Broadcast and Multicast Communication Enablers for the Fifth-Generation of Wireless Systems

Deliverable D5.1
Content Delivery Vision

Version v1.1

Date: 2017/11/30

Document properties:

Grant Number:	761498
Document Number:	D5.1
Document Title:	Content Delivery Vision
Editor(s):	Nivedita Nouvel (Broadpeak)
Authors:	David Gomez-Barquero (UPV); Athul Prasad (NOK), Steve Appleby, Tim Stevens, Rory Turnbull (BT); Nivedita Nouvel (Broadpeak); Peter Sanders (One2Many); Tuan Tran (Expway); Waqar Zia (Nomor)
Contractual Date of Delivery:	2017/11/30
Dissemination level:	Public
Status:	Final
Version:	1.1
File Name:	5G-Xcast D5.1_v1.1_web

Abstract

This deliverable sets out the vision of a flexible, self-organizing content delivery network which can combine unicast, multicast, broadcast and caching to deliver content cost-effectively at scale. It outlines the broad operation of such as system in a network-agnostic manner and identifies and proposes the core functional elements and the key technology components work together in an overall content delivery system to help realise the 5G-Xcast content delivery vision.

Keywords

5G, caching, broadcast, content delivery, multicast, unicast, LTE Broadcast

Disclaimer

This 5G-Xcast D5.1 deliverable is not yet approved nor rejected, neither financially nor content-wise by the European Commission. The approval/rejection decision of work and resources will take place at the Mid-Term Review Meeting planned in September 2018, after the monitoring process involving experts has come to an end.

Executive Summary

There is a growing demand from operators to deploy converged video delivery solutions across all their networks (fixed and mobile) and across all the screens, to save on operational and equipment costs.

Focusing on the Media & Entertainment vertical, this document reviews relevant business models and technology, and suggests approaches to deliver content in a flexible way; this flexibility being apparent in both the delivery mode (point to point and point to multipoint) and delivery type (fixed and mobile). We refer to these approaches as the Content Distribution Framework (CDF).

Different business models are applied when it comes to monetising video delivery. They are important to keep in mind since the revenue flow can greatly impact the success of a service. In addition, we can learn from the success or failure of relevant technologies. This leads to the conclusion that the CDF should adopt the following principles:

- It should be a generic data delivery framework, and should not rely on the business case from one vertical.
- It should have a CDN for global coverage, with point-to-multipoint-capable networks at the edge for efficiency.
- The CDN/network interface should be kept as simple as possible, and applications on end devices should not need modification.
- It should treat PTM as an internal optimisation capability for network operators which does not need to be exposed as an externally consumable network service in its own right.
- The use of a PTM capability should be dynamic, following consumption patterns and business rules.
- The use of local, or on-device storage to allow synchronous, pre-emptive delivery of content efficiently using a PTM capability would be advantageous.
- Application-layer intelligence is preferred over network features, particularly where the requirements on latency and packet loss are relatively relaxed, such as in typical media delivery scenarios.

Table of Contents

Executive Summary 1
Table of Contents 2
List of Figures 4
List of Acronyms and Abbreviations 5
1 Introduction 7
1.1 Objective of the document 7
1.2 Structure of the document 7
2 Video Delivery Market 8
2.1 Video services 8
2.1.1 Linear Television 8
2.1.2 Video-on-Demand (VoD) 9
2.1.3 Non-linear television 10
2.2 Mobile video traffic: market size assumptions 10
2.3 Business models 10
2.3.1 Operator centric business model (On-net model) 11
2.3.2 Transparent caching business model 11
2.3.3 Operator CDN business model 12
2.3.4 Content provider centric business model (OTT Model) 12
2.3.5 Mobile video delivery business model 13
2.3.6 Mobile Broadcasting Business Case 15
2.3.7 Impact of net-neutrality 16
3 Media delivery technologies – Past and Present 17
3.1 The rise of HTTP 17
3.2 Broadcast to mobiles 18
3.2.1 DVB-H 19
3.2.2 MediaFLO 19
3.2.3 ISDB-T One-Seg 20
3.2.4 T-DMB 20
3.2.5 LTE-Broadcast 21
3.3 Lessons learnt 23
4 Media delivery technologies – Future 24
4.1 ATSC 3.0 24
4.1.1 Overview 24
4.1.2 DASH-IF ATSC broadcast profile 25
4.2 Information Centric Networking 25
4.2.1 Terminology: ICN, CCN & NDN 25
4.2.2 CCN Basic Description 26

4.3 Role of NFV 27
4.4 Multi-access Edge Computing 27
4.5 Multi-link (ML) 28
4.6 MPTCP 29
4.7 QUIC 30
5 5G Perspectives and Content Delivery Framework Principles 32
6 References 35

List of Figures

Figure 1. Euro 2016 – England vs. Wales- peak audience resulting in peak of network resources consumption. 8
Figure 2. Channel change at halftime – irregularity in consumption patterns that leads to reserve unnecessary resources for some periods. 9
Figure 3. The Grand Tour – shows that even VoD content highly time-sensitive. 9
Figure 4. Expected growth of different types of content consumption on mobile devices. 13
Figure 5. High growth of live video on mobile. 13
Figure 6. AT&T and Verizon LTE mobile download speeds. 14
Figure 7. Hybrid (broadcast/broadband) delivery in ATSC 3.0 using MPEG-DASH. 24
Figure 8. ATSC 3.0 protocol stack. 25
Figure 9. Mobile edge computing framework [23] adapted to multi-access and potential related use cases. 28
Figure 10. 5G architecture with Multiconnectivity (ML-GW: multilink gateway; ML-MW: multilink middleware). 29
Figure 11. Video broadcasting using MPTCP. 29
Figure 12. The location of the QUIC in the TCP/IP protocol stack. 30
Figure 13. Unicast / Multicast / Broadcast convergent video delivery. 34

List of Acronyms and Abbreviations

3GPP	3rd Generation Partnership Project
ABR	Adaptive Bit rate
ACK	Acknowledgement
ALC	Asynchronous Layered Coding
ALTSCV	Alternative Services
API	Application Programming Interface
ARQ	Automated Repeat Request
ATSC	Advanced Television Systems Committee
BC	Broadcast
CCAP	Converged Cable Access Platform
CCN	Content Centric Networking
CDN	Content Delivery Network
CDNaaS	CDN-as-a-service
DOCSIS	Data Over Cable Service Interface Specification
DVB	Digital Video Broadcasting
eMBMS	enhanced Multimedia Broadcast/Multicast Service
EC	European Commission
FDT	File Delivery Table
FEC	Forward Error Correction
FLUTE	File Delivery over Unidirectional Transport
HEVC	High Efficiency Video Coding
HTTP/2	Hypertext Transfer Protocol 2.0
ICN	Information Centric Networking
IETF	Internet Engineering Task Force
IP	Internet Protocol
LTE	Long Term Evolution
mABR	multicast Adaptive Bit rate
MAC	Media Access Control
MBMS	Multimedia Broadcast/Multicast Services
MC	Multicast
MPEG	Moving Picture Experts Group
MPTCP	Multipath Transmission Control Protocol
MOOD	MBMS Operation On-Demand
NACK	Negative Acknowledgement
NDN	Named Data Networking
NMS	Network Management Systems
NORM	NACK-Oriented Reliable Multicast
NRL	Navy Research Laboratory
OSS	Operations Support System
OTT	over-the-top service
PARC	Palo Alto Research Centre
PDCP	Packet Data Convergence Protocol
PoC	Proof of Concept
PoP	Point of Presence
PTM	Point to Multipoint
PVR	Personal Video Recorder
QoE	Quality of Experience
QoS	Quality of Service
QUIC	Quick UDP Internet Connections
RFC	Request for Comments
RLC	Radio Link Control

ROUTE	Real-time Object delivery over Unidirectional Transport
RTCP	Real-time Transport Control Protocol
RTSP	Real-time Streaming Protocol
RTT	Round-Trip Time
SDI	Serial Digital Interface
SMPTE	Society of Motion Picture and Television Engineers
STB	Set-Top Box
TCP	Transmission Control Protocol
TLS	Transport Layer Security
UDP	User Datagram Protocol
UE	User Equipment
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
WP	Work Package
xDSL	Digital Subscriber Line

1 Introduction

1.1 Objective of the document

In this document, we articulate a set of properties that we are required in an ideal content delivery framework. Our focus is almost exclusively the Media and Entertainment vertical, although we will also touch on other uses of a content delivery framework, such as large-scale software distribution.

We will begin with a discussion on demand patterns for content and their evolution. Whilst we can be certain that the overall demand for content will continue to rise dramatically for the foreseeable future, we are much less certain of the underlying patterns of this demand. This leads us to assert that our content delivery framework should be flexible, autonomous and fast to be configured for different demand patterns.

We will derive further properties of our framework by learning from history. Previous content delivery technologies which have sought to add value to the network, have required moderately complex interfaces across organisational boundaries. We will use the rise of HTTP as the one-size-fits-all ubiquitous protocol to argue that our content delivery framework should allow cross-organisational interfaces to be extremely simple.

A review of some of motivation behind some of the newer technologies and standards for content delivery will provide us with a sense of how to exploit these within our framework.

Finally, we will conclude by outlining the main attributes and components of our ideal content delivery framework.

1.2 Structure of the document

The document is structured as follows:

- Chapter 2: Review of the video delivery market: services, market size and business models.
- Chapter 3: Review of past and present technologies for video broadcast on mobile networks, and analysis of lessons learnt.
- Chapter 4: Technologies to consider for the future.
- Chapter 5: 5G perspectives and content framework vision.

2 Video Delivery Market

In this chapter we provide an overview of the video delivery market, by reviewing the video services that a video distribution infrastructure has to deal with (each application bringing its own technical constraints), share some figures about the current market size and its future projection, and explore the different business models that are at stake.

