience, but will not congest the network since it is using standard HTTP -- same as all other applications downloading content over the network. This is one of the reasons that HTTP Adaptive Streaming is especially well-suited for streaming to mobile devices.

Policies should be based on providing proportional fairness (however that is defined) to the mobile clients, and probably most, if not all of them, are using HTTP to download content. A HAS client is subject to these policies, just like all other applications, and if the HAS client tries to be greedy and download too high of a bit rate of content then the end user suffers, not the network, as the playback of content will skip and/or stall. However since the HAS client is using HTTP to download, it will not be able to greedily take over the network, and the network can control the total amount of traffic that the mobile device receives.

The performance of HTTP Adaptive Streaming with some implementations that include poorly designed HAS clients may suffer due to the following issues.

Unfairness and Instability: The HAS client continually senses the available bandwidth and adjusts the bit rate of the requested chunks to ensure the playout buffer does not run dry. The problem is that a client only senses the available bandwidth when it is downloading a chunk. This can result in unfairness. For example, if one client is already downloading a chunk at a high bit rate, then the second client will sense only limited capacity and will chose to download a small chunk. This imbalance in bandwidths can remain for some time. Similarly, two clients can oscillate in their selection of chunk bit rate requests as their downloads periodically overlap and are distinct. Proper network based policies can mitigate these problems.

Under Utilised TCP throughput: Some issues still exist with the efficiency of delivering HAS over the reliable TCP transport layer that will need to be overcome to better utilize the resources of the bearer channels. When comparing TCP versus HAS throughput, it appears that TCP theoretical throughput is often greater than HAS TCP throughput, in some cases, by as much as twice. The unutilized TCP capacities exist in HAS thus limiting the performance of HAS, thereby reducing the quality of experience for viewers. Looking at the HAS performance with respect to round trip time delay (RTT), RTT dramatically impacts the TCP throughput and further reduces the HAS bandwidth HAS client. Figure 12 compares the performance of HAS TCP versus theoretical TCP [22].

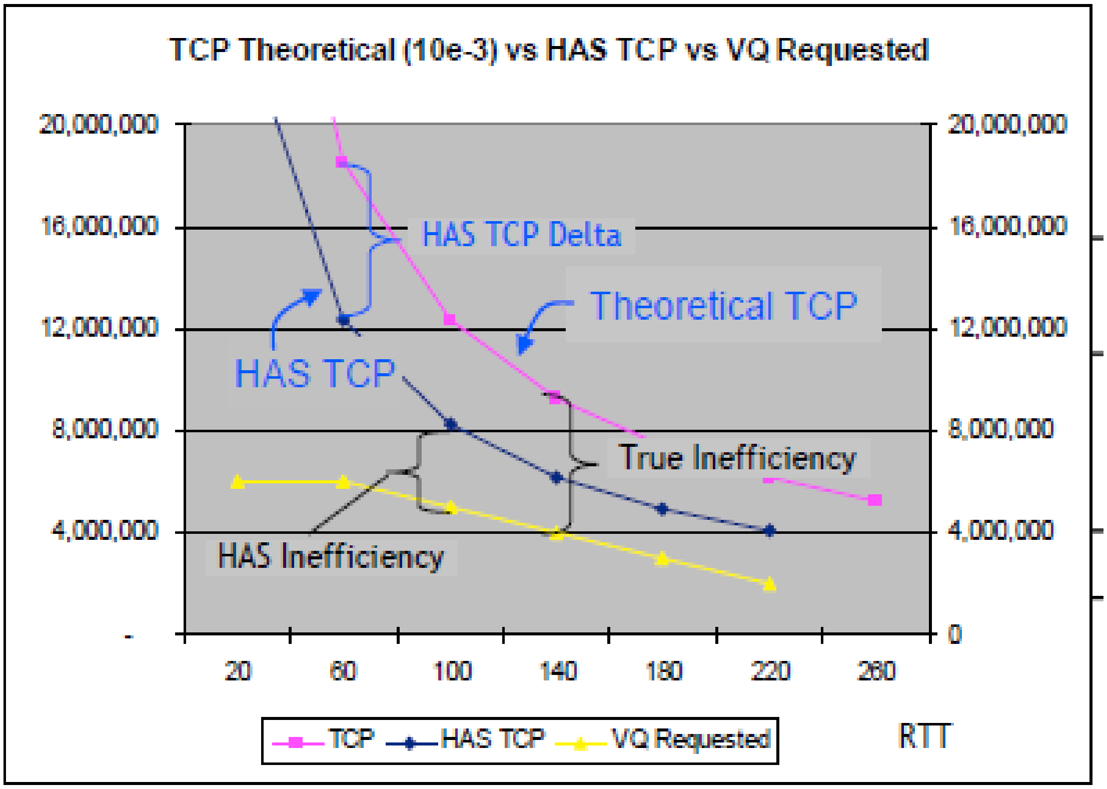


Figure 12: Performance of HAS TCP versus Theoretical TCP [22]

The reason for the lower HAS TCP performance is the interaction between the HAS rate determining algorithm (RDA) and the TCP congestion avoidance algorithm. Since HAS relies on the measure of the available bandwidth to decide which bit rates chunk to request, HAS clients either present a delay before starting playing the video or start immediately playing the lowest available quality. As video chunks get downloaded, the bandwidth estimation is adjusted. This slow start estimation coupled with a possibly conservative approach by the HAS RDA greatly impacts the end-user experience, since reaching the optimum rate for the current network conditions can take up to 20-30 seconds. After the initialization phase, HAS is then prone to suffer from the “downward spiral” [23], shown in Figure 13, whereby the client ends up requesting the lowest bit rate when competing flows or network congestion result in bandwidth fluctuations. This spiral towards the worst quality comes from a constant under-estimation of the available throughput due to the combination of several factors, like periodic request pauses to let drain the playback buffer, repetitive TCP congestion window reset, slow start TCP ramp-up and smaller video chunks requests over time. Determining the fair share of band-width available at the bottleneck is precisely the role of TCP, but this information is lacking at the HTTP layer.

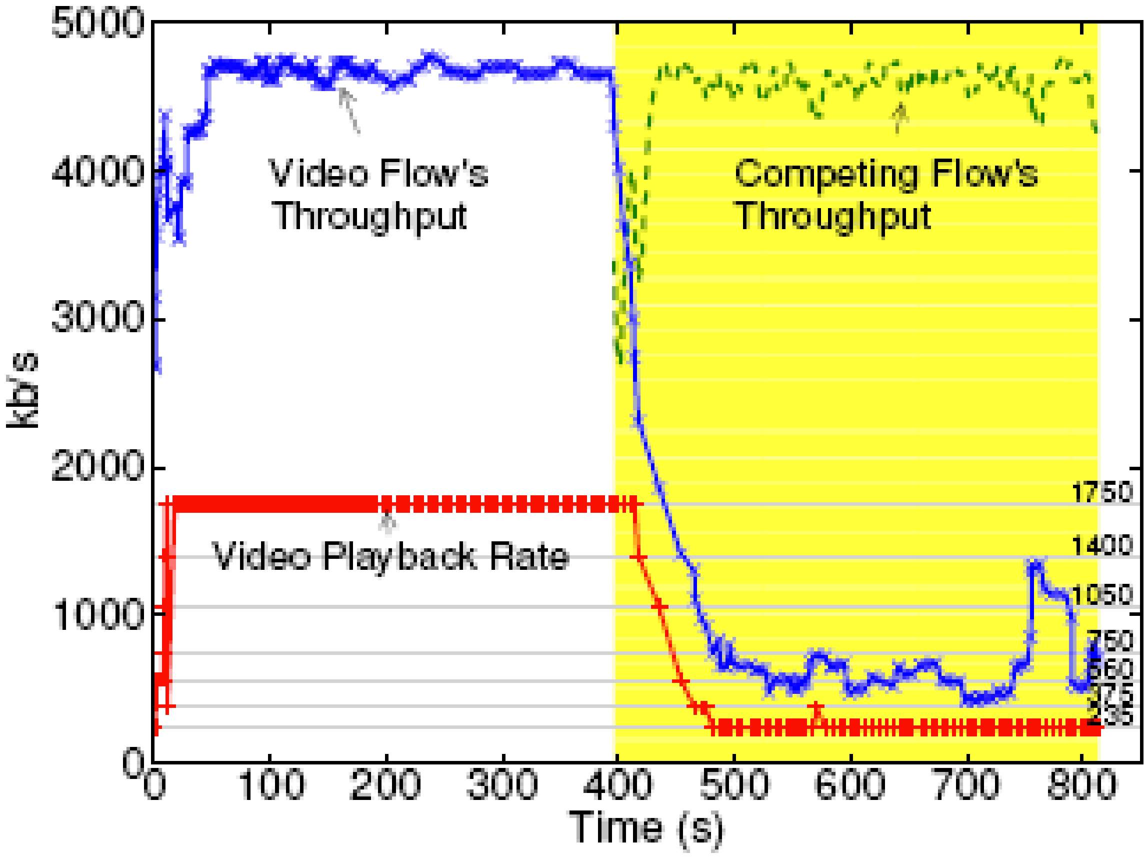


Figure 13: Downward spiral behavior of video throughput [23]

Clearly, the performance of HAS is very dependent on specific client implementations, and well designed client HAS implementations can get around some of the issues and help provide good Quality of Experience for mobile video users.

Finally, some versions of TCP cannot distinguish between packet loss caused by congestion and packet loss that results from noise (transport errors). As a result, TCP often cuts window size in half when it should not, thus decreasing its efficiency over noisy channels such as wireless links. Service providers could also take advantage of Early Congestion Notification (ECN) and Selective Acknowledgements (SACK) mechanisms to make TCP more robust. ECN helps improve efficiency in a wireless environment by allowing TCP to partially distinguish between packet loss due to errors and packet loss due to congestion. Instead of waiting for packet loss to signal congestion, TCP is notified when a buffer is filling up. With Selective Acknowledgements (SACK), to avoid excessive retransmission of packets, the TCP receiver more carefully indicates which packets were lost, rather than retransmitting everything, starting with the first lost packet that was reported. This approach is especially valuable when the packet's round-trip time (RTT) is large. Additional details are available in RFC 3481 [30].

Wireless links can also be protected by link layer mechanisms that are designed for unreliable links. Forward Error Control (FEC) and Hybrid Automatic Repeat ReQuest (HARQ) help reduce packet loss between the endpoints of the link. However, they also introduce additional delay and overhead bit rate. Service providers need to evaluate how well this trade-off meets their own unique requirements.

HAS/TCP interactions is an area of active research by various authors including [11], and an interested reader is referred to various contributions from these authors and others. Careful network conditioning, new mechanisms/policy and improved client rate determining algorithms can help overcome many of these issues to reduce their effect and loss of performance on E2E QoE. Through subjective testing, it has been shown that HAS delivers an acceptable video quality over a range of typical network conditions. Thus, HAS offers a cost-effective way to deliver more video content to more customers, while reducing costs through efficient use of network resources.

4.4 MONITORING OF VIDEO QOE

In order to determine a mean opinion score (MOS) value, you need to get a number of people to view and rate a set of clips on a scale 1 (poor) through to 5 (excellent). The Video Quality Experts Group (VQEG) has published guidelines on how to measure MOS. The problem is that this is a subjective measure, and the radio network is a dynamic delivery network. The available radio resource assigned to a terminal varies continually depending on the radio condition of the terminal and on all the other terminals in the same cell. Hence, an objective measure of QoE is needed.

There are three classes of monitoring Video QoE as follows.

1) **Full Reference** – In this approach, the final displayed video after encoding, transmission and decoding is compared with the original source material. This allows for very accurate measurement of impairments. However, it requires access to the original material and involves significant amount of processing. It is used by broadcasters to ensure the integrity of their encoding and transmission equipment where high quality is essential.
2) **Reduced Reference** – In this approach, some information about the original material is used. For example, checking the integrity of a watermark inserted in the original before any processing. The degree to which it has survived is an indication of video quality.
3) **Non-Reference** – In this approach, no information about the original material is used. Quality is assessed by analyzing the video itself and/or the underlying network transport. This can be done in real-time with small processing effort. It can identify blocky and jerky videos.

There are two common Full Reference objective measurement techniques, PSNR and SSIM. Peak-Signal-to-Noise Ratio (PSNR) is a simple measure of image quality based on the error between the original and final decoded image. PSNR measures the fidelity or how similar the sequence is to the original. It is an objective measure which is easy to compute and is well understood. However, the correlation to subjective quality measure is poor. PSNR only considers the luminance component and neglects the chrominance component, which is also important to human perception. Furthermore, it does not do a good job of comparing different sequences.

Structural SIMilarity (SSIM) is a more advanced objective measurement. While PSNR measures errors between the original and optimized image, SSIM measures the structural distortion, luminance and contrast differences between the two images. The idea behind SSIM is that the human vision system is highly specialized in extracting structural information from the viewing field but is not specialized in extracting errors. Thus, a measurement on structural distortion should give a better correlation to the subjective impression. However, the mathematical complexity of this calculation is comparatively high and the SSIM calculations might vary from that of the subjective measurements.

Both PSNR and SSIM require access to the source material. Therefore, they can be used for operators who are providing their own video service. They are not suitable for OTT video streamed through the mobile network. This is becoming increasing popular with the extension of TV Everywhere to mobile devices and LTE, thus making it feasible to receive good quality content.

Non-Reference approaches can give good results. For professional content, a common approach for streamed content is with some form of HTTP Adaptive Streamed (HAS) content (DASH, Apple HLS, Microsoft Smooth Streaming).