2.1 Video services

2.1.1 Linear Television

The recent increase in over-the-top (OTT) media consumption has challenged network service providers (NSPs) to support more content, more devices, and more users with limited infrastructure resources. While developments in adaptive bit rate (ABR) technology enable telecommunications and cable television / fixed network operators to deliver video content to a growing number of new screens and applications, predicting the popularity of live video content is challenging, as major news and sporting events are often fast-breaking, causing a volatile peak in video consumption. If millions of users simultaneously watched a live event, the peak in video consumption would require a significantly large investment for an operator using a traditional unicast delivery method, where each viewer counts as a unique unicast session. The exponentially growing availability of 4K content, which requires a large amount of bandwidth to be streamed, appears as a new issue for network operators.

Graphs below (courtesy of Matt Stagg, BT/EE) provide evidence for dynamic reconfiguration of network capabilities.

It is possible to define an average popularity for a channel that may drive the default allocation of point-to-multipoint (PTM) resources. This allocation has to be dynamic and respond quickly to the demand in order to take into account rapidly changing conditions.

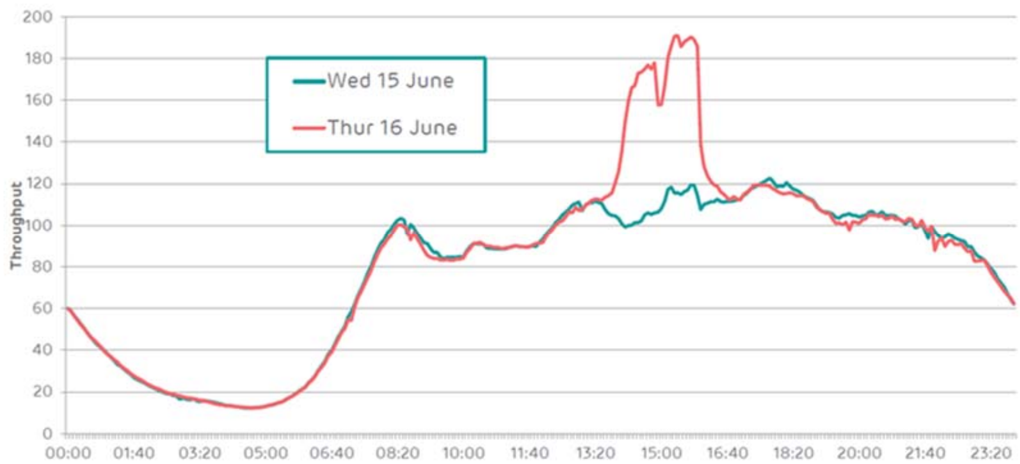

Figure 1. Euro 2016 – England vs. Wales- peak audience resulting in peak of network resources consumption.

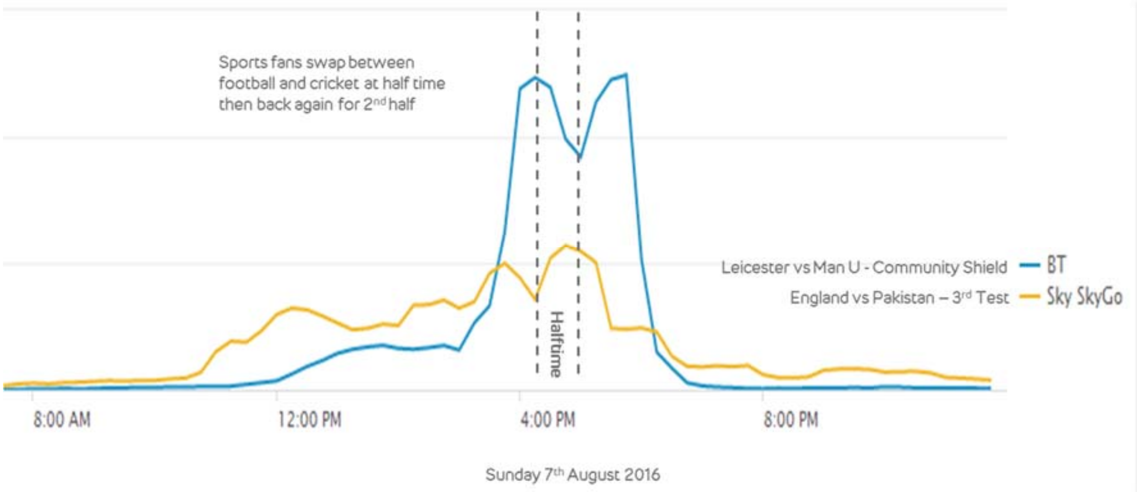

Figure 2. Channel change at halftime – irregularity in consumption patterns that leads to reserve unnecessary resources for some periods.

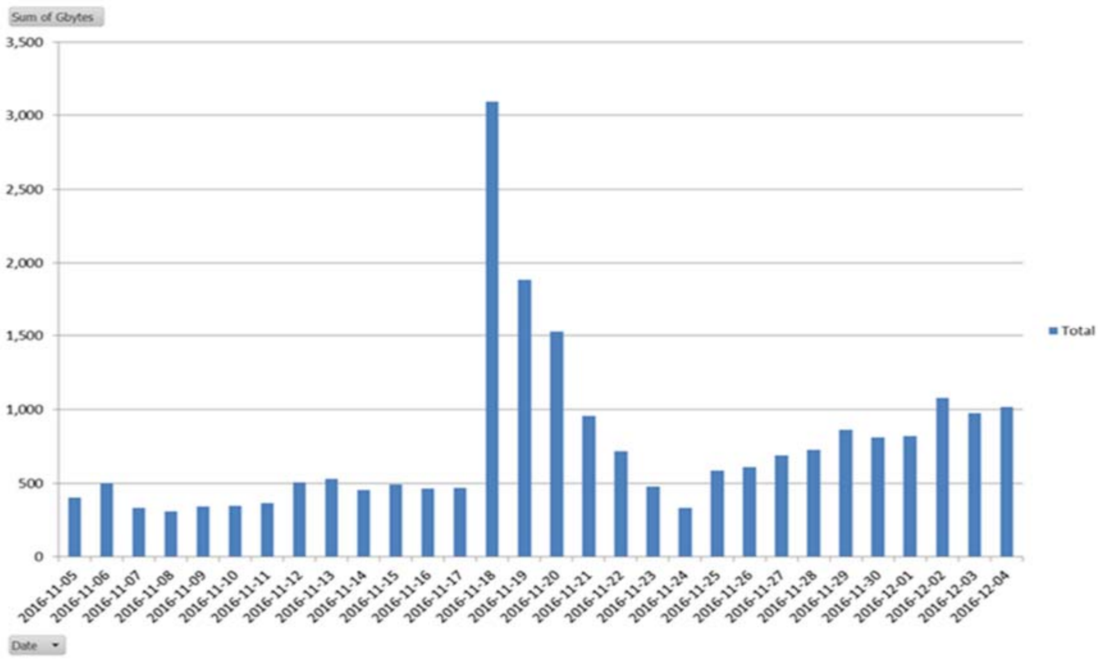

Figure 3. The Grand Tour – shows that even VoD content highly time-sensitive.

2.1.2 Video-on-Demand (VoD)

Several factors have contributed to the rise in VoD consumption. Free services such as movie trailers have seen great success: trailers enable subscribers to see movie clips for free before choosing them. There has also been an increase in pay-per-view on-demand services, enabling end-users to watch premium HD VoD content at any time. This creates several issues for NSPs. When subscribers consume a lot of bandwidth to watch free content, such as trailers or catch-up TV content, it incurs substantial costs for NSPs. Thus, NSPs need to find a way to save bandwidth, especially when subscribers do not pay to watch the content. In addition, NSPs need to figure out a method of providing subscribers who are not eligible for broadband with access to HD videos on-demand.

What characterizes VOD consumption is the existence of some very popular content that will concentrate most of the requests and long-tail contents that are watched by very few people.

The content caching strategy has to take this into account for finding the best trade-off between storage requirements and delivery cost.

2.1.3 Non-linear television

Non-linear TV collectively refers to all the applications that consist of consuming content that was available on a linear television channel after the broadcast period.

It covers network time-shifting, start-over, catch-up TV and cloud-PVR, which can be programme- or time-based.

A further constraint on caching may be imposed for cloud-PVR applications in cases where maintaining a private copy of each recording per end-user is mandatory for legal reasons: no common copy can be shared or cached.

These use cases have an impact on the global design of the video delivery solution, in particular for the content caching aspects and for the switching between unicast and multicast streaming modes that they may induce.

2.2 Mobile video traffic: market size assumptions

The following figures are derived from Cisco Visual Networking Index [1]:

- Globally, mobile video traffic will grow 8.7-fold from 2016 to 2021, a compound annual growth rate of 54%.
- Globally, mobile video traffic will reach 38.1 Exabytes per month by 2021, up from 4.4 Exabytes per month in 2016.
- Globally, the number of video capable devices and connections will grow 1.7-fold between 2016 and 2021, reaching 7,108 million in number.
- Globally, video capable devices and connections will be 61.4% of total mobile connections by 2021, compared to 50.6% in 2016.
- Video will be 78% of global mobile data traffic by 2021, compared to 60% at the end of 2016.
- Live Internet video will account for 13 percent of Internet video traffic by 2021. Live video will grow 15-fold from 2016 to 2021.

There is a growing demand from operators to deploy converged video delivery solutions across all their networks (fixed and mobile) and across all the screens, to save on operational costs and on equipment costs. But the replacement or RTSP video delivery that is currently mostly used to address main screens, with HTTP ABR delivery first requires that issues mainly related to live video delivery be solved:

- Latency from the live edge.
- Scalability for peak events.

2.3 Business models

Different business models are applied when it comes to monetizing video delivery. They are important to keep in mind since the revenue flow can greatly impact the take-up of the service (as will be seen in Section 3).

There are also some specificities associated with the business models related to mobile video delivery and broadcasting.

Eventually, net neutrality and its local variations in terms of legal constraints must be integrated in the business approaches.

2.3.1 Operator centric business model (On-net model)

Network operators have been keen to monetize the delivery of video services on their infrastructure. Telecommunications and cable television operators have started deploying their own Content Delivery Network to provide to their subscribers with linear channels and on-demand videos negotiated with content-right owners. This model was first available in an IPTV or Cable TV context and has been extended to HTTP Adaptive Bit Rate streaming technology to address new screens (mobile devices and connected TVs). The principle of popularity-based caching was implemented by technology providers for these operators, who own data centres in strategic locations where they can deploy streaming servers. They also negotiate bandwidth with transit providers and sign peering agreements between them, in order to extend their reach. They can also rely on 3rd party networks to address more consumers.