As described in Section 2.5, in HAS, the content is divided into short non-overlapping intervals of between 1 and 10 seconds long. Each interval is then encoded in multiple bit rates which are then sent over the radio interface in chunks. The HAS client uses HTTP to request chunks. By monitoring the sequence of HTTP requests, it is possible to reconstruct the chunks asked for in an individual session. From this, it is possible to reconstruct the entire stream. Work by the NGMN P-SERQU project [24] shows how to map the quality level of the requested chunks to a predicted MOS value for the session. This approach works even for encrypted content.

3GPP in their recommendation 26.247 has defined the following parameters which should be monitored for HAS content. The sections below refer to the sections in 3GPP 26.247.

- *List of HTTP Request / Response Transactions* (Section 10.2.2). This records both when the request is made and which quality level is requested. It allows a reconstruction of the quality of a session including estimating if there was a hesitation in playout.
- *List of Representation Switch Events* (Section 10.2.3). Quality of Experience is affected by changes in quality. With HAS, these changes can be abrupt and very noticeable.
- *Average Throughput* (Section 10.2.4). For any one codec, the higher the bit rate the better the image quality. However, popular codecs such as AVC (H.264) are highly non-linear.
- *Initial Playout Delay* (Section 10.2.5). If the delay in requesting an on-demand event and payout is much more that 2 seconds, QoE is significantly reduced. For channel change to live content, it is sub 1 second.
- *Buffer Level* (Section 10.2.6). If the buffer level in the client drops to zero, it results in a freeze which is very annoying and should be avoided.
- *Play List* (Section 10.2.7). This is the information on the content being played.
- *MPD Information* (Section 10.2.8). The information in the manifest file is used to reconstruct a session.

Work is ongoing in the NGMN Project “Service Quality Definition and Measurement” P-SERQU [24] to determine the Video Quality of HAS in a mobile environment. The aim of P-SERQU project is to evaluate the end user acceptance of streamed video over a next-generation (LTE) wireless network using HAS technology. Progressive download is excluded from the studies based on the observations of industry direction that favors HAS for paid video streaming services. From the outcome of the test approaches proposed in P-SERQU project, it is expected that it will be possible to correlate network level impairments with end user QoE.

Monitoring video at the RAN level is generally not feasible. This is because of the interaction of TCP congestion avoidance algorithm and the HAS rate determination algorithm. Also, it is often difficult to get access to the necessary data in the eNodeB. Main factors that influence end user QoE of streamed video are identified [24], including device and media factors and network factors. Monitoring of Video QoE through monitoring the quality of HTTP chunks rather than individual radio and TCP parameters is thought to be the preferred viable option.

Since HAS uses HTTP which runs on TCP, IP packet losses due to network impairments are masked by automatic retransmission. So there should not be any visual and audio artifacts caused by IP packet loss unless the period of extreme link impairment resulting in total outage exceeds the buffer time. Rather, the effect of network impairment is to reduce the TCP bandwidth. The HAS client adapts to the reduced TCP bandwidth by selecting chunks of a lower bit rate. Hence, there should be no pause in playout while the client rebuffers as long as the link outage doesn't exceed the buffer time. However, the end user will see a change in video quality as the RDA optimizes the highest bit rate is can pull through the network while maintaining a buffer of video (typically 30 seconds deep). The lab tests in P-SERQU [24] should show how much the impairments impact HAS/TCP bandwidth, and the mass tests that are a part of that study will reveal how much the changes in the HAS chunk bit rate results in reduced MOS.

4.5 SPECIFIC ATTRIBUTES THAT CONTRIBUTE TO IMPROVED VIDEO QOE

In principle, a higher network bandwidth results the higher the QoE. However, RAN bandwidth is a scarce shared resource. Therefore, it is sensible to choose encoding bit rates which provide adequate image and audio quality on the target devices. For most content, this is 200-500 kbps for smartphones and 500 kpbs to 1 Mbps for tablets when using H.264 video encoding and stereo audio. Of course, screen resolutions are improving so we can expect increased demands for improved quality, especially if people use their tablets to stream to TVs.

A video stream lasts for several minutes up to hours. This puts significant pressure on the demand for sustained bandwidth. Playout freezing must be avoided. This means that when using adaptive streaming, at least the lowest quality level must be maintained despite poor SINR and high cell congestion.

Maintaining an even bit rate at the RAN level can help eliminate variable quality levels in the viewed video. Furthermore, minimising the impact of dropped IP packets which results in either image impairment (RTP/UDP based video transports) or reduced throughout leading to reduced quality (TCP based transports) is critical. Hence, using GBR or QoS to favor forwarding of video related packets will improve QoE but at the expense of best effort data traffic.

Cell handover places particular demands on mobile video. The algorithm used in the buffering or playout of video from a source in the cloud can largely influence:

- network bandwidth use;
- battery usage associated to usage of resources;
- QoE as end users either wait for buffered video (as video plays) or suffer a delayed start in video play (in case of larger playout buffers).

Through a combination of techniques that are available now, and are being developed, it is expected that mobile video QoE will be acceptable and sufficient to support a large number of mobile video users.

4.6 NEW/UPCOMING CODECS (HEVC)

High-Efficiency Video Coding (HEVC) is currently being prepared as the newest video coding standard of the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG). The first edition of the HEVC standard was finalized in January 2013, resulting in an aligned text that will be published by both ITU-T and ISO/IEC. Additional work is planned to extend the standard to support several additional application scenarios including professional uses with enhanced precision and color format support, scalable video coding, and 3D/stereo/multiview video coding.

HEVC is a successor to the current state of the art video coder - H.264/AVC video coding standard. Since AVC was first standardized in 2003, some previous trends have sharply accelerated and totally new trends have appeared which are creating even stronger needs for coding efficiency superior to H.264/MPEG-4 AVC's capabilities. The traffic caused by video applications targeting mobile devices and tablets, as well as the transmission needs for video on demand services, are imposing severe challenges on networks. There is an increased desire for higher quality and resolutions for mobile applications.

The results of subjective tests for Wide Video Graphics Array and HD sequences indicate that HEVC encoders can achieve equivalent subjective reproduction quality as encoders that conform to H.264/MPEG-4 AVC, by using approximately 50 percent less bit rate. The HEVC design is shown to be especially effective for low bit rates, high-resolution video content, and low-delay communication applications.

HEVC adheres to the same hybrid video coding structure as H.264/AVC. It uses spatial and temporal prediction, transform of the prediction residual, and entropy coding of the transform and prediction information. Improved coding efficiency gains in HEVC compared to H.264/AVC are achieved with minimal increase in decoder complexity. It is estimated that HEVC decoder complexity increased 20 percent-30 percent comparing to High Profile of H.264/AVC, depending on architecture. In some respects, HEVC is even more implementation friendly than H.264/AVC. One of the H.264/AVC bottlenecks for hardware implementations is CABAC decoding. Context-adaptive binary arithmetic coding (CABAC) is a form of entropy encoding used in H.264/MPEG-4 AVC video encoding and is a lossless compression technique. CABAC is highly sequential and has strong data dependencies which make it difficult to exploit pipelining. Numbers of modifications were introduced in HEVC CABAC design to enable its parallel decoding. HEVC has been also designed to better support parallel processing architectures by adding parallel processing tools: Tiles (picture partitioning into rectangular non-overlapping regions) and Wavefront Coding.

5. ARCHITECTURES FOR MOBILE VIDEO DELIVERY

5.1 CELLULAR (HET-NET) AND WI-FI OFFLOAD ARCHITECTURES

In response to the rise in traffic demand, driven largely by the expected growth in video traffic, more radio spectrum is becoming available and more spectrally efficient technologies like LTE are being introduced. Operators are engineering their networks to provide increased capacity and coverage while maximizing the use of their limited existing spectrum assets. LTE is being deployed in areas where traffic demand is high. Such areas are typically already served by existing 2G/3G networks that support a mix of voice and packet data. LTE allows operators to use new and wider spectrum to complement existing 2G/3G networks with higher data rates, lower latency, and a flat IP-based architecture.

Considering the significant increase in traffic demands anticipated within the next few years, it is clear that the improvements that LTE macrocells alone bring will not suffice to meet demand. Operators will need to address the capacity crunch through a combination of more spectrum, increases in spectral efficiency, cell densification via small cells and heterogeneous networks (HET-NET).

Both HSPA and LTE technologies are near the Shannon bound for the cases of slowly-varying channels for which fading can be compensated by practical channel estimation algorithms. Link-layer performance has therefore little room for further improvements, and present investigations are more focused on increasing the experienced Signal to Noise ratios (SNRs) in the system. One way for improving the SNR is to increase the density of sites by introducing a collection of small cells under the eventual coverage of the macrocell, equipped with low power base stations at specific locations. These small cells can increase capacity and coverage by enabling better SNR values in the nearby, resulting in heterogeneous networks with enhanced area spectral efficiency [13].

Heterogeneous networks are characterized by the coexistence of macrocells and smaller metro cells, potentially increasing the number of cells tenfold or more in some cases. Heterogeneous networks utilize diverse technologies and base station types to improve spectral efficiency per unit area. They provide ubiquitous coverage along with the high-bandwidth capacity to deliver a superior QoE.

The move towards heterogeneous networks is being supported by work done within 3GPP. Within 3GPP specifications, LTE-Advanced wireless networks are being defined to improve spectral/spatial efficiency and improve the overall throughput by shrinking cell size via deployment of a diverse set of base-stations. LTE facilitates the deployment of heterogeneous networks since LTE has characteristics, as below, that ease the deployment of small cells that will result in more ubiquitous usage of small cells:

- Spectrum flexibility – LTE leverages more frequency bands so more spectrum is available;
- Improved spectrum utilization - flexible carrier bandwidth, support of carrier overlapping partly or totally with another carrier, support of carrier aggregation;
- Flexible resource management – carrier resource flexibility and the easing of mixing various types of cells using the same frequency resources;
- Full IP architecture - native support of flat IP architecture with low latency and standardized communication interfaces (S1, X2);
- Improved coordination - through better handover and load balancing abilities and through interference management features such as eICIC (enhanced Inter-Cell Interference Coordination) and its variants in Rel-8/Rel-9 and Rel-10. These features allow more users to be offloaded to the picocells to leverage the additional air interface resources, while still being able to address the additional interference to those users who may be in poor SINR conditions.
- Separation of the control plane from the bearer plane in LTE allows for independent scaling of corresponding network elements, aligned with the unique traffic demands.

In a broad sense, heterogeneous networks comprise not just of 4G macros and small cells but also of 2G, 3G and 4G cellular technologies along with their respective radio access options namely, macros and small cells, on same or different carriers, as shown in Figure 14. 3G Metro cells help alleviate congestion by providing capacity offload in dense urban environments with existing device ecosystems. They may also be used to provide indoor coverage, for enterprise users and for home users, where the high building penetration loss may result in the inability of the macro to provide adequate indoor coverage. Generally, video traffic is consumed by stationary/pedestrian users and hence the deployment of small cells to provide the air interface access for such users will significantly help alleviate macro congestion.

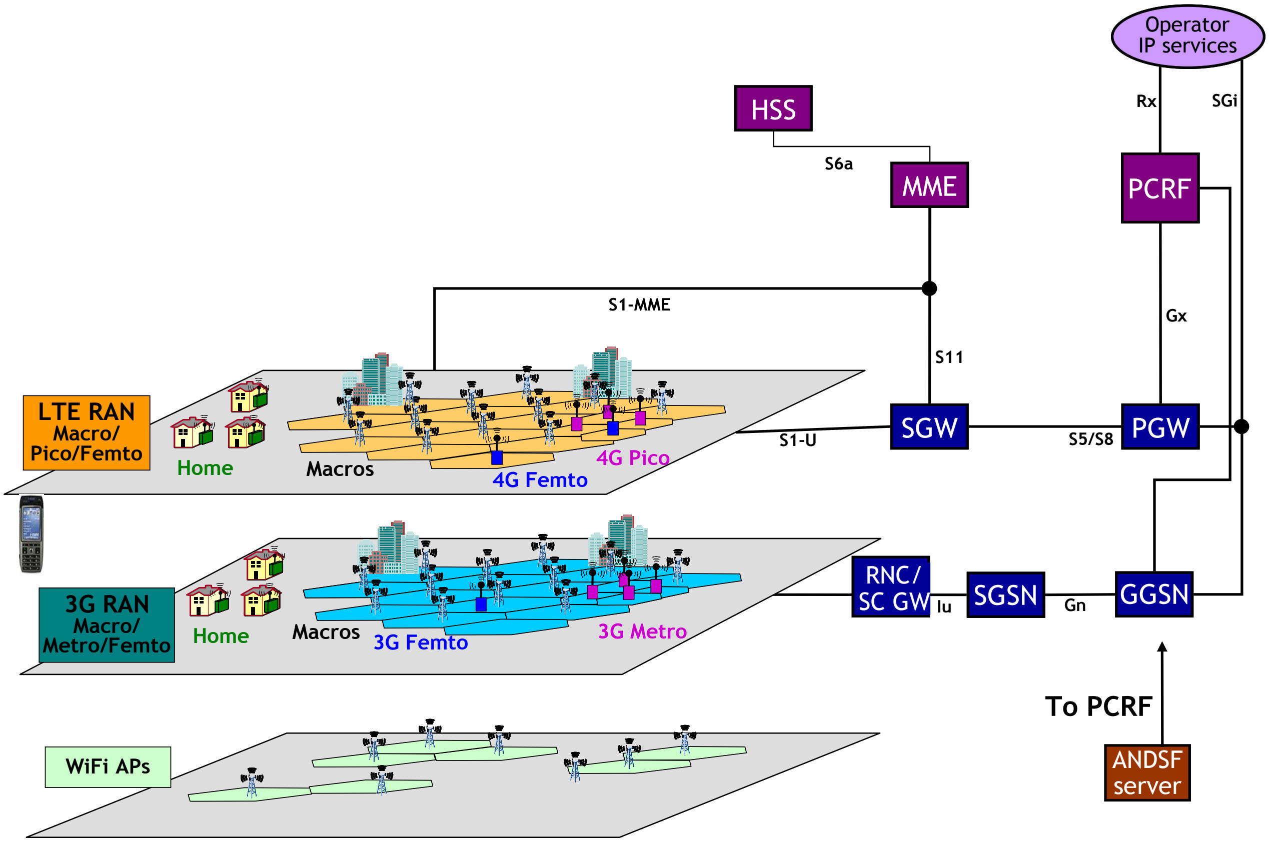


Figure 14: HET-NET Architecture

Through pooling of radio resources and through traffic management, the UE can be assigned to the appropriate technology, carrier and cell type depending on its mobility, service required and its location. In doing so, the QoE of all the various users is best met and a mix of technologies and access options can efficiently satisfy capacity and coverage needs in all environments.

With HET-NETS, the planned deployment of macrocells with a large RF coverage area (for the coverage of urban, suburban, or rural areas) are complemented with small cells with targeted RF coverage for coverage extension or network capacity enhancement, mainly in hotspot scenarios. Still the available capacity is far short of the capacity needed. To address the need for higher air interface capacity availability, mobile network operators are under increasing pressure to embrace small cells and Wi-Fi traffic offload solutions. Wi-Fi offloading from cellular to Wi-Fi networks with unlicensed spectrum can further help improve the performance (QoE) of the cellular users by conserving precious cellular resources. Users can be offloaded from cellular to Wi-Fi in an area of Wi-Fi coverage, or when the traffic on the cellular carrier exceeds a congestion threshold, or based on some specific policy mechanism using 3GPP based ANDSF mechanisms.

5.1.1 VIDEO USERS IN A HETEROGENEOUS NETWORK

With the introduction of heterogeneous networks, cell sites will no longer just be on towers – they will be located on lamp posts, utility poles, and on the sides of buildings. A variety of access types, fiber, copper, and microwave, will be needed to connect cell sites to the backhaul network. New backhaul topologies will be required to support LTE and the shift to heterogeneous networks. All must be able to support multiple types of wireless services including HSPA+, LTE and Wi-Fi.

Video users will more often than not be either stationary or move around the network at low mobility. If near a hotspot location, such users can be offloaded to a nearby picocell thus freeing up the macro capacity and allowing the pico to provide high data rates to the video user. The Quality of Experience for the video users served by picocells is likely to be higher than the corresponding QoE of video users served by macrocells due to proximity of the pico to the user and hence better RF propagation characteristics, including lower path loss. The bit rate and the RTT variations between macro, pico and Wi-Fi is dependent on the specifics of the network deployment including choice of backhaul, spectrum etc. Video users who are not nearby a picocell but are in a Wi-Fi hotspot location can be offloaded to the Wi-Fi network, depending on the operator policies, assisted by Wi-Fi ANDSF mechanisms and HotSpot 2.0. Traffic management across existing 3G and 4G networks can further facilitate the optimal use of radio resources (both bearer and signaling) across the various radio access technologies, frequencies, and cell types for both data intensive users such as video users, and signaling intensive users.

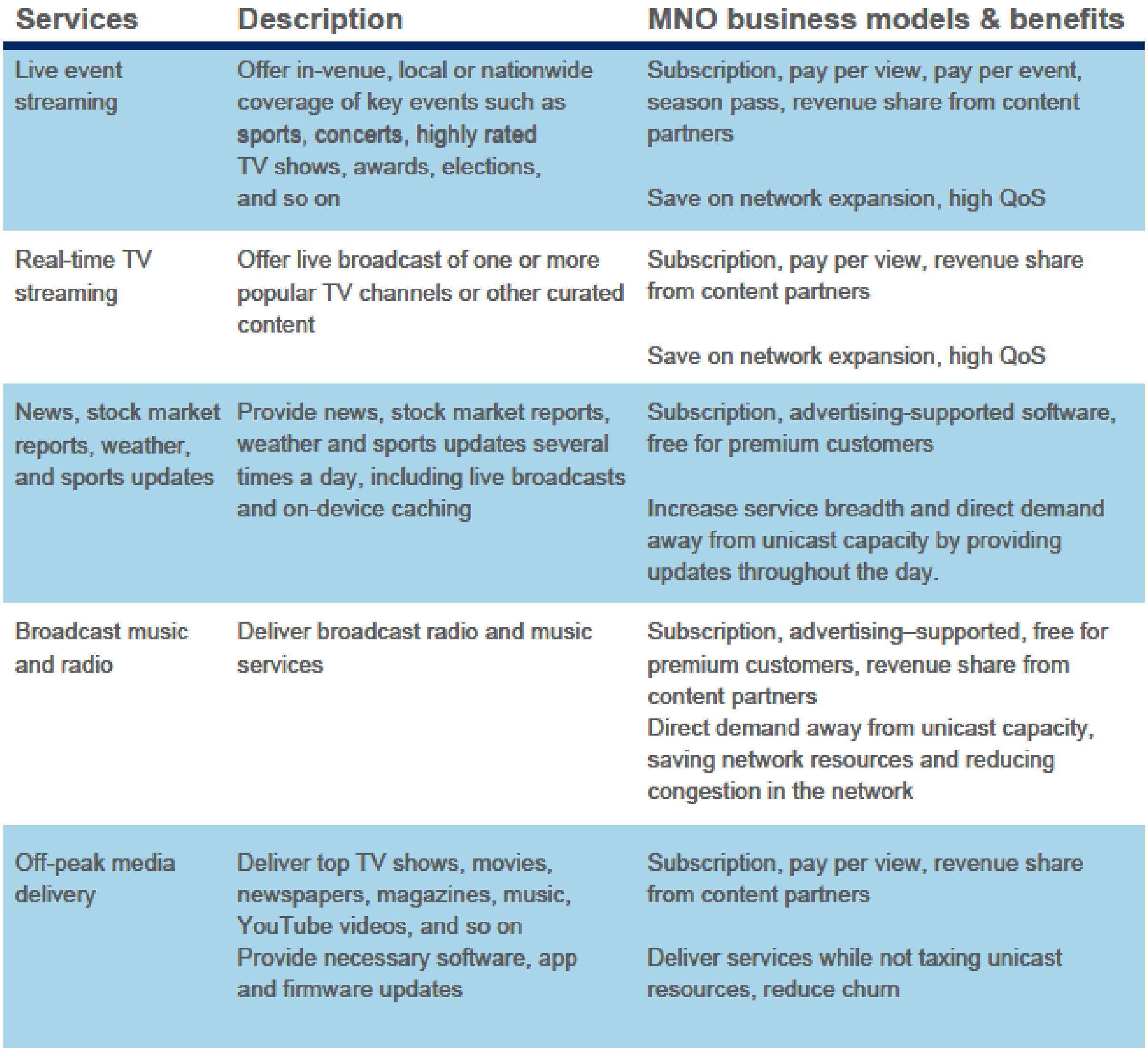
Video users moving around the network at low mobility will experience a drop in the SINR as they move through the interference zones between the macro and the picocells. However, the time taken for the messages in the handover signaling to go through will likely be acceptable, in spite of the SINR degradation, due to the low user mobility. Hence the video quality is not likely to suffer for users moving through the network at low speeds. At high speeds, degradation in video quality is more likely to be pronounced due to higher packet drops which then need to be corrected at upper layers. Higher speed video users may choose to remain on the macro cellular layer without handing over to the picocells, if it is a priori known that the user will remain in the range of the picocell for a limited period of time. Above a certain user speed, good video QoE may be difficult to sustain due to the demands on throughput.

5.2 LTE BROADCAST ARCHITECTURE

It is well established that there are significant spikes in mobile data traffic in many markets across the world due to major sporting events, breaking news, highly watched TV programs and celebrity events.
LTE broadcast, also known as Evolved Multimedia Broadcast Multicast Service (eMBMS), extends existing LTE/Evolved Packet Core (EPC) systems with an efficient point-to-multipoint (PMP) distribution feature. It allows operators to dynamically allocate network resources to offer highly flexible broadcast services thereby alleviating the strain caused by demand for popular content over unicast. The returns are particularly efficient when content demand is confined to a dense venue or is oriented to real time mass consumption. It is a Single Frequency Network (SFN) broadcast / multicast mode within LTE known as eMBMS which was defined in 3GPP Rel-8 and Rel-9. Enhancements to certain aspects continue in Rel-10 and Rel-11.

Using LTE broadcast, MNOs can handle these spikes in mobile usage and also offer premium content service in venues, city centers or nationwide, to create new revenue streams with well-controlled media quality and efficient scalability. Delivering video streams to hundreds of users in a cell site by using LTE broadcast will utilize almost the same bandwidth as a single user with the same video quality. The guaranteed quality of experience can increase subscriber loyalty and deliver significant service differentiation compared with the competition. There are multiple business models that MNOs can use, identified in Table 3 below.

Table 3: Potential use cases and business models for LTE Broadcast services [31]



Services	Description	MNO business models & benefits
Live event streaming	Offer in-venue, local or nationwide coverage of key events such as sports, concerts, highly rated TV shows, awards, elections, and so on	Subscription, pay per view, pay per event, season pass, revenue share from content partners Save on network expansion, high QoS
Real-time TV streaming	Offer live broadcast of one or more popular TV channels or other curated content	Subscription, pay per view, revenue share from content partners Save on network expansion, high QoS
News, stock market reports, weather, and sports updates	Provide news, stock market reports, weather and sports updates several times a day, including live broadcasts and on-device caching	Subscription, advertising-supported software, free for premium customers Increase service breadth and direct demand away from unicast capacity by providing updates throughout the day.
Broadcast music and radio	Deliver broadcast radio and music services	Subscription, advertising–supported, free for premium customers, revenue share from content partners Direct demand away from unicast capacity, saving network resources and reducing congestion in the network
Off-peak media delivery	Deliver top TV shows, movies, newspapers, magazines, music, YouTube videos, and so on Provide necessary software, app and firmware updates	Subscription, pay per view, revenue share from content partners Deliver services while not taxing unicast resources, reduce churn

There are two aspects to eMBMS. There are a set of modifications to the radio physical layer that enable SFN operation within the already defined LTE operating modes and is implemented at the modem level. eMBMS is designed to operate within the mobile operator's LTE network infrastructure and dedicated LTE radio spectrum. There is also an upper layer eMBMS framework that contains tools to enable services across the broadcast physical layer, often referred to as a broadcast service layer. The service layer includes support for file delivery and forward error correction (FEC).

The primary benefits of eMBMS relate to the increase in capacity that can be achieved via SFN operation, and the one to many aspect of broadcast distribution. These two aspects combined can provide a significant capacity lift for any content that is being viewed simultaneously by more than one user per cell/sector. The specific capacity numbers supported over eMBMS depend on sub-frames allocated for eMBMS configuration, and network configuration. Figure 15 shows a high level architecture for eMBMS.

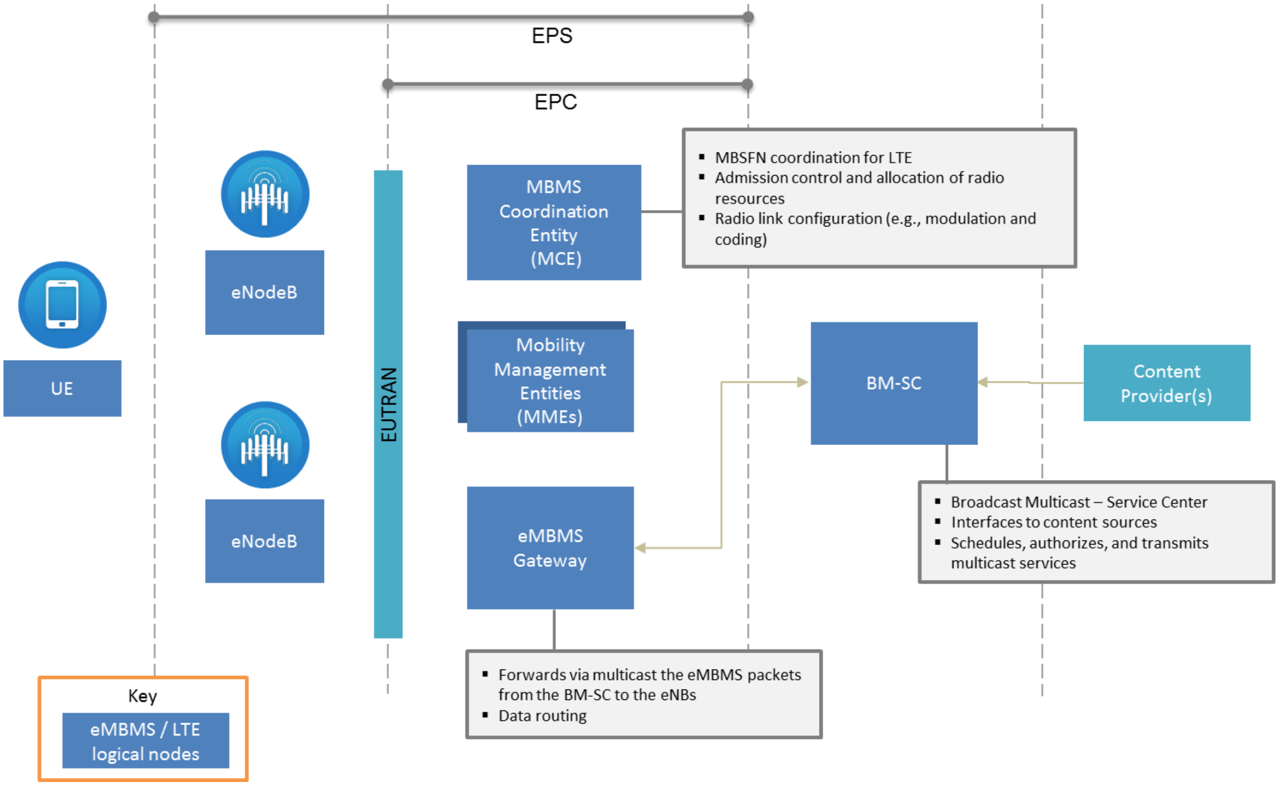

EPS
EPC
UE
eNodeB
eNodeB
EUTRAN
MBMS Coordination Entity (MCE)
MBSFN coordination for LTE
Admission control and allocation of radio resources
Radio link configuration (e.g., modulation and coding)
Mobility Management Entities (MMEs)
BM-SC
Content Provider(s)
Broadcast Multicast – Service Center
Interfaces to content sources
Schedules, authorizes, and transmits multicast services
eMBMS Gateway
Forwards via multicast the eMBMS packets from the BM-SC to the eNBs
Data routing
Key
eMBMS / LTE logical nodes