Telecommunications and cable television operators can also be mobile video operators, if they manage a 3G/4G infrastructure. They usually try to provide “cross terminal services” that allow consuming the same video content on any device and with shared bookmarks, which is a good way of preventing subscriber churn.

They have the full control of their infrastructure including the last mile, and are therefore in the best place to optimize the video delivery quality. They can also access multicast resources that are key for scalable live video delivery. Their position is however challenged by content providers who want to have a direct access to their viewers.

A key challenge for operators is to come up with a fully converged solution, allowing them to address all networks and devices with minimal operation cost.

Cloud PVR is thus an important evolution for them, especially if they can handle it in a shared storage context (use the same copy of the content to address all the users who have required it instead of producing one copy per user). Its development greatly depends on local regulation policies concerning content rights.

2.3.2 Transparent caching business model

Transparent caching consists, for an operator, to cache unmanaged content on their network, based on its popularity, in a way that is transparent to the content provider.

The benefit of transparent caching is twofold: it reduces ingress bandwidth usage for the operators, and increases quality for the content owners and end-users. Transparent caching is deployed in the network by intercepting end user HTTP traffic at routers or switches, and all requests are directed to internal caching servers. The mechanism for caching operation is as described below:

- A user requests the content.
- A router or switch on the network determines the HTTP packets and redirects them to the transparent cache placed in the network instead of forwarding them to the origin server for delivering content directly upon end-user request.
- The transparent cache inspects the situation and checks whether the requested content is already cached or not.
- If the object/content is stored in the cache, the transparent cache will deliver the content from the cache server itself.
- If the object is not available in the cache server, the system will send the request to the origin server. The content will get stored in the cache for future use.

The main issue with transparent caching is that it does not work with HTTPS streaming which is growing in its usage, due to increasing concerns related to users' privacy (for e.g., Apple and Google have made it mandatory for their apps). One of the key HTTPS objectives is to guarantee that the content is served by an authenticated server. By bypassing this mechanism, the usage of transparent caching triggers the display of a warning message on modern user agents.

Moreover, some content providers are not keen on having their content cached, without them being able to control this behaviour. They can use dynamic URLs or specific tags to prevent the caching of their content assets, especially if, like Netflix or Google, they provide their own caching system. Some CDNs also adopt this behaviour in order not to see their traffic reduced by third-party caching.

This model is thus destined to disappear.

2.3.3 Operator CDN business model

The Operator CDN approach allows operators to become the local-level CDN provider of their own network, maximizing the monetisation of their infrastructure.

They can sign a contract with content providers to allow them to access their infrastructure in a privileged way, including the multicast capabilities of their network if available.

This approach has many benefits: it provides revenue to operators allowing them to finance the growth of their infrastructure; it brings high quality content to end-users and a better service to content providers (especially for live content if multicast ABR technologies can be used).

But it requires that content providers sign contracts with possibly a large number of operators, if they want to benefit from their services in a large region or in several countries and the approach raises the debate about net neutrality.

2.3.4 Content provider centric business model (OTT Model)

Content distributors and Over-The-Top video service providers such as Netflix, Hulu and Amazon Prime Video continue to take a larger share of the value chain and are providing solutions that are pushing the bounds of existing communications infrastructure to affordably deliver content to consumers. This represents a shift in delivery expectations and customer experience that has previously been driven by telecom providers.

Most of them are using CDN-as-a-service (CDNaaS) in order to have a global reach. CDNaaS providers, like Akamai, Level3, Limelight help content providers to address their end-users all over the globe. They have their own data centres, points of presence, and transit agreements with operators to carry content across the globe.

With this approach, however, content providers do not get a commitment on the last-mile traffic quality that only network operators can control. And they usually cannot access multicast resources for their linear channels, since most operators do not allow 3rd party content providers to use them. This means that they have to pay the CDNaaS for every single viewer of a live channel. In case of events generating a peak audience, this can easily get out of control.

Some content providers such as the BBC, Netflix or Google have therefore built their own CDN in order to keep the control of the entire end-to-end delivery chain. They provide their own servers to operators (for free) who can deploy them in their infrastructure to optimize the quality of the videos. This allows content providers to reduce greatly their operational expenditure, but they need to invest in R&D and in capital expenditures (CAPEX) to build their own solution. This approach is not accessible to smaller content providers, and network operators will likely limit the number of third party infrastructure nodes that they are willing to deploy in their network – even if they are provided for free.

2.3.5 Mobile video delivery business model

Mobile operators deliver ever increasing quantities of video, including live video, to mobile devices. These are currently mainly delivered in unicast. Cisco VNI research (and others) show huge quantities of video are and will be delivered over the mobile networks, putting a great load over these networks, resulting in both reduced performance and reduced profits.

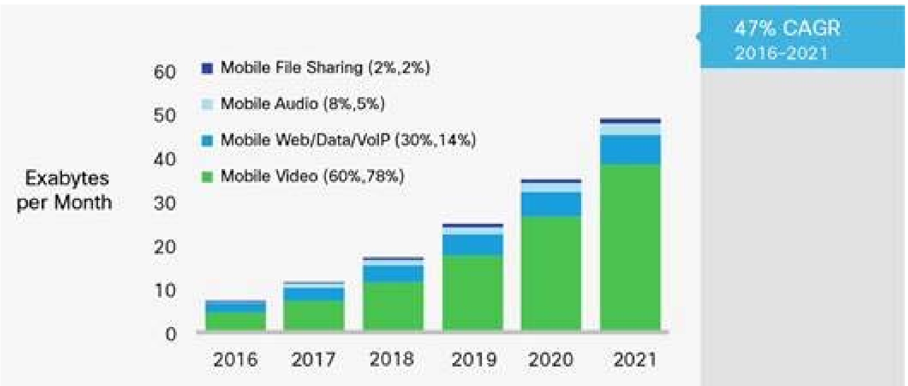

Figure 4. Expected growth of different types of content consumption on mobile devices.

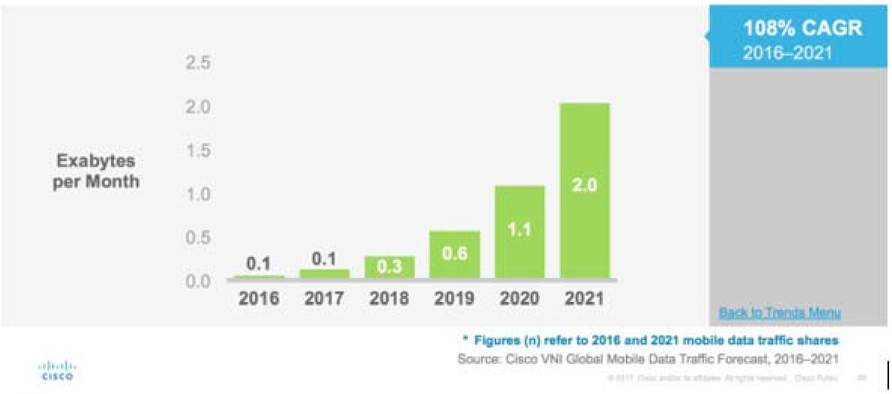

Figure 5. High growth of live video on mobile.

Recently, in 2017, USA mobile operators have taken several steps to handle these problems in their networks. Recent reports (2017) show USA operators have issues with video streaming [2]. Attributed to over-utilisation of their networks, either because of more users and/or because of unlimited data packages which are mostly demanded and used for video streaming, these US mobile network operators (MNOs) have been either actively throttling video streaming speeds and/or suffering from overall reduced service speeds.

The reduced speed experienced by the USA subscribers is shown in Figure 6 [3].

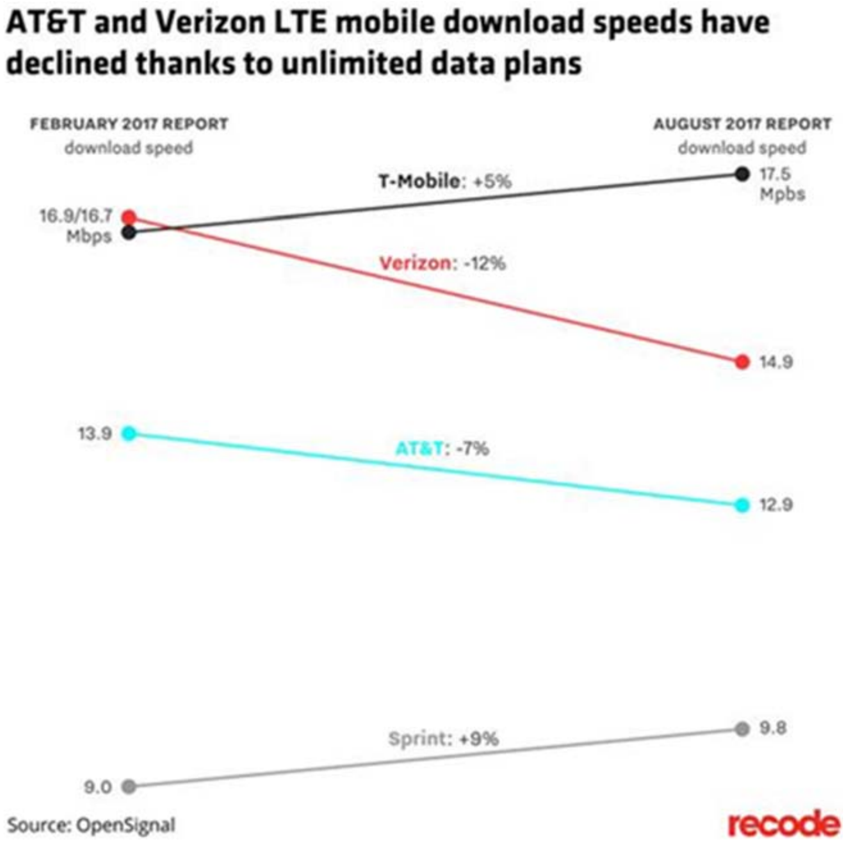

Figure 6. AT&T and Verizon LTE mobile download speeds.