Figure 15: eMBMS/LTE Broadcast Architecture

The components in the architecture outlined in Figure 15 below are:

- **MCE** - MCE use Session Control Signaling (received from MME) to initiate configuration of radio resources to be used for broadcast transmissions.
- **eMBMS GW** - The MBMS-GW provides functionality for sending/broadcasting of MBMS packets to each eNodeB transmitting the service. The MBMS-GW uses IP Multicast as the means of forwarding MBMS user plane data to the eNodeBs. The MBMS GW performs MBMS Session Control Signaling (Session start/stop) towards the E-UTRAN via MME.
- **MME** - The MME provides Session control of MBMS bearers to the E-UTRAN access. It transmits Session control messages towards multiple E-UTRAN nodes eNodeB/MCE and it receives MBMS service control messages and the IP Multicast address for MBMS data reception from MBMS GW.
- **BM-SC** - The BM-SC handles eMBMS sessions (start, stop) and is responsible to deliver user plane media to the MBMS-GW.

The ecosystem associated with LTE Broadcast is under development, and is expected to accelerate when LTE Broadcast service is commercially launched. Some of the challenges for the industry include:

- Understanding the tradeoffs between dedicating spectrum for multicast technologies like LTE Broadcast and smart use of unicast video over RF.

- Monetizing of LTE Broadcast with a selection of video services that differentiate multicast over unicast.

- Building an ecosystem of content, spectrum and devices to address this optimized way of delivering video over RF.

The reader is referred to other papers that address the LTE broadcast issues in further detail [25, 26].

5.3 VIDEO DELIVERY ARCHITECTURE

Figure 16 shows a representative CDN architecture. A Content Delivery Network (CDN) is a system of distributed caches containing copies of data placed at various points in a network so as to maximize bandwidth for access to the data from clients throughout the network. A client accesses a copy of the data nearest to the client, as opposed to all clients accessing the same central server. This avoids bottlenecks near that server. Content types include web objects, downloadable objects (media files, software, and documents), applications and real-time media streams.

Video delivery begins with content capture and ingestion. The video that is created at the time an event is captured is compressed and transmitted to one of the content ingestion sites where video is archived and meta data related to the video is created and stored. The content is then prepared for delivery by coding the video into multiple different versions corresponding to different codec technologies, container formats, resolutions, and content variances (e.g. trailers or short clips, embedded captions), etc. Each of these versions may also be encrypted according to the Digital Rights Management technology adopted. Appropriate content portals or catalogues that exhibit the content availability are also updated to reflect the new content. Policies are defined for access restrictions and advertising policies that, respectively, control access to the published contents and advertisement placements.

Content distribution to local video headends follows the content ingestion step. In this step appropriate versions of video created during content ingestion step are sent over high speed networks to different video headends that serve the video to the end consumers. Video delivery begins in the origin server in video head end but typically involves additional CDN nodes that cache the content that are popular for delivering to the end users. CDN nodes may also be responsible for transcoding content into format appropriate for a particular device. Description of transcoding is provided in Section 6.4.2.

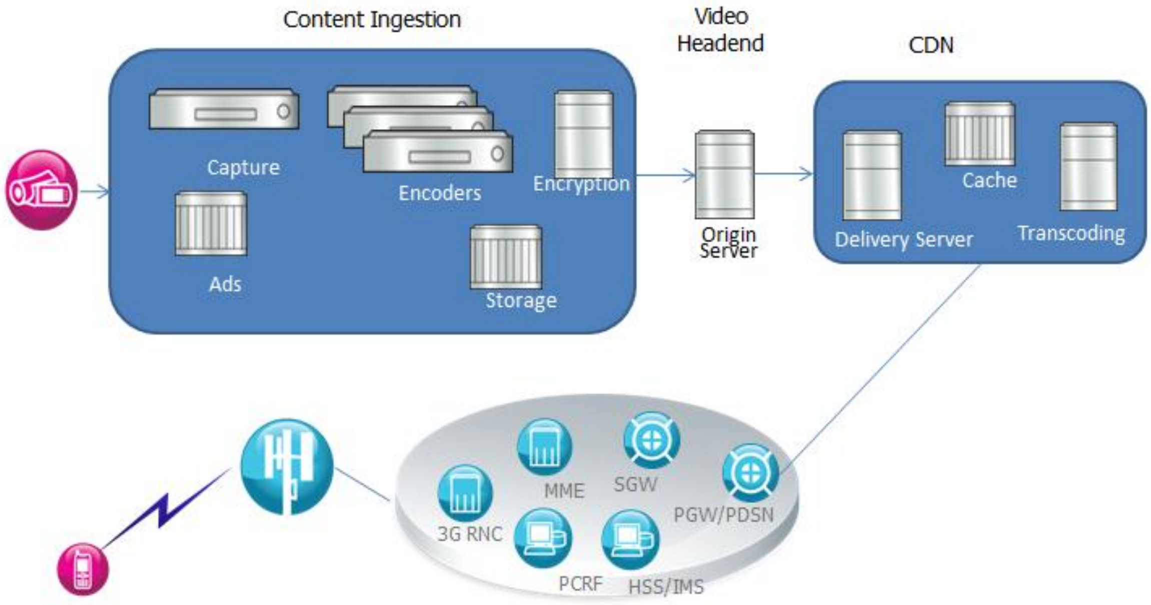


Figure 16: Content Delivery Network (CDN) Architecture

Request for content from the end device is redirected to the CDN's delivery server. The CDN server processes the request from end device, extracts device information from a subscription data base and picks appropriate content format to deliver the video, transcoding if necessary. The delivery server maintains the session with the end device streaming the video as the end device consumes the content. The video server may invoke APIs provided by the mobile core to establish QoS for the session.

5.3.1 MOBILE OPERATORS AND THEIR MIGRATION TO IMS

Mobile operators have the unique opportunity to consolidate multimedia services, and provide converged services by leveraging their IMS Core resources. Figure 17 shows the rollout of a 4G Americas operator's IPTV service.

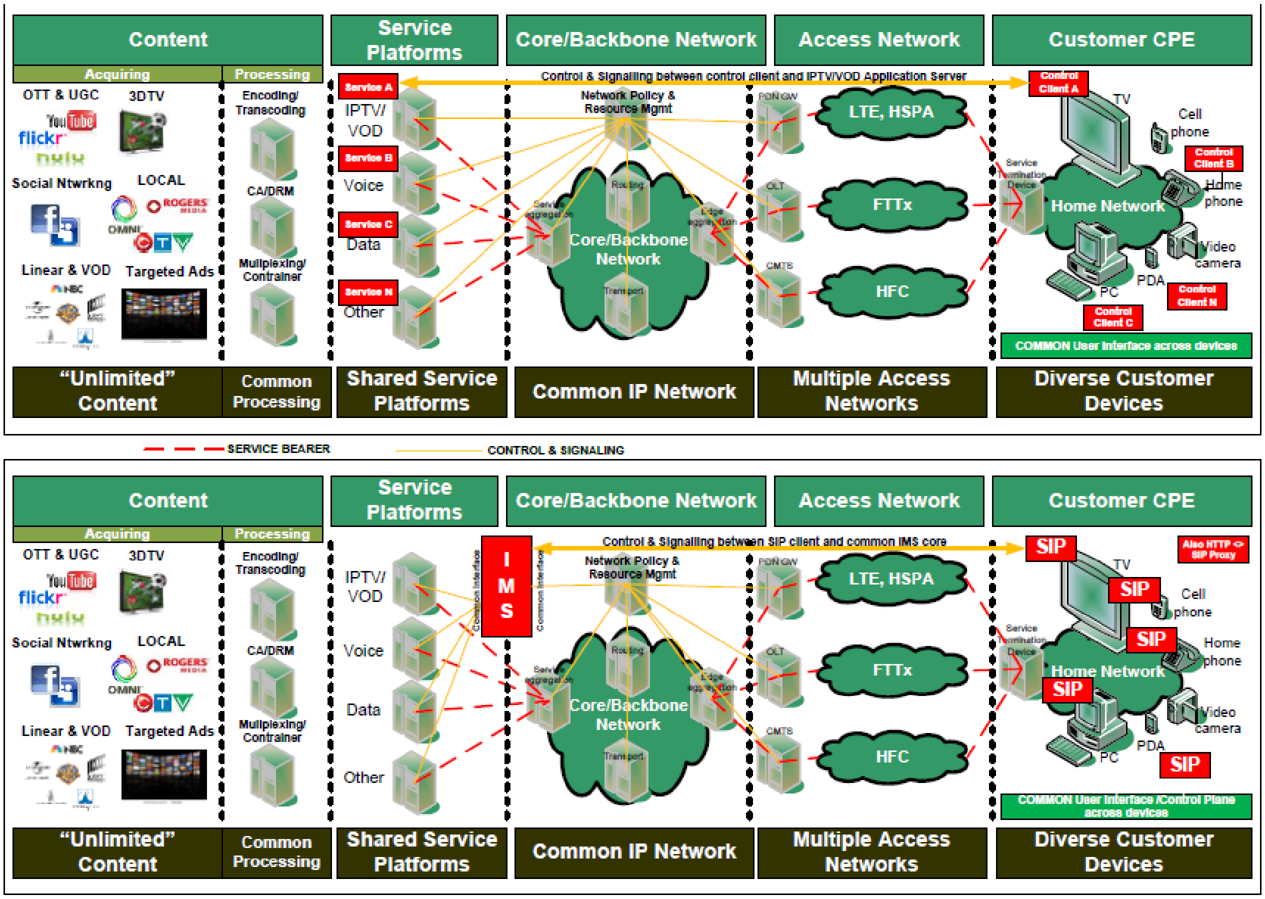


Figure 17: Rollout of an operator's IPTV service

The top part of this diagram represents a view where IPTV is rolled out as a separate service, with Mobile, Cable and FTTx access networks. The bottom half of Figure 17 represents the next generation of this strategy and direction, to converge all services to a common core and control plane, to maximize converged services delivery, enabling advanced video applications like teleconferencing, distance learning, video security, and easily moving video entertainment between platforms.

The mobile operator has a significant task in acquisition, processing and packaging of the videos sources they intend to deliver to their subscribers. Along with the negotiations, agreements, and contracts, they will need to have a Contract Rights and Management System (CRMS) to help them automate and provide business rules to their Content Management System (CMS). The CRMS provides rules for how the content can be used, where, when and how the content is delivered, exceptions, exclusions, and other details beyond the scope of this white paper. Some operators will do this function with a great deal of manual steps. However, they will need to progress to automation, or an outside service providing that automation, to keep up with both demand and new content available.

The CMS provides Metadata Management and ingests from multiple sources, like Internet Movie Data Base (IMDB), Rotten Tomatoes for details about the content itself. Rotten Tomatoes is a website devoted to reviews, information, and news of films. Electronic Program Guides from sources like Tribune, Rovi, validates, transforms, enriches, and normalizes the content and asset metadata, for use in Workflow Automation. The CMS, using workflow automation, business rules from the CRMS, operator rules, manages encoder and transcoder farms, manages scramblers and watermarking for Content Access/Digital Rights Management (CA/DRM), and manages Quality Assurance (QA) systems. The rules engines trigger workflows based on rules and triggered based on metadata.

Linear and VOD content themselves may come from many sources, such as satellite signal reception, off-air, direct feed fiber, Internet, and local file servers, and transport mechanisms such as Aspera and Signiant. Real-time processing of encoding (MPEG4 AVC, MPEG4 SD and HD Profiles, future HEVC), transcoding of MPEG2 to MPEG4, creation of Single Program Transport Streams (SPTS), including traditional splicing for Ad Insertion is also required. The CMS can be involved in the file based encode/transcode to multi-bit rate (MBR) profiles, containerization using Packagers to HLS, MPEG-DASH or other operator supported containers, as well as pre-encryption/watermarking of content.

The Content Access and Digital Rights Management System can protect the linear video services using Bulk Encryptors by Content Access client type. Offline (file based) VOD services can use scramblers by DRM client type. The encryption can also be applied in the packagers such that various containers could use different encryption systems depending on the client type being served.