Active steps taken by the MNOs include specifically capping video speeds. For example, Netflix speeds on Verizon Wireless appear to be capped for some customers [4]. While that current cap apparently was set at 10Mbps, which may be sufficient for mobile devices HD video, it is not enough for other devices, and especially for 4K video quality.

Other US operators set the cap much lower. For example, T-Mobile unlimited data plan only allowed 480p video, and then only in response to their competitor (Verizon), was forced to add HD, though seemingly with 720p resolution rather than full HD and limited bandwidth [5].

It is worth noting that these attempts by the MNOs to control their networks performance degradation and business models due to video streaming are also tough on regulatory issues. Capping specific applications or providers might infringe Net-Neutrality and similar regulations where applicable. It would hence not be a valid step in some markets.

5G-Xcast broadcast, multicast, multi-link and caching technologies target these issues with traditional MNO business models related to video. These issues may be resolved with 5G-Xcast, not because of the 5G capacity itself (which brings in more expensive resources), but because of the non-linear nature of the resource allocation provided by using these technologies where there's a "short tail" of the same content being watched by several viewers.

Also, higher QoS/QoE and network performance can be achieved with 5G-Xcast multi-link technology that can offer seamless transition between coverage areas and its seamless support of the option to serve basic video in the multicast and higher quality video to relevant users in unicast (users that have the bandwidth at that moment).

2.3.6 Mobile Broadcasting Business Case

The expenses associated with building, maintaining, and servicing a mobile broadcasting network can be high, but that does not mean that a viable business model is not possible [6]. Broadcast is economically advantageous because the infrastructure cost is amortized over those customers who receive a multicast-eligible service on a multicast-capable device. As the user base grows, the cost per subscriber falls. Note that this is unlike unicast cellular architectures, where the cost per supported user is fixed. Hence, everything comes down to consumer demand and population density.

Reaching the break-even point to recover the costs for infrastructure depends on the user population and the average revenue per user (ARPU). For a given technology, the business will be thus more likely to succeed in populated areas and where there is a high consumer demand. This fact may prevent achieving nation-wide coverage, limiting the deployment to urban areas. The main business aspects when introducing a new mobile broadcast technology are the necessary CAPEX and operating expenditures (OPEX) together with the consumer demand. Basically, the high-level questions are:

- How much investment is needed to deploy a mobile broadcast network, and what is the annual operational cost?
- To what extend are consumers willing to pay for mobile broadcasting services? Should they even be aware of the technology that is used to bring the content to them, and have the unicast / broadcast notion exposed in the offer they subscribe to?

A complete answer to these questions requires considering many aspects related to technology, business and regulations. But it is clear that nowadays paying for mobile TV content is not a mass impulse, despite the fact that early DVB-H and MediaFLO trials were encouraging to operators and willingness to pay was high [7]. Today, it is commonly accepted that mass-market demand for mobile multimedia entertainment is conditioned to the provision of these services at low cost.

The successful deployment of new mobile broadcasting systems is also conditioned by several business and regulatory related aspects. The main problem is that broadcast and cellular networks belong to different industries, each of them developed in a different business environment under very different value chains and revenue models. Mobile broadcasting might give rise to several business models with various possible roles [8]. The balance between TV broadcasters, cellular network operators and broadcast network operators will depend on the local regulatory regime. A good description of the country-specific implementations considered at DVB-H/MediaFLO times can be found in [7]. Assuming that no player would be able to control the complete value chain in most countries, broadcasters and cellular network operators may need to work together, but need a suitable business model in an extremely complex business environment, each playing to their strengths.

A collaborative model could be achieved with the presence of a wholesaler of the mobile broadcasting service acting as datacast operator. This model leaves potential retailers negotiate a sustainable business case enabled by cooperation [9]. The key objective is to have one neutral service provider (wholesaler), controlling a single mobile broadcasting network and sharing the scarce spectrum resource with several cellular network operators and content providers. The wholesaler can be a consortium of cellular operators, a broadcast network operator (due to its neutral position in the market and technical competence), or any third party. The service provider is responsible for the technical service provision, and acts as a facilitator for the cellular operators in the aggregation of content and the usage of broadcast transmission capacity. This model enables fully interactive services and a common bill for users, since the cellular operators bill their customers for the mobile broadcasting service. Moreover, it also overcomes the inefficient usage of spectrum resources because cellular operators share a common service package, being also possible to differentiate their service offerings to some extent.

2.3.7 Impact of net-neutrality

Different Internet markets are structured differently, both from a regulatory perspective and in terms of the ecosystem of companies that collectively provide Internet services. From the perspective of 5G-Xcast, it is useful to understand some of the views on net-neutrality to understand where it might have an impact on our approach to architecture.

Net-neutrality is a somewhat ill-defined principle, that all traffic on the Internet should be treated equally. Specifically, an Internet access provider should not be able to prioritize their own services over a competitors', nor should they prioritize one service provider's service over another, perhaps because there is a commercial arrangement.

It is very difficult to write down a precise definition of what the really means, particularly in a legally-enforceable way.

One argument against the requirement for net-neutrality legislation is that competition will deal with it [10]. The argument is essentially that, if services that consumers wish to use are being blocked by their ISP, then they will simply move to another ISP.

In the US, in May 2017, the Federal Communications Commission (FCC) voted to withdraw net-neutrality legislation introduced during the Obama administration [11].

In Europe, following EU Regulation, the Body of European Regulators for Electronic Communications (BEREC) has published guidelines which explain to the National Regulatory Authorities (NRAs) how to interpret the legislation [12] BEREC's summary of the EU requirements is;

"ISPs are prohibited from blocking or slowing down of Internet traffic, except where necessary. The exceptions are limited to: traffic management to comply with a legal order, to ensure network integrity and security, and to manage congestion, provided that equivalent categories of traffic are treated equally. The provisions also enshrine in EU law a user's right to be "free to access and distribute information and content, run applications and use services of their choice". Specific provisions ensure that national authorities can enforce this new right."

In the more detailed guidelines [13] BEREC makes it clear that traffic can be prioritized (and by implication other traffic slowed down), according to the generic QoE requirements of the traffic. That is, video could say, be prioritized in a different way from audio, but it would not be legal to prioritize one video service provider's traffic over another.

Within 5G-Xcast, we should not concern ourselves too much with the regulatory requirements of any one region or point in time, but we should ensure that our approach is flexible enough to adapt to different regulatory environments. In many cases, the approach that 5G-Xcast takes will be independent of net-neutrality. Compliance with local net-neutrality legislation will often depend on how an operator deploys technology, rather than the technology itself.

3 Media delivery technologies – Past and Present

The objective of this section is to inspect several technologies that have been specified and/or deployed to handle video delivery. A specific focus will be made on PTM technologies.

The success or failure of these technologies will be discussed to assess what lessons can be learnt from them in the perspective of structuring requirements for the 5G standards.

3.1 The rise of HTTP

In the past, TV-quality video was treated as a special type of service. Bandwidth requirements and timing constraints meant it was a significant engineering challenge to deliver such video at scale. However, increasing network bit rates and decreasing storage costs mean that TV-quality video is becoming less 'special'. This is not to say that the bandwidth and storage required by video services does not still present a challenge, just that there is now less of a need for a fully, vertically integrated TV distribution platform. Instead, Media and Entertainment requirements can now be included in the list of requirements for generic networks.

Currently, digital terrestrial and satellite television services are delivered using MPEG-2 Transport Stream multiplexes. These transport streams are designed specifically for broadcast TV requirements and, whilst it is possible to deliver non-media payloads such as EPG data, these are intended to supplement the TV service.

In contrast to this, Internet protocols tend to be generic. HTTP now dominates as the protocol of choice for non-conversational services. Even services such as video streaming, which would previously have used a specialized protocol, now use HTTP. However, HTTP is designed for downloading files, and its underlying TCP transport protocol means that although the files are guaranteed to arrive intact, the bit rate and delay are highly unpredictable. This unpredictability does not make HTTP a good candidate for a video streaming protocol. It is possible to argue that, whilst HTTP can be used for almost all (non-conversational) services, it is also sub-optimal in most cases.

A few years ago, an Internet protocol stack was created for video streaming, consisting of UDP, RTP, RTCP and RTSP. In combination with network QoS, this addresses the requirements of bit rate control, timing, session setup, etc. However, it has not in fact found widespread use for video streaming. Conversely, HTTP addresses none of these concerns, yet is now used exclusively.

So why is HTTP ubiquitous? It is attractive because it is generic, reusable and economically scalable. Content delivery networks need only implement the one protocol stack, and, whilst performance may be tuned according to the mix of services, fundamentally the use of HTTP dramatically simplifies network infrastructure: one example is that with HTTP, the server does not need to maintain connection state for its many clients, since state management becomes the client's responsibility. Further, reuse of the single protocol stack means that all services benefit from its capabilities. For example, Transport Layer Security, caching etc. can be used 'for free' by any service that uses HTTP. Additionally, HTTP-based protocols will traverse standard firewalls and survive NAT-translation, whereas more esoteric protocols may not. Eventually, they bring the possibility of dynamically adapting the quality of the video to the available bandwidth through multi-layer encoding and streaming.

The unpredictable throughput of TCP and the concatenation of files that represent the media segments into a single stream are dealt with at the end points. For video streaming for example, a combination of increased buffering on the client devices and adaptive bit rate streaming reduces the occurrence of stalling and therefore permits a sufficient degree of quality control.

The question then arises: is there a wider advantage to replacing specialized TV & video distribution protocols with Internet-type protocols, and how could this be done?

There are a number of advantages to using a non-specialized protocol stack:

- It is much easier to carry any data, not just media,
- The same protocol stack could be used for different network types (fixed, mobile, over-air broadcast),
- A common protocol stack will facilitate multicast carriage of otherwise unicast data and will facilitate dynamic switching between multicast and unicast,
- A common protocol stack will facilitate switching between network types (e.g. between fixed and mobile),
- If the protocol is designed to carry HTTP-like data, then it can easily interface with content service providers, end devices and network caches or CDNs.
- It allows to reduce the firmware footprint in terminal devices by using common generic software libraries.

Our conclusion is that the primary task of a globally-scalable unified content distribution system would be to deliver HTTP responses efficiently. The interfaces between the organisations en route should be based on HTTP GET request/response pairs.

We note though, that HTTP itself is evolving. HTTP/2 has recently been standardized and appears likely to be widely adopted. Separately, QUIC is a UDP-based transport protocol for HTTP (and other application protocols) which is currently being standardized and is already in use by Google. Indeed, using HTTP/2 and QUIC could have advantages for multicast delivery.

Our conclusion is that the best way to deliver content efficiently and globally is to use HTTP, in conjunction with CDNs where needed, in line with current practice. It is worth re-stating that it is not at all obvious that HTTP would be the dominant 'one-size-fits-all' protocol. We need to keep in mind the lessons learnt from this before adding additional complexity to the network or to the interfaces between the network operators and the content service providers.

3.2 Broadcast to mobiles

Several technologies have been specified and to some extent deployed to address point-to-multipoint communications video delivery. This section reviews the ones that have had the biggest impact. In particular, DVB-H (Digital Video Broadcast – Handheld), MediaFLO (Forward Link Only), ISDB-T (Integrated Services Digital Broadcasting – Terrestrial) OneSeg and T-DMB (Terrestrial – Digital Multimedia Broadcasting). The only relative success stories among the first generation of mobile broadcasting systems are ISDB-T in Japan and T-DMB in South Korea. But in these two cases mobile broadcasting services are a demand from the national regulators, and it should be pointed out that the services are not yet commercially profitable. Only free-to-air mobile television services are available, and the incomes achieved with advertising do not compensate the costs for running and maintaining the networks. An important advantage of ISDB-T is that mobile services are provided in-band, using a narrow portion of the channel bandwidth, while the rest of the bandwidth is dedicated to the transmission of stationary services. This way, there is no need to deploy a dedicated mobile broadcasting network.

LTE Broadcast, finally, is a technology standardized by 3GPP that provides transport features for sending the same content information to all the users in a cell.

3.2.1 DVB-H

The most relevant mobile broadcasting technology in Europe was DVB-H [14]. After a slow start, and with a clear endorsement of the European Commission, several DVB-H networks were deployed across Europe. Italy was the first country to launch commercial services in 2006. Commercial DVB-H services were also on air in the main urban areas of Finland, Switzerland, Austria, the Netherlands, Hungary and Albania, but they have been progressively switched off.

One of the major concerns about the roll-out of DVB-H was the network infrastructure cost. Due to the fact that DVB-H terminals suffer from much more severe propagation conditions than DVB-T, a considerably large number of new sites is required for the installation of additional DVB-H transmitters or repeaters ("gap fillers"), complementing the existing broadcasting towers and forming very dense Single Frequency Networks (SFN), which implies very large investments in infrastructure.

In addition, the assumption that investments in a dedicated mobile broadcasting network, such as DVB-H, would be justified by market demand for mobile TV services proved to be wrong. Consumers were not prepared to pay monthly fees for mobile TV services, and the demand was not sufficiently high to offset the costs associated with building and operating the networks. Receiving devices also failed to materialize in mass-market numbers, and the receivers that appeared on the market were limited in choice and did not appeal to consumers. Furthermore, there was no regulatory framework which would have obliged all players in the service chain (broadcasters, broadcast network operators, cellular operators) to work together.

3.2.2 MediaFLO

MediaFLO is a proprietary technology, developed by Qualcomm [15][16]. It was designed to provide mobile multimedia services, including traditional real-time content as well as non-real-time. MediaFLO was designed aiming to be overlaid upon a generic cellular network with independent network infrastructure. The receiving devices would also be furnished with separate modules for mobile broadcast and cellular access.

MediaFLO services were commercially launched in 2007 by different partnerships formed by Qualcomm and cellular operators in the USA. The network was deployed with high power UHF transmitters and provided coverage to approximately 70 million inhabitants of major US cities. The system was not commercially successful and the network was switched off in 2011.

MediaFLO was beset by commercial and technical challenges. Perhaps the most significant was that the technology required entirely separate infrastructure to support it. When MediaFLO was launched it was not compatible with supporting 3G networks – MediaFLO being OFDM-based whereas 3G was W-CDMA and CDMA.

Another key issue when MediaFLO launched, was the scarcity of available devices that were equipped with an appropriate receiver. During this time, the iPhone was still in its 2G iteration – the mobile data revolution was about to take off but was not sufficiently established to support mobile broadcast services. Those users that could have used MediaFLO therefore, would have suffered a poor viewing experience in limited bursts (usually just 5-10 minutes of footage a day).

The issues were not purely limited to user experience either. MediaFLO also used spectrum that fell outside of that used for mobile networks – this made it costly to support and deliver. This explains why the length of mobile broadcast footage on offer was so limited and why the number of compelling supporting business models were so limited.

Faced with this list of obstacles and problems, operators, vendors, OEMs and regulators had no choice but to go back to the drawing board.

3.2.3 ISDB-T One-Seg

The ISDB system was designed for high quality audio, video, and associated data. Originally, it was developed by and for Japan. In 2006 Brazil adopted a slightly modified version of the standard, known as ISDB-T_B[1], which has been the terrestrial broadcast choice in most countries of South America.

The ISDB-T standard is based on MPEG-2 for audio/video coding and multiplexing, and it also includes advanced video coding H.264 for mobile services [16]. The system uses BST-OFDM (Band Segmented Transmission – OFDM) modulation, frequency-domain and time-domain interleaving, and concatenated Reed-Solomon and convolutional code.

The BST-OFDM modulation scheme was designed for hierarchical transmission and partial spectrum reception. In ISDB-T, a 6 MHz bandwidth channel is divided into fourteen segments. Thirteen segments are used for data transmission and one segment is used as guard band. In the thirteen data segments, it is possible to transmit up to three simultaneous linear services in three hierarchical layers with different transmission parameters and hence a specific robustness. The most common configuration employs one segment for mobile services (this is usually referred to as One-Seg), and twelve segments to transmit HDTV programs for fixed reception.

If compared to DVB-T, ISDB-T employs the same error correction scheme. Hence, their performance is very similar in static channels. However, ISDB-T makes use of physical time interleaving which makes ISDB-T outperform DVB-T in mobile conditions.

Commercial One-Seg services were launched in Japan in 2006. As of February 2012, approximately 124 million cellular phones had been sold. Despite such good figures, it should be pointed out that the service is not yet commercially profitable. But broadcasters are required to provide One-Seg services by the national regulator.

3.2.4 T-DMB

T-DMB was the first commercialized mobile broadcasting technology in the world in December 2005 [17]. It was designed to provide different types of mobile multimedia services including audio and video service (with a target resolution equivalent to CIF or QVGA for small size displays, 2-7 inches), interactive data service as well as audio-only services.

T-DMB is based on the Eureka-147 Digital Audio Broadcasting (DAB) system, which was modified in order to suit the more restrictive requirements of a system delivering video and associated audio and multimedia information.

Compared to other first generation mobile broadcasting systems, T-DMB has a lower spectral efficiency because DAB was originally designed for audio and its related data service. The maximum data rate of T-DMB in 1.5 MHz is approximately 1.6 Mbps (the effective throughput only 1 Mbps).

T-DMB (along with ISDB-T) is the only success story of first generation mobile broadcasting technologies. The advantages of T-DMB are a relatively simple service deployment due to the structural inheritance of the Eureka-147 DAB system that has been in commercial service for the past decade in many European countries, and that

[1] The main differences of ISDB-T_B compared to ISDB-T are source code (H.264 for video and MPEG-4 HE AAC for audio); middleware (Ginga); and adoption of VHF/UHF bands for channels 7 to 13, and 14 to 69, respectively.

the cost for establishing T-DMB network is relatively low due to simplicity of the system and its optimal frequency band. Since the commercial launch of T-DMB, more than 55 million receivers had been sold in South Korea as of June 2011.

3.2.5 LTE-Broadcast

LTE-B (a.k.a eMBMS) is a technology standardized by 3GPP that allows a potentially infinite number of user equipment (UEs) to consume the same content at the same time using a fixed amount of network resources. To enable LTE-B, it is important to harmonize all three mutually dependent aspects: devices, network and content source. The following aspects are recognized to play a role it the lack of commercial success of LTE-B:

Lack of compatible handsets

The current device ecosystem is weak from the device perspective. Only a few chipset vendors support LTE-B technology hence not all devices are compliant. Furthermore, the lack of interest from a major device maker like Apple leads to a weak ecosystem. Currently, LTE-B receiver firmware is included in some Android based devices such as Samsung Galaxy S6, S7, or Bittium Touch Mobile, but may not be activated by the network operator. This lack of interest from vendors may also result from an uncertain benefit vs cost ratio, a situation that may change if other drawbacks are addressed in 5G networks.

Lack of standardized APIs

No standardized APIs between the application and the eMBMS middleware exist at device side. APIs between content provider and eMBMS at the network side (i.e. BM-SC) were standardized only in 3GPP release (Rel) 14. The lack of standardized APIs prevented the application/service developers from developing in large scale because of the extensive integration efforts.

Lack of flexibility in management of MBMS Service Areas

From the network perspective, the coarse granularity of MBMS service area which is statically configured prevents flexible deployments of broadcast services. The problem has been recognized by 3GPP. Rel'14 specification addressed this issue by the introduction of geographical area property associated with MBMS session on the interface between eMBMS (BM-SC) and content provider.

Inflexible cell resource allocation

When eMBMS uses MBSFN transmission, broadcast radio resources are segregated from unicast radio resources. Unused broadcast resources can be used only by UEs in certain transmission modes, which impacts negatively resource utilisation and system performance. Up to 80% of radio resources can be allocated to MBSFN transmission in a shared MBMS/unicast-mixed cell (i.e. Rel'14 FeMBMS/unicast cell). When eMBMS uses MBSFN transmission, the eMBMS architecture does not provide mechanisms for adjusting the amount of resources allocated to MBMS and unicast or performing statistical multiplexing within 3GPP domain based on the level of MBMS traffic.