While some mobile systems may use multicast delivery for linear feeds, most VOD content from the content processing is delivered to the Content Delivery Network (CDN) origin server, and then the CDN caching system, streaming systems and content routers, are used to deliver unicast streams. While the Mobile operator uses their wireless RAN, these same streams may be delivered over other access networks, for example on a DOCSIS Cable/HFC network of a Cable Operator through the CMTS, or via an FTTX network through the OLT.

The mobile operator will also need an IPTV “headend.” This Converged Video Application System (CVAS) will support the middleware and core TV Applications. The server-side middleware will include, at a minimum, an Electronic Program Guide and VOD title catalog, channel maps, defined packages and entitlements, and user management and authentication. This headend or CVAS will have applications to support Linear TV, VOD and nPVR. It will also provide remote authorization; enable advertising (system wide or targeted). We have only mentioned a few of the 100s of features of a modern IPTV headend. The headend or CVAS will also provide (directly or interfaces to) Service Provisioning Interface, Service Management Interface and Service Billing Interface; these are not “seen” by the users, but are the most important to the mobile video provider, to provision, manage, and most importantly, receive revenue.

The mobile operator could possibly have a future advantage in converged services using video as a key element in the feature. Service policy can be part of the Server Middleware/Application, as an application server on the IMS core, and by leveraging that operator's IMS core and connected network policy, the mobile operator can use SIP session management, identity management, profile management, using a common control plane to also deliver the video service, and converged applications with voice. The operator can also then leverage the network policy bandwidth and resource management, access control management, security and firewall management, as well as portions of the OSS systems. The common control plane at the clients (SIP or HTTP to SIP Proxy) helps to reduce the present barriers to multimedia application convergence.

As can be seen by this high level description, the mobile operator must be prepared to take on a complex video acquisition, content processing and packaging system, as well as comprehensive headend for server- side video applications and middleware, system management, provisioning, user management and interfaces to billing systems. Therefore, it is reasonably likely that many mobile operators will try to buy their video content as a managed service rather than take on wholly new areas of expertise and experience that normally only exist in Cable Multiple System Operators (MSO).

5.4 CONTENT MANAGEMENT IN CLOUD ENVIRONMENT

Content management delivery in a cloud environment leverages some of the edge caching techniques addressed in Section 5.1. Section 5.4 addresses some of the architectural principles in content management architecture in a cloud environment.

5.4.1 CONTENT MANAGEMENT ARCHITECTURE AND DELIVERY

From the perspective of the content consuming device, low latency and optimized user experience in video delivery is maximized when the server providing the content is close to the access network edge. This requires a good understanding of the market and traffic patterns related to video delivery, as described in this section.

a) Application of traffic profile and expectations to video delivery results in location of caching servers close to video consumption hotspots.
b) End user experience for video over wireless is maximized when the edge delivery architecture is aligned with the location of wireless hotspots to maximize economies of scale in resources to deliver video traffic over both content delivery networks (CDN) and radio access networks (RAN).
c) Hardware and software in the cloud is usually distributed for redundancy and scalability. Hardware could be dedicated to CDNs or leased based on traffic patterns to allow for more flexible architectures.
d) In networks where video content is just being introduced, efficiencies in the video delivery may not be fully exercised. There may be multiple copies of the same stream being sent to consumer devices. These are usually observed in emerging deployments of video. As traffic and demand for the same stream of video increases, e.g. multiple hits on a YouTube video, edge solutions are implemented that repeat the stream over the access to manage content so that video stream duplication between the cloud and the access network is minimized. This usually suffices for access networks that have not been overwhelmed with video delivery.
e) With evolution of access technology to LTE, it is now possible to use eMBMS techniques over mobile broadband by which repeated streams in a geographical area are broadcast to reduce use of duplicated, dedicated video bearers.

A combination of the last two approaches results in well-managed video content in a cloud environment, where the Content Management System (CMS) in the cloud accomplishes the following objectives:

- It identifies areas in access network where streams are being repeated, and optimizes by streaming a high bandwidth instance to that area.
- The access network at that location uses edge caching to minimize the number of repeated streams towards the cloud.
- The access network uses broadcast over RF to minimize use of radio resources dedicated to copies of video content to different devices simultaneously.

5.5 POLICY MANAGEMENT ARCHITECTURE

The 3GPP policy and charging control (PCC) provides a rich policy and charging environment to support dynamic control of QoS, charging and policy by application servers. Figure 14 shows a standard PCC architecture. The PCC functions include the following:

The PCRF (Policy and Charging Rules Function) provides policy control and flow based charging control decisions. The PCRF determines policy rules in a multimedia network in near real-time. It operates at the network core and accesses subscriber databases and other specialized functions, such as a charging system, in a centralized manner. The PCRF aggregates information to and from the network, OAM systems and other sources (such as portals) in real-time to create rules and then automatically making policy decisions for each subscriber active on the network. Therefore, it allows a network to offer multiple services and QoS possibilities. Because it operates in real-time, the PCRF has an increased strategic significance and broader potential role than traditional policy engines. PCRF can be integrated with different platforms like billing, rating, charging, and subscriber database or can also be deployed as a standalone entity.

The PCEF (policy and charging enforcement function) is implemented in the PDN gateway. It enforces gating and QoS for individual IP flows on the behalf of the PCRF. It also provides usage measurement to support charging.

In addition, an online charging system (OCS), not shown in Figure 14, provides credit management and grants credit to the PCEF based on time, traffic volume or chargeable events. An off-line charging system (OFCS) receives events from the PCEF and generates charging data records (CDRs) for the billing system.

The support for dynamic control of policy and charging will enable applications to make multiple requests per service session. Triggers associated with changes in location, network access, device status and security will generate additional requests. The key message is that the PCC capabilities brought together in 3GPP Rel-7 and enhanced with Rel-8 are sufficiently powerful to provide dynamic control of charging and QoS on a per flow and per subscriber basis. This architecture will allow support of new business models.

6. SURVEY OF OPTIMIZATION TECHNIQUES/MITIGATION STRATEGIES

6.1 OBJECTIVES OF WIRELESS VIDEO OPTIMIZATIONS

Video users require significant amount of resources to be made available throughout the network to ensure that the QoE of those users is being met. Even with the introduction of LTE and LTE-Advanced and the improved spectral efficiencies realized thereof, air interface resources will be at a premium due to the need to support a larger number of video users at higher data rates along with the other non-video users. Within the network, the backhaul and the core networks could also become bottlenecks.

Wireless video optimization techniques play an important role in ensuring that the QoE for both guaranteed-bit rate and non-guaranteed-bit rate video users are being met. The goal is to maintain the average throughput delivered to the clients over the time scale of several video chunks in the face of changing radio conditions and congestion conditions to the best extent possible, and only allow gradual changes to the average throughput. Different devices with different screen sizes, such as a tablet and smartphone, but in the same radio conditions, should be treated differently. A video user should always get better throughput than a data user when both have the same radio conditions and yet the extent of priority for video over data client should be configurable. The resource partitioning between data and video clients should adapt to the number of clients of each type to ensure that the data clients are not starved by the video flows. These optimization techniques must be realized with minimal impact to the UE battery life, must be dynamic enough to accommodate fast variations in RF and network loading, and must be devised so they operate within multivendor network environments. Furthermore, the impact of these optimizations must not adversely affect the QoE of other non-video users such as voice users. The following sections describe techniques for wireless video optimizations both in the radio and in the core networks.

6.2 END-TO-END QOS SUPPORT IN LTE FOR VIDEO DELIVERY

6.2.1 LTE QCI AND OTHER NETWORK QOS TECHNIQUES

QoS management in LTE is characterized by an advanced Policy and Charging Control (PCC) architecture, whose main change with respect to UMTS is the introduction of QoS Class Identifiers (QCI). QCIs are scalars used as reference for the definition of specific packet forwarding behavior, that link to a set of parameters pre-configured by the operator at the eNodeB such as scheduling weights, admission thresholds, link level protocol configuration, etc. QoS is controlled per service data flow, with each service data flow being a virtual connection that carries data plane information. Service data flows sharing the same QoS policy are in turn carried over bearers which represent suitable data paths between the user and the packet data network. Applications may trigger multiple service data flows for specific needs, each with their own QoS parameters.

In LTE Rel-8, a total of nine QCIs are standardized as shown in Table 4. Each QCI addresses a specific data flow and resembles the PDP context of GPRS and UMTS. A bearer is always associated with a QCI.

Table 4: Standardized QCI characteristics [from 3GPP TS 23.203]

QCI	Resource Type	Priority	Packet Delay Budget (NOTE 1)	Packet Error Loss Rate (NOTE 2)	Example Services
1	GBR	2	100 ms	10^{-2}	Conversational Voice
2		4	150 ms	10^{-3}	Conversational Video (Live Streaming)
3		3	50 ms	10^{-3}	Real Time Gaming
4		5	300 ms	10^{-6}	Non-Conversational Video (Buffered Streaming)
5	Non-GBR	1	100 ms	10^{-6}	IMS Signalling
6		6	300 ms	10^{-6}	Video (Buffered Streaming) TCP-based (e.g., www, e-mal, chat, ftp, p2p file sharing, progressive video, etc.)
7		7	100 ms	10^{-3}	Voice, Video (Live Streaming) Interactive Gaming
8		8	300 ms	10^{-6}	Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.)
9		9			

Each QCI belongs to one of two groups either GBR (guaranteed bit rate) or non-GBR (non-guaranteed bit rate). GBR bearers try to guarantee a minimum bit rate as required by the applications and usually force eNodeBs to reserve the necessary radio resources throughout the life of the connection, after admission control criteria have been met. GBR bearers are especially suitable for conversational and streaming multimedia applications. Non-GBR bearers do not impose any minimum bit rate and are therefore could be used for unpaid video traffic delivery.

QCI priorities enable differentiated treatment of packet flows when congestion appears at any point of the network, with a higher preference for lower QCIs (QCI 1 being the highest priority). Scheduling strategies at the eNodeBs should take these priorities into account in congestion situations where multiple data flows compete for a limited set of radio resources. QCI 9 with the lowest priority corresponds to the default bearer, which is always set up after initial power up thus allowing "always-on" capability. This bearer does not preclude simultaneous establishment of additional bearers with higher-priority QCIs, possibly for combined data services.

Mobile video delivery can greatly benefit from the traffic differentiation that can be achieved through use of different QCIs. Streaming video that may require a minimum data rate should be assigned to one of the GBR bearers corresponding to QCI 2 or QCI 4. Voice over IP on LTE (VoLTE) is supported through QCI 1. Non-premium best-effort data services can be well served by QCI 9, while QCI 8 could be reserved for premium-grade TCP-based services (like high quality progressive video), with a higher priority but not very tight delay restrictions. Speech services with QCI 2 would be given special attention (possibly through semi-persistent scheduling) and would not be under-served due to "hungry" video users if taken care of by appropriate scheduling strategies. Streaming video applications with bit rate constraints should be given any of the GBR bearers for tighter delay requirements. Premium-type video subscriptions could be favored through the use of QCIs 6 and/or 8 for TCP-based progressive video services.

In addition to the parameters shown in Table 4, each bearer is provided with additional parameters such as the Aggregate Maximum Bit Rate (AMBR), which limits the maximum user throughput at both the core and radio access networks. This enables additional service differentiation between users connected to the same Access Point Name (APN) and with the same QCI.

The network can also be protected from non-GBR data-intensive bearers that could collapse the network through use of AMBR. The PDN-GW performs rate policing based on AMBR per APN for both UL and DL traffic, while the eNodeB performs rate policing based on the AMBR per terminal. Maximum bit rate of GBR bearers is also limited at the PDN-GW. Optimization of these values can prevent episodes of network congestion due to greedy applications.

Prior to the introduction of QCIs, 3GPP defined a set of so-called traffic classes for UMTS [3GPP TS 23.107]. The four UMTS traffic classes are, in descending order of delay sensitivity: Conversational, Streaming, Interactive and Background. Conversational and Streaming classes are more applicable for GBR applications, such as video streaming, while Interactive and Background are reserved for best-effort TCP services with less delay requirements (including progressive download video). HSDPA provides an additional QoS control by defining Scheduling Priority Indicator (SPI), a number in the range [0, 1… 15] available at the NodeB for scheduling purposes [27]. A high SPI could be allocated to video streaming and other real time services, while other best-effort applications could be assigned a low SPI value. However, SPI values are only visible at the NodeB and cannot be given other proprietary “meanings” from the core network, therefore limiting its applicability.

6.2.2 LTE GUARANTEED BIT RATE BEARERS

3GPP has standardized a number of different QoS classes so that an application can invoke a bearer of an appropriate QoS class. In order to maintain the data rate at a constant level, video should be delivered over a Guaranteed Bit Rate (GBR) bearer. For a GBR bearer, the average data rate is guaranteed to be above a certain target GBR value by allocating the required amount of resources to the device to maintain the minimum target rate. For example, if the spectral efficiency drops to half, then twice the amount of bandwidth (physical resource blocks) or time is allocated to the user over the time scale of a second to maintain the minimum average rate over this time period. GBR bearers are assigned a higher priority compared to the default or best effort bearer. Resources are first allocated to GBR bearers up to a maximum allowed design limit and only then are remaining resources are allocated to best effort bearers. Thus, the video quality of a video stream carried over a GBR bearer will be superior to a best effort video bearer for the same radio conditions.