Reserved cells

The MBSFN transmission may create a significant interference in cells adjacent to a MBSFN area. Interference from the adjacent cells to MBSFN area is also undesirable because high variation of SINR within MBSFN area will impact the coverage and/or spectral efficiency of MBSFN transmission. Reserved cells at the edge of MBSFN area may need to be configured to transmit at restricted power or not transmit at all to address these issues.

Long switch times with MooD

MBMS operation on demand (MooD) enables switching between unicast and broadcast. MooD was standardized in Rel'12. The performance of MooD suffered from rather long switching time caused by LTE radio access network, which required at least 5.12 seconds to modify the configuration of MBMS until Rel'14. In Rel'14, the MBMS control information in LTE radio network can be changed within 10 milliseconds. As usual, it is up to the network configuration to find an optimal balance between signalling overhead and desirable performance.

Unclear billing functionality

The Broadcast/Multicast Service Centre (BM-SC) is able to generate charging records for the content provider's billing system based on data provided for broadcast/multicast and MBMS service users through MBMS user service consumption reporting. However, the business model around LTE-B charging is not clear. Details of LTE-B solutions would be proprietary.

Lack of Multicast Backhaul

eMBMS architecture relies on IP multicast transport in the user plane. Mobile backhaul networks do not currently support IP multicast transport – it is a significant upgrade to support multicast natively in the backhaul. Solutions to date have not been scalable.

Lack of interoperability testing

LTE-B technology has not yet matured to the stage where it is possible to mix and match vendor components.

The above weak points have prevented a strong LTE-B ecosystem in the past. However, starting from Rel'13, emergency agencies and governments pushed forward to use LTE-B to replace the old technologies such as P25 and TETRA for Public Safety solution based on LTE (PS-LTE). In the same release, a second MBMS delivery mode called SC-PTM (Single Cell Point to Multipoint) was introduced to offer a flexible coverage area at the granularity of cell level instead of MBMS Service Area, however, without the advantage a Single Frequency Network offers. Furthermore, 3GPP has made significant improvements of LTE-B in Rel'14. The interface xMB (specified in 3GPP TS 26.346 and 3GPP TS 29.116) between the content provider and the MNO is standardized. In addition, 3GPP has also completed the normative work on standardized APIs between the middleware and the application on the UE side (see 3GPP TS 26.347).

Other important features that both content provider, MNO, UEs benefit from are MooD (MBMS operation on Demand) and Service Continuity. The earlier indicates a seamless switching between unicast and broadcast based on the popularity of the content while the latter ensures a seamless experience perceived by the UE when moving within, in or out of broadcast coverage. Trial of MooD is planned for 2017 by Telstra and Ericsson using middleware provided by Expway [18].

Last but not least, it's important to have a broader range of use cases beyond the traditional ones such as sport and concerts to make LTE-B to take off. Recently, V2X and IoT verticals have shown strong needs of broadcast for their use cases. For instance, a large number of IoT devices in a given area may need to get urgent software/firmware update due to security risk. In a short period of time when autonomous cars become more popular, broadcast mode will allow mass software update to be performed, to distribute trusted certificates bundles, etc. These two verticals have been actively pushed in 3GPP. The single cell point-to-multipoint (SC-PTM) transmission of eMBMS has been adopted for Narrow Band Internet of Things (NB-IoT) UEs and Bandwidth reduced Low complexity and CE stands for Coverage Enhancement UEs.

3.3 Lessons learnt

So what can we learn from the above technology descriptions that help us create our content delivery framework?

- The content delivery framework should be **a generic data delivery framework**, and should **not rely on the business case from one vertical**.
- We should have **a CDN for global coverage**, relationships with CSP etc., but **point-to-multipoint network at edges** for efficiency.
- The **CDN/network interface should be kept as simple as possible**, and applications on end devices should not need modification.
- The attractiveness of producing devices that support point-to-multipoint should be increased, while **hiding the use of PTM from as much of the ecosystem as possible**. Network operators could offer certain services, or certain qualities, only to those handsets which can support PTM thus making PTM indirectly attractive to consumers and therefore adding value to PTM capable handsets.

So, the 5G-Xcast approach is for the CSP to sell the services and the NSP to worry about whether unicast or multicast is used on grounds of infrastructure savings.

4 Media delivery technologies – Future

The objective of this section is to investigate technologies that could be leveraged to optimize the delivery of video content within a 5G context.

A more technical and in-depth description of these technologies will be provided in deliverable D5.2.

4.1 ATSC 3.0

4.1.1 Overview

ATSC 3.0 is a suite of around 20 standards for a next-generation IP-based broadcast television in the United States and other territories [19]. The suite consists of specification on the transmission, audio, video, captioning, watermarking, and security aspects.

The important design choice for ATSC 3.0 standards was that they took the approach of enabling hybrid broadcast/broadband delivery from its roots; it's not an afterthought as with previous IP-based broadcast technologies. Here MPEG-DASH is a key enabler of this hybrid delivery, as depicted by the following figure. Hybrid DASH content delivery is enabled by the broadcast and broadband paths.

The Broadcast File Transport leg is realised by ROUTE protocol. On the unicast leg of the chain, standard DASH content, broadband distribution network, and players are used. This aspect may be the most significant part of the services layer in ATSC 3.0 which has far reaching impact and hence we focus only on the DASH side of the services delivery in ATSC 3.0. It should also be noted that the ATSC 3.0 suite of standards was not created in a vacuum and its services layer is influenced by the existing example of 3GPP DASH delivery over FLUTE [20].

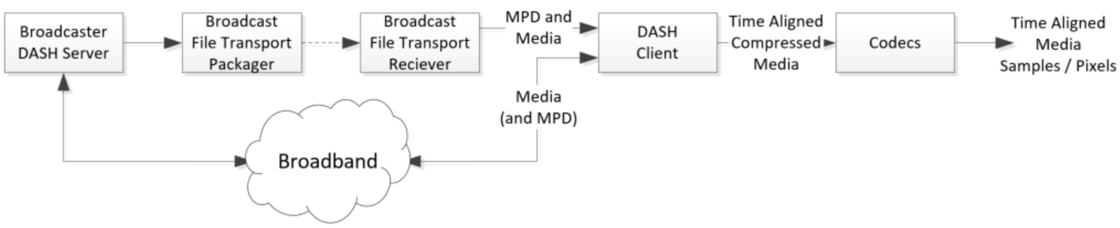

Figure 7. Hybrid (broadcast/broadband) delivery in ATSC 3.0 using MPEG-DASH.

Time synchronisation between the client and the server has to be achieved by a common UTC source. The main strengths of this design are:

- DASH enables time synchronisation between individual media content components regardless of how they were delivered: broadcast, broadband, or locally cached.
- Content protection frameworks that work with DASH can also be used in this environment.
- All key audio/video/subtitles formats are supported as with DASH.
- Ad-insertion mechanisms already enabled with DASH for unicast work as well in broadcast and hybrid environments since time and media synchronisation framework is the same.

Taking a closer look at the technological realisation, DASH unicast is realised via traditional HTTP/TCP/IP on the right-hand side while the broadcast is realised via ROUTE Transport towards the left of the unicast leg. An important design element is the HTTP proxy on the client-side that provides the interface between ROUTE and DASH.

ROUTE extends FLUTE [21], where the latter is designed specifically for file delivery; hence important optimisations were enabled by ROUTE. On the one hand, it enables the same features as providing source as well as FEC repair data. ROUTE allows for template-based instantiation on the client side using Layered Coding Transport (LCT) information. This eFDT significantly increases efficiency of object delivery by eliminating the need of sending the complete FDT with every object. eFDT can be sent by different means to the client, either on the same LCT channel as media, on a different channel, or in the Service Layer Signalling SLS. Other enhancements include the possibility of an optimized Media Delivery Event (MDE) that can be exploited by clients for even faster tune-in and channel change times.

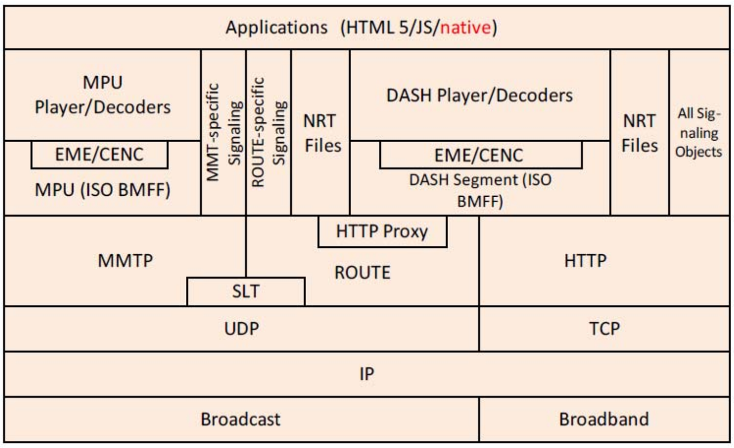

Figure 8. ATSC 3.0 protocol stack.

4.1.2 DASH-IF ATSC broadcast profile

DASH delivery is enabled in ATSC 3.0 by defining the ATSC broadcast profile in DASH-IF [22]. This work done jointly by ATSC, and DASH-IF defines a carefully pruned profile of DASH that is ideally suited to work with ATSC 3.0 broadcast. This is a key enabler for interoperability between the content creation and consumptions sides.

4.2 Information Centric Networking

4.2.1 Terminology: ICN, CCN & NDN

ICN is the umbrella term that is used to refer to the broad research direction of a content/information/data centric approach to network architecture. Under this common ICN topic, several different architecture designs have emerged in recent years. Of these, two are of particular interest to 5G-Xcast:

- Named Data Networking (NDN), which itself evolved from Content-Centric Networking (CCN).
- Hybrid ICN (hICN), an initiative by Cisco that aims to mix traditional IP networks with Content Centric Networks so that items may be carried in either network without the client being aware of the transport. This hybridisation is likely to be important in 5G.