In addition to a target GBR value, 3GPP specifies a Maximum Bit Rate (MBR) not to be exceeded by a particular bearer. The scheduler at the base station will avoid reserving more resources than needed in order to not exceed the MBR value even in good radio conditions. LTE-Advanced (as defined from 3GPP Rel-10) enables setting an MBR value higher than the GBR, but this is not allowed in LTE Rel-8 and Rel-9 where MBR is constrained to be equal to GBR.

6.2.3 MAINTAINING CONSTANT VIDEO QUALITY WITH GBR SERVICE

As discussed in Section 4, frequent changes in video rates are detrimental to perceived video quality. Hence for video delivery, it is likely necessary to maintain the data rate even when the wireless channel fluctuates. Wireless channel variations result in instantaneous data rate changes because of the adaptive coding and modulation. With adaptive streaming, the rate determination algorithm in the client may try to follow the delivered data rates resulting in quality fluctuations. In order to maintain a constant average data rate to the client, it is not sufficient to maintain a minimum guaranteed bit rate as in the case of a GBR bearer because higher rates above the GBR target could be achieved when excess resources are available. Thus the GBR bearer will experience rate variations above the GBR target rate.

For RTP/UDP streaming, the maximum bit rate (MBR) can be set for the bearer with the target MBR value set equal to the guaranteed bit rate (GBR) value, which is presumably set to the RTP/UDP streaming rate. With GBR equal to MBR, a constant average data rate will be delivered to the RTP/UDP client by ensuring that the scheduler allocates an appropriate amount of resource to maintain the RTP/UDP streaming target rate and provide a good user experience. With this solution, since the scheduler always provides the same download rate to the RTP/UDP client, network utilization may not be optimized under varying network conditions.

For HAS streaming, the GBR can be set to a rate that provides the minimum download rate needed for an HAS client to provide the minimum acceptable quality play back experience to the user. A MBR can be set for the bearer, with the target MBR value set to a greater value to cap the highest quality that the HAS client provides to the user. The data rate delivered to the HAS client will then be within the specified GBR to MBR range, thus ensuring that the scheduler allocates an appropriate amount of resource to maintain at least the GBR rate and at most the MBR rate. With this solution, the GBR can generally be set lower than what would be needed for RTP/UDP streaming, the MBR can be set higher than what would be needed for RTP/UDP streaming, and the HAS client can choose from multiple available playback rates between GBR and MBR. This provides the scheduler flexibility to optimize network utilization, while at the same time allowing the HAS client to provide to the end user with the highest quality streaming solution possible for current conditions within the GBR and MBR constraints. A well-designed HAS client will do its own averaging and choose an appropriate playback rate at each point in time to provide the highest quality playback while avoiding stalls or skips.

6.2.4 DIFFERENTIATION AND FAIRNESS WITH GBR SERVICE

Video optimization using GBR and MBR also allows differentiating between different classes of users if desired. For example, if “Gold” users pay more to watch video than “Silver” users, the GBR target of Gold users could be set higher than that for the Silver users. This will ensure that in most cases Gold users get better quality than Silver users when adaptive streaming is used. Furthermore, setting the same GBR value across all users of a given class ensures that the same throughput is delivered. The GBR can be set according to the device characteristics and screen size (specified in a HSS/SPR) to achieve the same quality for different screen sizes. For example, for the same class of users, a tablet device could have GBR equal to MBR set at 400 kbps while for a smartphone, the GBR target could be 200 kbps. Because of the smaller screen size, the same quality can be achieved with the lower throughput for the smartphone compared to the tablet.

6.2.5 LIMITATIONS OF GBR BEARER

While in principle GBR and MBR offer a means to control video quality, there are number of caveats to consider. At the time the flow is admitted, a user may have reasonable radio conditions allowing for a good GBR target. However if the radio conditions worsen, maintaining the GBR can consume an excessive amount of resources leaving the remaining data flows starved. Conversely, setting the GBR target too low can make it irrelevant because even the data flows will achieve that throughput. Thus it is unclear how to pick an appropriate value for GBR that will result in a reasonable amount of resources being devoted to the GBR flow. Network operators could shy away from using GBR for this reason.

To avoid the problem of excessive consumption of resources, a cap on the consumption of resources by the GBR flows is usually imposed. This makes the GBR targets not achievable under congestion. However, the side effect of that is that these GBR flows will get throughputs according to the resource allocation policies of the underlying scheduler. This can result in highly unequal throughputs allocated to the various GBR flows with some flows meeting their targets while some other flows falling substantially below their targets. In other words, simply setting the GBR targets does not necessarily maximize the aggregate quality of the video flows when there is congestion in the network. A more desirable solution would be to have more equitable allocation that allows more video flows to reach a slightly lower target thereby accommodating more video flows.

6.2.6 MICROBURSTING

Microbursting is a technique where servers shoot out data as fast as possible (i.e. at a high data rate) within a short period of time to each end user and thus fill up the client buffers. However, challenges with microbursting remain. The problem with microburst is to allocate the time to burst and permitted size. With many independent flows in the same RAN, sharing the bandwidth in a fair manner is non-trivial - especially when the sources are from servers outside the operator’s network. Protocols such as TCP, which try to achieve fairness between multiple flows, could also going to be compromised by bursting.

6.3 RF SPECIFIC OPTIMIZATIONS FOR VIDEO DELIVERY

Video optimization at the air interface, typically performed at the base station or eNodeB and the video client on the device, is designed to address a number of requirements such as:

- maintain video quality in the face of time varying radio conditions;
- manage video quality as the cell becomes increasingly congested;
- differentiate video quality according to the subscription level;
- manage video quality during handoffs;
- ensure fairness when multiple users are served.

We will first discuss optimization at the base station, and then address client optimization. Techniques described in this section may have been described earlier and are summarized here since they constitute RF specific optimizations for video delivery.

Optimizing video flows at the base station has the inherent limitation that the base station does not have information about the video characteristics such as the type of video or the kind of scene that is being carried in a particular set of packets, the manifest file of the video stream, or the screen size of the device. The manifest file holds all the information about the video file, its duration, the streams, the codec info, and a list of all the available chunks. The manifest file may be encrypted and, thus even with DPI, cannot be available to the base station. Any information related to the video has to be signaled to the base station by upstream nodes. If proprietary signaling is not to be used to maintain interoperability, then the only information that can be provided to the base station are the radio bearer characteristics defined in the standards associated with the bearer that is designated for carrying the video flow for a specific device. Several schemes in the literature, that optimize the stream based on providing higher rate or reliability for the I frames compared to the P and B frames or the ones that re-encode the video into base layer and enhancement layer according to channel rates, are not pragmatic.

On the other hand, radio-related information is readily available at the base station. The base station knows for example the number of users being served, the number of packets waiting to be delivered for the different flows, and the signal-to-interference ratio on the radio link to the different devices. The main method of video optimization at the base station is by controlling the data rate delivered to the device.

Client side optimization has information about the video, such as the manifest file in the case of adaptive streaming. In the case of HAS, the rate determination algorithm runs on the device and it can be optimized specifically for the wireless environment. A well-designed HAS client should be able to work well and share fairly with other HTTP flows without knowing about the other devices that share the network. The network fairly provides as much data downloading capacity as it can to each end device. The HAS client on the device then decides which bit rates to download at each point in time to provide the best quality experience to the end user (i.e., no stalls or skips and a playback quality that pretty much matches the current download rate of the HAS client).

6.3.1 GUARANTEEING MINIMUM VIDEO QUALITY ON THE AIR INTERFACE

Wireless channel quality is inherently time-varying because signal propagation is affected by the movement of the transmitter, receiver or surrounding objects. Adaptive coding and modulation is a standard feature in cellular wireless networks to deal with channel time-variations. With adaptive coding and modulation, the data rate transmitted to the device is adjusted according to the channel quality. As a result of these data rate variations, it is possible that the data rate dips too low to maintain the video quality. This could result in the video stalling, or the video quality becoming unacceptably low for the case of adaptive streaming. Well-designed HTTP Adaptive Streaming clients would be able to handle this issue. In addition, video optimization at the base station that targets maintaining a minimum data rate to the video clients in the face of channel variations can substantially improve video quality by preventing stalling. Since video quality depends on only average data rate over time scales of a second it is sufficient to maintain the average data rate delivered to the device at a constant level. This video optimization is accomplished through setting of Guaranteed Bit Rate bearers for video streams.

6.3.2 BROADCAST/MULTICAST OF VIDEO

When the video is of interest to a number of subscribers at the same time then multicasting the video over the air using eMBMS in LTE (or MBMS in 3G) is substantially more efficient than sending separate unicast transmissions to each client. This is because wireless is fundamentally a broadcast medium. Whatever is transmitted wirelessly can be received by multiple receivers in the range of the transmitter. Additionally, the concept of Single Frequency Network (SFN), where the same signal is transmitted by multiple adjacent base stations, improves performance by eliminating out-of-cell interference. SFN can be employed when a number of users in multiple cells are watching the same content at the same time, as can broadcast of live sports events or news. eMBMS is described in detail in Section 5.2.

Traditionally, a number of users watch the same video in Linear TV. However, Linear TV is not commonly used on mobile devices. The real application for eMBMS delivery of video to smartphones and tablets is in stadia where specialized content such as action replays is watched by a large number of people at the same time. With a number of small cells covering the stadium, eMBMS with SFN could be substantially better than unicasting to different users.

6.4 NON-RF SPECIFIC OPTIMIZATIONS FOR VIDEO DELIVERY

There are several techniques above the RF layer, besides LTE QCI, which can be used to optimize the wireless video viewing experience. These range from adapting the video quality to the available network bandwidth, minimizing the impact of video on the network through to encouraging change of user behavior to use the network when it is less heavily used. This section provides a survey of the main techniques of non-RF specific optimizations for video delivery.

6.4.1 DASH (HAS)

All HTTP Adaptive Streaming (HAS) methods, including Dynamic Adaptive Streaming over HTTP (DASH), allow the video quality to dynamically adapt to the available bandwidth, and is a non-RF specific method of optimization. The client senses the available bandwidth and makes the decision on what quality level to pull for the next chunk (e.g. 2 seconds of video), as has been described in Section 2.5.

There are several attractions to using HAS. Firstly, it is based on TCP. This means that lost packets are automatically retransmitted, and the TCP congestion avoidance ensures a fair share of the available bandwidth. A further attraction of HAS to the content providers is that it does not require any support from the network. This means the same content can be delivered over any network. The only requirement from the network is that it can sustain the bandwidth for at least the lowest video quality. However, this means the end user will experience variable QoE depending on congestion and their radio conditions. A final attraction is that HAS is based on HTTP. This allows reuse of web technology reducing the cost of deployments. It also enables support for HAS content through web caches.

6.4.2 BIS SCALABLE VIDEO CODECS (SVC)

Several codecs support the ideal goal of scalability. Scalable Video Coding (SVC) [28] allows representing video into different versions with different frame rates, spatial resolutions and/or quality levels (fidelity to the original video) by a single video stream. The content is encoded using a base layer and one or more enhancement layers. The base layer provides a complete encoding of the content at the lowest acceptable quality. The client only needs to receive the base layer encoding. However, it can improve the quality if it also receives one or more of the enhancement layers. These can improve the spatial (higher image quality) and/or temporal (frame rate) of the decoded image.

An important property as defined in MPEG Scalable Video Encoding (SVC) is that it is easy to separate out the base layer and the individual enhancement layers – even where the video is encrypted. This reduces storage overhead of SVC as a single file can contain the base layer and all the enhancement layers. The overhead for the delivery server to deliver on-demand the base layer and any of the required enhancement layers is low. Unfortunately, there is an overhead using SVC. For MPEG-4 SVC this is over 10 percent per enhancement layer. As 5+ enhancement layers are required for a smooth transition from low quality to high quality, this overhead can become unacceptable, especially when this is applied to the most congested part of the network namely, the wireless RAN.

When SVC is used in DASH, an additional level of optimization is realized. The different SVC layers can be split and distributed into different "representations." If users have enough throughput they can request all the representations with all the layers, but when their throughput is limited, they can omit request to data from the representations containing the highest layers [29].

6.4.3 TRANSCODING/TRANSRATING/RATE CAPPING

While techniques like HAS and SVC handle the dynamic variation of the network bandwidth efficiently, it doesn't work with majority of the video content watched on the mobile network today. This is because majority of today's video content is delivered over HTTP-PD (Progressive Download). So there is a need for special techniques like dynamic transrating and dynamic transcoding on the mobile network to address this content. Otherwise operators will not be able to efficiently address the video data tsunami.

Transcoding involves dynamically changing the encoding of the video stream to adapt to the available bandwidth. The video coding technology is modified based on the knowledge of the device. A device profile is maintained and, based on the requesting device and its software configuration, an appropriate technology is selected and the video is transcoded from one format to another. The source video may be prepared for a screen size that is large with a fast refresh rate, and thus coded at a bit rate that is unnecessarily high for a wireless device with a small screen and lower refresh rate. Transcoding may be required to change the resolution and frame rate of the video to make it suitable for the smaller screen. Other factors that can be included in selecting the transcoded video frame rate and resolution are subscription tier of the subscriber requesting the video, and other policy based compression techniques.

There are several approaches that support changing the resolution and fidelity to reducing the frame rate without needing a full decode/encode. This is known as transrating. Bandwidth estimation is performed to determine which video flows need transrating. Video streaming can thus adapt to congestion.

Transcoding is heavy on processing requirements to perform decode and re-encode in real-time. Also, it requires the content to be unencrypted. It is less attractive than HAS since many professional content providers are reluctant for uncontrolled changes of the encoding and hence end user QoE of their content. Transrating is usually less demanding of compute power, and yet is able to give a good range of qualities. Hence it is an attractive approach where content providers permit.

High value (i.e. paid for) content is likely to be encrypted, thus making transcoding difficult. This leaves lower value content including user generated video sites as candidates for transcoding. The operator has to decide on balancing the cost of transcoding against the improved QoE for such content. In addition, they must consider how the content providers will react when more of their content is changed in quality in the network. They may respond to providing the content in multiple bit rates to help mitigate the need for network transcoding.

Rate capping is the process by which an intermediate node responsible for video optimization measures the available data rate and adjusts the sending rate of a progressively downloaded video (without altering the video content) such that the downloaded portion of the video does not run too much in advance of what is played, thus avoiding a lot of wasted bandwidth when the user stops watching. The consequence of this technique is that a large play-out buffer is required. During times that the video rate is higher than the shaped rate, the video player needs to be able to rely on the video stored in this buffer.

6.4.4 CONGESTION-AWARE OPTIMIZATION

The impact of increased video traffic is felt the most on the RAN. It is in the RAN where many video usability problems start surfacing as the cells get congested. As more users crowd a cell, the effective network bandwidth experienced by each user decreases. As a result, video playback, particularly HTTP-Progressive Download (HTTP-PD), starts experiencing buffering issues. Adaptive Streaming protocols handle this problem by requesting a lower quality video, which by structure would need lower bandwidth for satisfactory playback. But with HTTP-PD the entire video is delivered as a single content segment. So if the rate of video delivery is lower than the playback rate, then the player would have to pause the video and buffer enough frames before resuming the playback. It is this scenario that the “congestion triggered optimization” technique hopes to address.

In simple terms, congestion triggered optimization is a mechanism where a video optimization engine dynamically starts optimizing (transrating/transcoding) video traffic during periods of congestion. Once a video flow is selected for dynamic optimization, the optimization engine would continuously modulate the compression rate (up/down) to ensure a smooth video playback. Congestion trigger can be achieved in two ways.

1) **External RAN probe based trigger.** With the external trigger approach, mobile operators typically deploy a RAN probe on the Gn interface (between SGSN and GGSN) and have it monitor the RAN usage messages flowing out of the RAN. When this probe detects a particular cell/site to be congested, it would notify a network element about the same. Typically this notification is propagated to interested elements via the PCRF. The PCRF would then issue a dynamic update to the video optimization engine. Then, the video optimization engine would identify all video flows happening on the congested cell and apply dynamic optimization to them.
   It is important to note that in this model the entire cell is being labeled as “congested.” Therefore the video optimization engine applies dynamic optimization to all flows pertaining to that cell.

2) **Flow based trigger.** In this model, the video optimization engine is deployed inline to the data traffic. The optimization engine would then continuously monitor the average bit rate of each video and compare it to the effective bandwidth experienced by the device for each flow. If the optimization engine recognizes a disparity, where effective network bandwidth is lower than video’s average bit rate, then the flow is immediately selected for dynamic optimization. Once a flow is selected for optimization, the optimization engine would continuously monitor the effective network bandwidth and tune the compression level appropriately to ensure smooth video playback. This model provides a comprehensive solution to tackle the Quality of Experience issue with the delivery of video traffic on mobile networks. A user who is experiencing poor network quality due to his distance from the tower would receive dynamic video optimization even though he or she is the only user on the cell. Similarly, a user who is a connected to a congested cell, as per RAN metric, but is not experiencing bandwidth disparity issues, would not have his video compressed. The user receives original video quality and does so without consuming any additional computing resources (transcode/transrate) in the network.

6.4.5 CLIENT OPTIMIZATIONS FOR HAS

The client plays a significant role in determining how the video gets delivered. In the case of adaptive streaming, the video rates and thus the quality of experience is also controlled by the client. The client can thus play a role in optimizing the video delivery specifically taking into account the nature of cellular wireless communication.

HAS clients can smooth the effect of wireless channel rate variations by increasing the averaging window over which bandwidth estimation is done for video rate determination. By introducing longer averaging time, fluctuations because of channel fading could be eliminated. Rate selection could be made more conservative compared to wireline in the interest of maintaining a more stable rate. The criteria for switching to the next higher rate could be tweaked to ensure that estimated bandwidth will really be available over longer periods of time and a fall back to lower quality will not be required.

In the case of progressive video, the client has the flexibility to either download the whole file in one shot and buffer it or periodically download small chunks. The advantage of the latter is that when the user does not watch the entire video, then the network has not been used unnecessarily. However, if the downloading of the individual chunks is spaced more than 10 seconds apart (or the number of seconds corresponding to the eNodeB timer) then the radio bearer will be released after each download and needs to be re-established for the next chunk. This causes unnecessary overheads which utilizes additional resources both in the network as well as the device. For the device, the battery drains faster when the download is done in multiple small chunks.

The client can in principle be also tweaked to control the download rate to lower value that what the network can provide so as to maintain a steady download over a longer time. This can be done for example, by using TCP receive window control mechanism. Once the receive window is full, then sender will not send any more data. The client can thus pace the download of packets.

In general, having a larger playout buffer can minimize video stalling during periods when the network data rate diminishes. Thus, having a larger playout buffer improves quality but can result in additional start up delay if the client waits to fill up the buffer. However, the startup delay can be mitigated for adaptive streaming by starting with low video rates that quickly fill the buffer and then gradually improving the quality.

When the device has access to the network through multiple radio technologies, then the client could dynamically select the access technology through which video chunks are downloaded. For example, the client could request a low video rate chunk through the cellular interface and once that is delivered and there is still time before playout it could request a higher rate version through a Wi-Fi interface.

6.4.6 CONTENT CACHING IN THE NETWORK

The RAN is not the only place where the video traffic has an impact. The core network and EPC has to be large enough to handle the demands for the large and growing number of video streams. One way to reduce the impact is to cache (i.e. store) the popular content close to the end user. A request for content which is cached is diverted to a suitable cache server and delivered from there. This saves core network bandwidth. It also reduces traffic on peering and transit links from the wireless network operator's network to the Internet in cases where the content is external.

Caching can be transparent or explicit. In the former case, all content is cached irrespective of source. It will remain in the cache as long as there are requests for it. With explicit caching, the wireless network operator agrees to cache content from a content provider. In this case, because the content will probably delivered with higher quality (especially if some of the RF enhancements mentioned in the previous section have been applied), the content provider probably pays for the storage and delivery of their content.

It is important to note that while content caching may address the backhaul congestion, it does not solve the RAN congestion issue. Furthermore, it requires a co-operating network core. If video content is cached at the edge and delivered from the edge, then mechanisms to modify existing billing mechanisms must be considered. In current architectures, billing is typically done by the PGW, which sits in the mobile core. Content caching in the PGW does not require a change to the billing mechanisms.

6.4.7 TIME SHIFTING OR SIDELOADING SUPPORTED BY CONTENT CACHING IN THE UE

Traffic in a specific cell sector typically varies over the day. Typically, cells in business districts will be heavily used during the day whereas cells in residential districts will see their usage peak during the evening. In most cases, the cells will be largely only very lightly used during the night. Hence this makes in an ideal time to deliver video files. Either based on the user's explicit selection or by predicting their likely interests, video files can be downloaded when the cell is little used. Time shifting or sideloading refers to downloading video content to the device during off-peak times or when the device is connected to a Wi-Fi access point and caching the content in the device for viewing later. While time shifting is not on-demand instant streaming, with a good recommendation engine, the end user experience can be almost as good. Indeed because the file is downloaded and not streamed, the quality is not affected by variable quality radio conditions. eMBMS can be utilized to efficiently deliver popular content which would be cached by a large number of users.

Time shifting technique is particularly suitable for non-streaming content such as YouTube videos that are based on progressive download. Pre-fetching content allows high quality video to be downloaded. Furthermore, the cost of downloading video will be substantially less when done over Wi-Fi or during off-peak times. The sideloading solution requires a pre-fetch client on the device that will talk to a sideloading server in the network that can direct the client to obtain specific pieces of content at the most appropriate time. At the direction of the sideload server, the client can then directly download the content from the appropriate video server. The sideloading server can utilize information from network probes and other information from the client such as proximity to Wi-Fi access point or battery life remaining to determine the opportune time and access technology to be used for pre-fetching.

With increasing amounts of memory available on the device, caching popular content on the device is becoming increasingly feasible. UEs such as smartphones and tablets now have substantial storage capacity. This allows content to be delivered to the UE when the network is lightly loaded and cached. When the content is requested, the request is directed to the local device avoiding traffic on the RAN. This has the further advantage of improved quality. To minimize the amount of content that needs to be cached, the users can opt in for sideloading and indicate to the sideloading server what class of content is of interest. For example, a user may indicate that new music videos of a particular artist is of interest and whenever there is a new video the sideloading server could include that in the pre-fetching list for that user. A personal recommender system that studies which videos the user watches and determines recommendations for the user can also be used by the sideloading server.

6.4.8 VIDEO USAGE BEHAVIOR

Yield management is the umbrella term for a set of strategies that enable capacity-constrained service industries to realize optimum efficiency from operations. The core concept of yield management is to provide the right service to the right customer at the right time for the right price. That concept involves careful understanding of service, customer, time, and price. The service can be defined according to the dimensions of the service, how and when it is delivered, and how, when, and whether it is reserved. The principles of Yield management have been applied to the airline industry and could possibly be applied to wireless networks particularly for video applications since they are bandwidth intensive. This concept could allow for effective use of operator spectrum resources, while best attempting to maximize user QoE.

7. RECOMMENDATIONS

7.1 RECOMMENDATIONS FOR VIDEO APPLICATION DEVELOPERS FOR EFFICIENT RADIO RESOURCE MANAGEMENT

Video application developers need to consider the challenges in operating over a network with mobile devices. This requires that applications are adaptable to changing access capabilities. This leads to the development of APIs from the CDN to the access network. Some guidelines for video application developers include:

1) Utilize APIs between the Application Layer and Access.

    - Mobile operators could provide the development space and environment for realistic simulations.

    - Drive the creation of APIs to demand QoS from the network.

    - App developers can differentiate themselves with access aware apps.

2) Understand benefits of network optimized applications and products such as:

    - A network aware application could operate better under busy network conditions than competing applications. Consequently, there could be an increase in the demand for network aware applications.

    - Develop applications for the maximum efficiency of the radio resources.

        - Utilize less data network resources than competing applications.

    - Make video applications more attractive to the end users.

3) Ensure that video servers deliver chunks of video in small regular intervals, not longer than approximately 10 seconds. Applications running over wireless networks need to be aware of the properties of the RAN. Both the video delivery algorithm and the network configuration must be optimized. Apple HLS typically uses 10 second HTTP chunks. With longer chunks, as each chunk is downloaded, there is a pause until the next one. Today's wireless networks implement inactivity timers about 5 seconds, and in many cases they have been optimized for more bursty types of traffic (especially when Rel-8 Fast Dormancy is not available). However, even if shorter chunks are used (which can be preferable for better rate switching and less end-to-end latency for live streaming), the client can still download several chunks consecutively before pausing to download the next set of chunks. The most optimal value of chunk size depends on, amongst other things, the eNodeB timer value settings. Video sources with significant inactivity periods between chunks of data can force the network to move the device to a dormant state, therefore wasting signaling resources and device battery. This will become especially harmful in high-speed wireless networks (like LTE), where the high bit rates available would allow for significant inactivity periods between video chunks.

4) Ensure that polling requests from the devices to the network are bundled to maximize use of seized resources. This will require device video application developers to synchronize their network information requests with other applications on the device. This requires that device

APIs to network access are leveraged. This also addresses the reality that video applications will likely incorporate user-specific advertising, special offers, subscription updates, and other types of notifications to the end user.

7.2 RECOMMENDATIONS ON BIT RATES FOR VARIOUS CODECS AND SCREEN SIZES/DEVICE TYPES

The choice of bit rate for the audio and video streams is a compromise between quality of experience and bandwidth used for streaming. The device screen size and viewing distance are a critical aspect in determining the bandwidth needed for encoding the video. The larger the screen and the closer it is viewed (up to the near point of the eye), the higher the bandwidth required. A video that looks poor on the larger screen of a tablet can be more than adequate, in terms of QoE, on a smartphone held at the same distance. Scenes with high motion or scenes with a lot of detail will require more bandwidth to minimize visual artefacts. Chaotic motion (grass blown in the wind, flames, waves, etc.) in scenes with a lot of detail are the hardest to encode and hence require higher bandwidth to avoid loss of definition. Finally, the compression efficiency of the video codec impacts the bandwidth needed to encode a video quality with a given fidelity.

At the time of writing, the most widely used video codec for smartphones and tablets is H.264 / AVC. Table 5 shows the recommended bit rates for different devices. 4G Americas recommends a minimum of 200 kbps for smartphone and 400 kbps for tablets. Note that these are average bit rates and variable bit rate (VBR) encoding can also be used. This allows a higher encoding rate to be used for short sections (a few frames) of high complexity and/or motion resulting in a higher overall quality. People often watch movies and TV programs on their smartphones and tablets using high quality headphones. These require a good quality stereo audio source. However, recognizing that radio spectrum is limited, 4G Americas recommends using AAC (Advanced Audio Codec) at 96 kbps.

Note that HD rates are included because of the option of delivery of the video stream over a mobile network and forwarding to TV using in home using DLNA or AirPlay over Wi-Fi. For example, with rural broadband it may be more efficient to use wireless technology to serve a dispersed community.