(Note that the term “content-centric network” is also sometimes used to refer to a network of data or files organized by their content or name (sometimes using metadata), rather than by geo-location or domain address. Most peer-to-peer file sharing networks use some form of content-centric networking, through the use of semi-equal peers. This is not the same as CCN as discussed here).

We start with an introduction to CCN since this came first.

4.2.2 CCN Basic Description

A CCN refers to a set of data or files that are organized by their content or name, rather than by location or domain address. Whereas IP-based networks specify a pair of locations for source and destination (typically named locations, mapped by DNS down to IP addresses), CCNs refer to media names: the media 'address' refers to the content, not its location. With CCN, each packet, or 'content object' as it is called, has a unique name, expressed as a "/"-delimited hierarchy. Files are broken down into collections of content objects depending on their size. The CCN header contains the content object's name, and a cryptographic signature, which allows modification of the contents to be detected. Routing is determined by the content object name only: there is no destination address in the CCN object. Copies of the same named item may exist in many locations, and the CCN architecture makes the decision about how to retrieve the object for the requesting client.

With conventional networking, IP connections are made between two end points and generally TCP or UDP is used as a protocol to transfer bytes to applications between the end points. The objects that are transferred are simply treated as bytes, wrapped up in IP packets, and they are normally completely opaque to the delivery infrastructure. Content Delivery Networks employ object caching as a technique to more efficiently deliver popular content. However, the growing use of connection-level encryption such as Transport Layer Security (TLS) means that caching at the packet level is impossible and difficult at the object level.

CCN enables applications to address and request data by name rather than location. As a result, the network can locate and retrieve the data dynamically from any potential source, leading to better support for mobility. For increased security, CCN secures and authenticates the data itself, rather than creating secure point-to-point connections to authenticated hosts. This security model enables opportunistic caching of data anywhere within a CCN network, resulting in better efficiency, availability and scalability.

The key concept is that a content object is an atomic unit and can be stored anywhere. It is referenced by its name, and because of the cryptographic signature, cannot be tampered with, so its absolute location is not critical (so long as it is accessible somewhere). It can be duplicated and cached providing redundancy and be available closest to its point of consumption.

The delivery model for CCN is a pull model. A request is made for a named content object by an application sending an *Interest* request to an attached CCN router. The router looks at the name to see if it has a cached copy which can be returned (delivered to the client application as part of the *ContentObject* response), otherwise the router forwards the request to one or more CCN nodes that may be able to source the object. CCN nodes generally incorporate item caches, thus caching becomes an integral part of the network fabric and not an overlay.

CCN can easily replicate standard caching models, but more importantly CCN makes it easy to add new models and be totally flexible in the way content is delivered.

ICN is particularly interesting for 5G-Xcast since the ICN architecture itself is responsible for locating the named objects, and for fetching them over the most appropriate underlying network. This means that content can arrive over fixed or mobile routes, or a mixture. Furthermore, initiatives such as hICN should allow hybrid delivery over a mix of conventional IP or CCN/ICN networks.

4.3 Role of NFV

Network Functions Virtualisation, or NFV, is a network architecture philosophy that utilises virtualisation technologies to manage core networking functions via software as opposed to having to rely on hardware to handle these functions. The NFV concept is based on building blocks of virtualized network functions, or VNFs, that can be combined to create full-scale networking communication services.

In Network Functions Virtualisation, virtualized network functions handle specific network functions that run in one or more virtual machines on top of the hardware networking infrastructure (both wired and wireless) such as routers, switches and servers.

The combination of NFV and Software-Defined Networking (SDN) provides the opportunity to make the transition from appliance-based, proprietary hardware implementations to software-based, open platforms.

Already, there has been tremendous focus from MNOs in the evolution of their mobile edge to incorporate virtualisation into their infrastructure; from cloud Radio Access Network (RAN) through virtual LTE evolved packet core and Gi-LAN virtual services. These technologies are crucial in refining their operational models to improve margins in existing businesses, and critical as MNOs attempt to capture new revenue streams, such as machine-to-machine and next-generation mobile virtual network operators, with economics that are substantially different from their current business.

Ultimately, as the base station and the LTE Evolved Packet Core (EPC) transition to virtual functions to improve elasticity, the underlying transport infrastructure between these virtual functions must be capable of adapting and scaling in the same manner. This will be achieved by incorporating SDN and NFV capabilities into the mobile backhaul network and integrating the intelligence and understanding of the RAN, backhaul and IP flows into a holistic understanding of the mobile network capabilities, conditions and constraints.

Once this level of visibility and cross-domain programmability is achieved, MNOs will be in the position to realise the economic (reduced CAPEX, OPEX) and operational (simplified management, improved time-to-market) benefits that SDN and NFV provide, and be in a better position to successfully deliver new consumer, enterprise and M2M services and applications.

NFV allows third parties to run their software within an operator's network. For our content framework, this provides a valid way to get content caching deeper into the network and also a means to connect the caches to multicast, since the multicast root could be co-located with the CDN cache locations.

4.4 Multi-access Edge Computing

Multi-access edge computing (MEC) is considered to be the natural evolution of access technologies for communication networks, evolving and converging with the recent breakthroughs in information technology and cloud computing. MEC is considered to be a key enabler for the access networks to support new services, business use cases and deployments. Potential use cases for MEC in the context of 5G-Xcast include:

- video analytics,
- provisioning of location-based services,
- enabling internet-of-things (including mass delivery of localized content and ultra-reliable, low-latency type of industrial IoT applications),
- virtual / augmented / mixed reality (VR/AR/MR) applications (services that include delay criticality),

- localized content distribution (broadcast within stadiums, concerts and other public events) and data caching (which could be service agnostic with key applications in media and entertainment).

An overview of the MEC framework, adapted from the mobile edge computing framework presented in [23] and potential related use cases, is as shown in Figure 9. Here the potential use case dependent edge-hosting of NFVI is also considered within the overall framework. Thus, MEC has multiple potential use cases and applications in the context of Xcasting in 5G, with the enablement of new businesses and support for verticals.

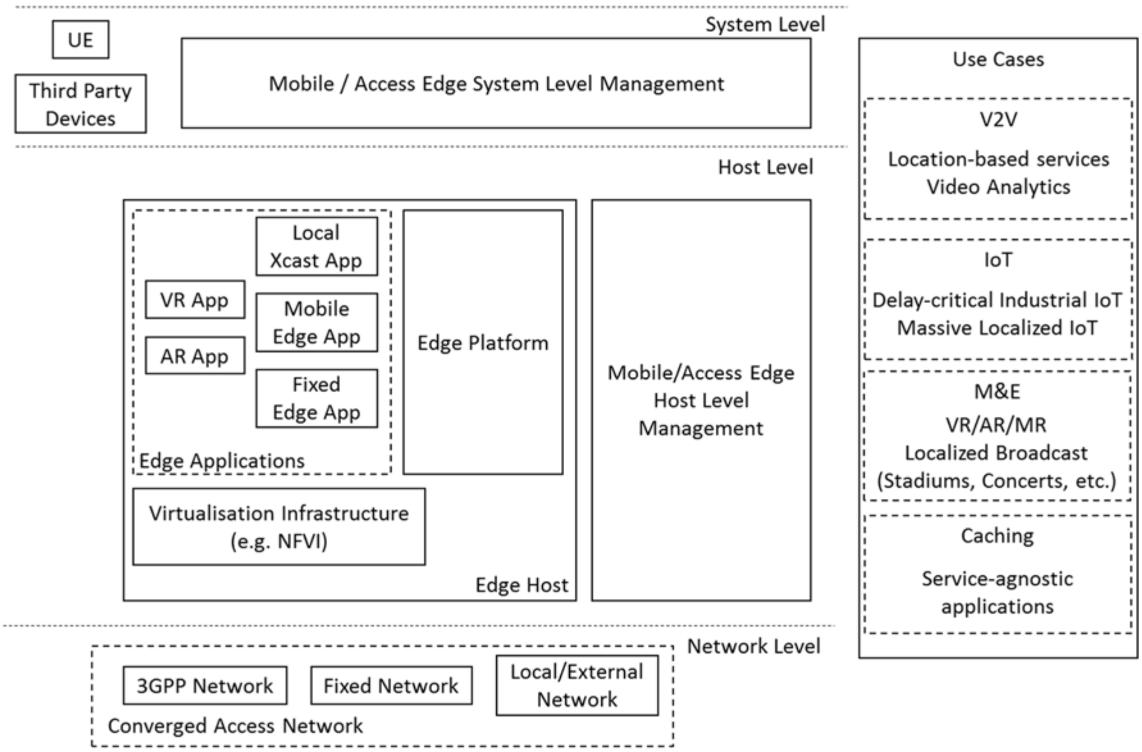

Figure 9. Mobile edge computing framework [23] adapted to multi-access and potential related use cases.

4.5 Multi-link (ML)

Multi-connectivity is the technology of using simultaneously multiple links to deliver content (IP packets). This can be used for a content delivery as an additional service for the content provider, especially in multi-homed content providers. In case of mobile networks, we usually discuss multi-link (instead of multi-connectivity) as a capability of a device to communicate via multiple wireless links in a mobile network. Figure 10 presents the usual exploitation of ML as an end-to-end tool for content delivery.

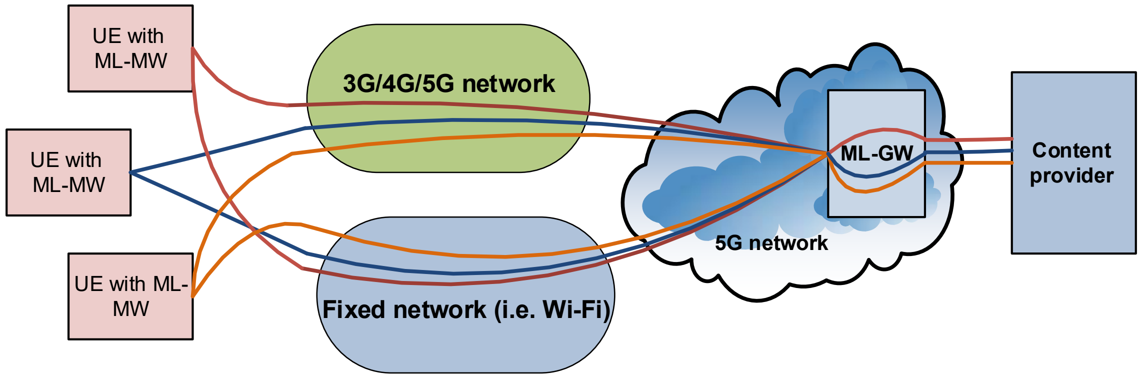

Figure 10. 5G architecture with Multiconnectivity (ML-GW: multilink gateway; ML-MW: multilink middleware).