Table 5: Recommended choice of bit rates for video streams

Video Quality Level	Video (kbps)	Total (kbps)	Comments
VQ1	200	296	Smartphone
VQ2	400	496	Tablet / Smartphone
VQ3	800	896	Tablet / Smartphone
VQ4	1400	1496	SD 480P TV screen through games console and STBs, Connected TVs, PCs and Tablets
VQ5	3150	3246	HD 72P TV screen through games console and STBs, Connected TVs, PCs and tablets
VQ6	7100	7196	Full HD 1080P, TV screen through games console and STB, Connected TVs, PCs with General Processor Units.

In many cases, the content provider will encode their content in multiple qualities. This allows the end user to manually select the desired quality taking into account their device, desired quality, and data plan. They may also select a lower video quality level because they will achieve more consistent video quality irrespective of the radio conditions. HAS dynamically selects the video quality based on the available network bandwidth. H.265 HEVC codec has just been ratified. Early tests indicate that it gives a 40-50 percent reduction in bit rate for the same image quality.

8. SUMMARY AND CONCLUSIONS

Video for entertainment and communication is experiencing significant growth, given the current plethora of mobile handheld devices enjoying enriched multimedia capabilities and as competitive forces are driving lower data rates on mobile carriers. The introduction of new technologies like LTE, with low latency and high throughput, will enable the evolution of video content from SD to HD and handheld form factors to larger screens, thus further increasing the data volumes in networks. People's expectations of quality and availability will also grow as the use of video on mobile devices becomes main stream. The increase in streaming content over mobile networks will provide significant challenges to the carriers to keep up with demand and to provide a Quality of Experience for these services that provides a competitive advantage over their competitors. Mobile video presents challenges unique to and different from wireline video. The growth of video will pose considerable stress on HSPA+ and LTE networks to handle both the volume of traffic and end user expectations. While some of these problems have already been addressed in wireline networks, wireless networks impose additional technical challenges due to varying interference and cell handover issues.

This white paper identifies technologies, techniques and architectures to mitigate the problems and recommends approaches to maximize the Quality of Experience in HSPA+ and LTE networks. Based on the data consumed by different streaming applications, we note that without more licensed mobile wireless spectrum allocations from governments throughout the Americas region, video streaming over mobile broadband wireless technologies such as LTE will continue to be challenged as the growth of mobile video demand increases at exponential rates. Technical solutions that can be used in tandem to address the challenges of mobile video delivery are discussed, including CDN design, Adaptive Streaming, cellular HET-NETs, Wi-Fi offload, eMBMS, etc. Optimization of mobile networks and QoE for streaming content is critical and will need to be the focus of the industry for years to come. In order to fulfill the mobile users' expectations, a set of network as well as client optimizations are being devised (both RF and non-RF based), in addition to the introduction of new and more advanced codecs.

Delivery of video chunks, through HTTP adaptive streaming, at various bit rates based on congestion is identified as the preferred option for mobile video delivery. Interactions between HTTP over Adaptive Streaming and TCP mechanisms are discussed in depth. Appropriate monitoring of the Quality of Experience in live networks can be conducted in order to optimize performance in the cases where it is feasible. Operators are currently striving for setting up QoE monitoring tools that could provide a deeper knowledge on the user’s degree of satisfaction. Recommendations for video applications developers can persuade them from wasting network resources or overloading the network with excessive signaling. A set of best practices will result in more efficient ways to utilize the scarce spectrum available.

This whitepaper provides different focus areas for different audience readers as summarized below.

For Operators

- Architecture choices and improvements in both access and content providing networks are described and analyzed.
- To work smartly with available and forecasted mobile broadband capacity in wireless networks, various ways to best utilize the wireless capacity are identified.

For Vendors

- Areas for cooperation and integration with other dependencies in the overall content generation and delivery network are identified.
- Areas for improvement and optimization in devices and networks are identified.

For Application developers and architects

- The effect of suboptimal design of video applications and their effect on mobile networks is addressed.
- Recommendations are provided on areas of improvement.

For Industry analysts

- The state of video on mobile broadband is described and trends are identified.

For Device manufacturers

- Evolution in video processing technology and codecs is addressed.

It is expected that through the support of various network architectures and optimizations, QoE monitoring metrics and techniques, mobile video will grow by orders of magnitude in the next several years.

LIST OF ACRONYMS

AAC	Advanced Audio Coding
ARQ	Automatic Repeat Request
CA	Content Access
CABAC	Context-Adaptive Binary Arithmetic Coding
CDN	Content Delivery Network
CMS	Content Management System
CRMS	Contract Rights and Management System
CVAS	Converged Video Application System
DASH	Dynamic Adaptive Streaming over HTTP
DRM	Digital Right Management
eMBMS	Evolved Multimedia Broadcast Service
FEC	Forward Error Correction
FLV	Flash Video
FTTX	Fiber to the last mile
GBR	Guaranteed Bit Rate
HARQ	Hybrid ARQ
HAS	HTTP Adaptive Streaming
HEVC	High-Efficiency Video Coding
HLS	HTTP Live Streaming
HSS	Home Subscriber Server
HTTP-PD	HTTP Progressive Download
IMDB	Internet Movie Database
IMS	IP Multimedia Systems
MPEG	Moving Picture Experts Group
MSO	Multi System Operators
NGMN	Next Generation Mobile Networks Alliance;
OTT	Over The Top
PCRF	Policy Control Charging Function
PPV	Pay Per View
PSNR	Peak-Signal-to-Noise Ratio
QCI	QoS Class Identifier
QoE	Quality of Experience
QT	QuickTime
RDP	Real Data Protocol
RNC	Radio Network Controller
RRM	Radio Resource Management
RTMP	Real Time Messaging Protocol
RTP	Real Time Protocol
RTSP	Real Time Streaming Protocol
RTT	Round Trip Time
SFN	Single Frequency Network
SIP	Session Initiation Protocol
SPR	Subscriber Profile Repository
SSIM	Structural SIMilarity Index
SVC	Scalable Video Coding
VOD	Video On Demand
VQEG	Video Quality Experts Group
WebRTC	Web Real Time Communication

LIST OF REFERENCES

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012–2017

http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html

[2] Ericsson Consumer Lab Report,
http://www.ericsson.com/res/docs/2012/consumerlab/consumerlab-tv-video-changing-the-game.pdf

[3] Ericsson AB, TV & Video, Changing the Game, 2012
http://www.ericsson.com/res/docs/2012/traffic_and_market_report_june_2012.pdf

[4] Ericsson Mobility Report, November 2012
http://www.ericsson.com/res/docs/2012/ericsson-mobility-report-november-2012.pdf

[5] Sandvine Global Internet Phenomena Report, 2H 2012

http://www.sandvine.com/downloads/documents/Phenomena_2H_2012/Sandvine_Global_Internet_Phenomena_Report_2H_2012.pdf

[6] NFL Super Bowl SLVII Live Stream Sets Viewership Records, February 5, 2012

http://www.nfl.com/superbowl/story/0ap1000000136438/article/super-bowl-xlvii-live-stream-sets-viewership-records

[7] BBC Player, Monthly Performance Pack, November 2012
http://downloads.bbc.co.uk/mediacentre/iplayer/iplayer-performance-nov12.pdf

[8] Ooyala, Global Video Index, Q3 2012
http://go.ooyala.com/rs/OOYALA/images/Ooyala-Global-Video-Index-Q3-2012.pdf

[9] ITU-T G.114 Telecommunication Standardization Sector of ITU (05/2003)

[10] 4G Americas White Paper, Mobile Broadband Explosion, August 2012
http://www.4gamericas.org/documents/4G%20Americas%20Mobile%20Broadband%20Explosion%20August%2020121.pdf

[11] Balakrishnan, Hari, Padmanabhan, Venkata N., Seshan Srinivasan and Katz, Randy, H. “A Comparison of Mechanisms for Improving TCP Performance over Wireless Links” August 1996

[12] Kuhlins, Christian, Ericsson & Thomas, Remi, FT Orange *LSTI, Boosting LTE for Mobile Broadband Deployment* http://www.slideshare.net/guest99ced7/lsti-mwc-presentationfinal, page 11.

[13] 4G Americas White Paper, Developing and Integrating a High Performance HET-NET, October 2012

http://www.4gamericas.org/documents/4G%20Americas%20-Developing%20Integrating%20High%20Performance%20HET-NET%20October%202012.pdf.

[14] 3G Americas White Paper, MIMO Transmission Schemes for LTE and HSPA Networks, June 2009

http://www.3gamericas.org/documents/mimo_and_smart_antennas_for_3g_and_4g_wireless_systems_May%202010%20Final.pdf

[15] Sesia, I. Toufik and M. Baker, *LTE, the UMTS Long Term Evolution: from Theory to Practice*, 2nd edition, John Wiley & Sons (2011).

[16] Alcock, S. and Nelson, R., “Application Flow Control in YouTube Video Streams”, *ACM SIGCOMM Computer Communication Review, 41*, 2 (2011).

[17] Ericsson Traffic Labs Measurements

[18] NGMN Alliance White Paper, Guidelines for LTE Backhaul Traffic Estimation, July 2011 http://www.ngmn.org/uploads/media/NGMN_Whitepaper_Guideline_for_LTE_Backhaul_Traffic_Estimation.pdf

[19] Evans, John and Filsfils, Clarence, *Deploying IP and MPLS QoS for Multiservice Networks, Theory and Practice*, March 2007

[20] Conviva Viewer Experience Report, Q1 2013, http://www.conviva.com/vxr/

[21] IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB*), Analytical method for objective scoring of HTTP Adaptive Streaming* (HAS), 2012

[22] Robinson, David C, Jutras, Yves, Cracium, Viorel, Bell Labs Technical Journal, Volume 16, Issue 4, Pages 5-23. *Subjective Video Quality Assessment of HTTP Adaptive Streaming Technologies*, March 2012.

[23] ACM SIGCOMM 12th Annual Conference on Internet Measurement, IMC, *Confused, timid, and unstable: picking a video streaming rate is hard*. Boston, MA, USA, November 14-16, 2012

[24] NGMN Alliance, *Plan to determine Video Quality of HAS in a Mobile Environment,* NGMN-SERQU HAS_TestPlan_Draft10_ChangesAccepted_Disclaimer” March 2012.

[25] Qualcomm Commissioned White Paper by iGR, *Content for All – The Potential for LTE Broadcast/eMBMS,* January 2013

[26] Qualcomm White Paper, LTE Broadcast, A revenue enabler in the mobile media era, February 2013

[27] Holma, H. & Toskala, A (eds.), *HSDPA/HSUPA for UMTS: High Speed Radio Access for Mobile Communications*, John Wiley & Sons (2006).

[28] ISO/IEC 14496-10 – MPEG-4 Part 10, Advanced Video Coding, Annex G Scalable Video Encoding

[29] W3C Web & TV Workshop, 3rd Annual, *Scalable Video Coding based DASH for efficient usage of network resources”* position paper, September 2011 http://www.w3.org/2011/09/webtv/papers/Sanchez_et_al_W3C.pdf

[30] RFC 3481 TCP over Second (2.5G) and Third (3G) Generation Wireless Networks

[31] Ericsson Whitepaper: “LTE Broadcast” http://www.ericsson.com/res/docs/whitepapers/wp-lte-broadcast.pdf

LIST OF FIGURES

Figure 1: Number of hours per week spent watching TV or Video on various devices [4] 7

Figure 2: Projection of mobile traffic trends in the United States 8

Figure 3: Global Shares of YouTube and Real-Time Entertainment Traffic[5] 9

Figure 4: Streaming Video usage by users on different days in 2009 [FLO TV Incorporated] 10

Figure 5: HTTP Adaptive Streaming 16

Figure 6: Message exchanges during HTTP Adaptive Streaming 17

Figure 7: LTE Spectral Efficiency as a Function of the System Bandwidth [10] 23

Figure 8: HET-NET architecture [13] 25

Figure 9: Effect of video playout buffers on Relative network load (X-axis) and Relative battery consumption (Y-axis) [17] 31

Figure 10: Typical backhaul architecture [18] 32

Figure 11: Cell throughput during busy time and quiet time [18] 33

Figure 12: Performance of HAS TCP versus Theoretical TCP [22] 38

Figure 13: Downward spiral behavior of video throughput [23] 39

Figure 14: HET-NET Architecture 45

Figure 15: eMBMS/LTE Broadcast Architecture 48

Figure 16: Content Delivery Network (CDN) Architecture 49

Figure 17: Rollout of an operator’s IPTV service 50

LIST OF TABLES

Table 1: Data Consumed by Different Streaming Applications[10] 21

Table 2: Performance comparison of 3GPP mobile technologies [10] 22

Table 3: Potential use cases and business models for LTE Broadcast services [31] 47

Table 4: Standardized QCI characteristics [from 3GPP TS23.203]…………………………………….…55

Table 5: Recommended choice of bit rates for video streams……………………………………….…67

ACKNOWLEDGEMENTS

The mission of 4G Americas is to promote, facilitate and advocate for the deployment and adoption of the 3GPP family of technologies throughout the Americas. 4G Americas' Board of Governor members include Alcatel-Lucent, América Móvil, AT&T, BlackBerry, Cable & Wireless, CommScope, Entel, Ericsson, Gemalto, HP, Nokia Siemens Networks, Openwave Mobility, Powerwave, Qualcomm, Rogers, T-Mobile USA, and Telefónica.

4G Americas would like to recognize the joint project leadership, great efforts and important contributions of Kamakshi Sridhar (Alcatel-Lucent) and Gautam Talagery (Ericsson) as well as representatives from the other member companies on 4G Americas.