Multi-link technology can be used to bond together various combinations of IP links:

1. 5G cellular network link of one operator with a 4G network link of the same or another operator.
2. 5G cellular network of one operator with a WiFi network link of the same or another operator.
3. 5G cellular network link of one operator with a cable, xDSL or satellite IP link of same or another operator.
4. Any number of IP links of same or different operators.

Currently multi-link bonding is implemented for unicast flows, exploring the features of different links. The end-to-end performance of each of the links between the content starting point and the end point (the bonding multilink middleware typically running on the end user device) have different results over each of these links between users, so that a simplistic multicast/broadcast over a common set of links will not work in many cases.

ML can provide interesting performance improvements on the core network, and can its interoperation with a CDN distribution network is worth pursuing. Furthermore, the exploitation of CDN edge servers as a ML-GW, exploring the CDN edge server as the source of ML connections is also an interesting option. These technologies will require proper support from either (or both) the network and transport layer.

4.6 MPTCP

Multipath TCP (**MPTCP**) is an ongoing effort of the Internet Engineering Task Force's (IETF) Multipath TCP working group, that aims at allowing a Transmission Control Protocol (TCP) connection to use multiple paths to maximize resource usage and increase redundancy for video content delivery, see Figure 11.

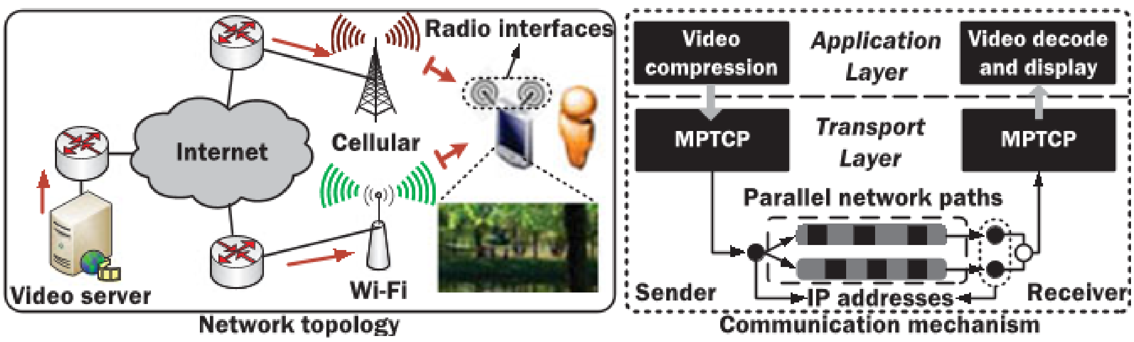

Figure 11. Video broadcasting using MPTCP.

Key goals for MPTCP are: to be deployable and usable without significant changes to existing mobile network infrastructure; to be usable by unmodified applications; and to be stable and congestion-safe over the wide range of existing Internet paths. MPTCP assumes that both peers are modified and that one or both peers have multiple addresses, which often results in different network paths that are at least partially divergent (however, note there is no guarantee that the paths are divergent at all).

4.7 QUIC

QUIC (Quick UDP Internet Connection) is a new multiplexed and secure transport atop UDP, designed from the ground up and optimized for HTTP/2 semantics (Figure 12).

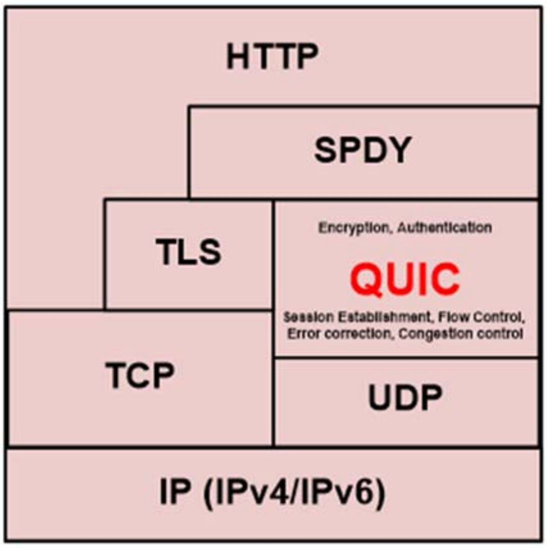

Figure 12. The location of the QUIC in the TCP/IP protocol stack.

While built with HTTP/2 as the primary application protocol, QUIC builds on decades of transport and security experience, and implements mechanisms that make it attractive as a modern general-purpose transport for video content delivery. QUIC provides multiplexing and flow control equivalent to HTTP/2, security equivalent to TLS, and connection semantics, reliability, and congestion control equivalent to TCP.

QUIC aims to be nearly equivalent to an independent TCP connection, but with much reduced latency.

An HTTP/QUIC connection carried over bidirectional unicast is defined as a conversation between two QUIC endpoints that multiplexes several logical streams within a single encryption context. This is a one-to-one relationship. Furthermore, QUIC connections achieve decoupling from the underlying network (IP and port) by means of a Connection ID. Use of a consistent connection identifier allows QUIC connections to survive changes to the network connectivity. The establishment of a QUIC connection relies upon an up-front, in-band exchange (and possible negotiation) of cryptographic and transport parameters (conveyed in QUIC handshake messages) and HTTP-specific settings (conveyed in HTTP/2 "SETTINGS" frames). Such parameters may be required or optional and may be used by either endpoint to control the characteristics of connection usage.

This concept of a connection does not suit the carriage of HTTP/QUIC over unidirectional network associations such as multicast IP. In fact, there is no requirement for either endpoint (multicast sender or receiver) to be in existence in order for the other to start or join this one-sided conversation. The term "connection" is misleading in this context;

therefore, was introduced an alternative term "multicast QUIC session" or simply "session", which is defined as the logical entity describing the characteristics of the anticipated unidirectional flow of metadata and data. Such characteristics are expressed as "session parameters". Advertisement of multicast QUIC sessions allows for the senders and receivers to discover a session and to form multicast IP network associations that permit traffic flow.

5 5G Perspectives and Content Delivery Framework Principles

In this section we present a discussion of the preceding sections of this document and draw conclusions about the broad outline of a content delivery framework to meet future needs.

Our first observation is that the Internet is replacing vertically-integrated delivery platforms. As broadband speeds increase, the Internet is increasingly viable as a commercial content services delivery platform. Further, the rise of global CDN platforms as well as media platforms, such as YouTube, Twitch etc., dramatically reduce the barriers to entry and reduce operating costs for many content producers. These global operators are rapidly adding services to their portfolios that resemble those of traditional broadcast TV and cable operators. It is entirely plausible that such platforms replace traditional broadcast infrastructure at some point in the future.

This growing significance of Internet-delivered media services leads us to conclude that the main role of a content delivery framework within 5G is to deliver this type of content as efficiently as possible. Such content is primarily responses to HTTP requests.

Certain patterns of content consumption pose particular problems. Content which drives sudden increases in network demand make it very difficult to plan capacity effectively and can lead to network infrastructure built around occasional large peaks. The inclusion of a PTM capability as the last leg of the delivery path can help manage network demand by exploiting the synchronous, efficient delivery that such a capability would offer. A key part of our framework is to combine global CDN networks with PTM at the edge of the network. Technologies, such as NFV allow caches to push ever closer to the edge of the network, however, they can do nothing to help with efficient spectrum usage and are unlikely to find their way into fixed access networks. It is on this last leg that PTM will remain very important.

The usage of multicast delivery technologies also allows solving the latency issue raised earlier, by removing the need for buffering important quantity of video on the player side to ensure a smooth quality of experience. The unpredictable bursts at stake in unicast delivery are replaced by a smoother and better controlled delivery in multicast. The 30s to 40s video buffering is hence not required anymore. Of course, an adapted player is required, that will request the appropriate number of chunks based on the user context.

Such an approach could be used to ensure efficient spectrum and last mile usage for linear content, but could also be used for pre-emptive delivery of other content, such as operating system updates or VoD assets if combined with local or on-device storage.

As we have argued earlier in this document and based on previous experience, we should seek to avoid the need for a multicast session to be ‘reserved’ by a content service provider or CDN operator as this complicates both the technical and commercial relationships involved. Instead, the multicast capability should be treated as an internal optimisation capability, dynamically selected according to demand. The result will be that it will be trivial for content service providers to benefit from multicast, as they will not need to make any changes to their normal unicast delivery infrastructure.

It should be made clear that this is very different from the normal way of exploiting PTM networks for linear content delivery. Normally, the whole content delivery chain, including the media encoders, the content protection the transport protocols, the client devices and applications using to play the content will all be different from those used by unicast content delivered over the Internet. By allowing the PTM capability to be a largely hidden, autonomously self-optimizing technology, we dramatically reduce the barrier to exploitation and we improve the infrastructure savings.

The novelty of this approach creates many challenges which will need to be overcome, in particular in relation to the interface between the CDN and the PTM network. For example, the PTM network needs to be inserted into the content delivery path in a way that is similar to caches, and, therefore would encounter the same problems as the transparent cache if we were to attempt to make the PTM network transparent to the CDN operator. A further complication to this interface could be that that content which can be delivered synchronously may need to be explicitly identified, possibly by the CDN operator or CSP.

We recognize that for a PTM technology to become a mainstream over-air distribution capability on mobile networks, it must be supported on handsets. This means that consumers must themselves see a benefit, so that they will see a PTM capability as a differentiator. This in turn will encourage handset vendors to add the technology. This could be primed by a network operator offering services which are exclusively available to those handsets which are capable of receiving PTM transmissions, such as live sport. We do recognize though, that there are efficiency gains to be achieved through the use of PTM in the backhaul network which could encourage an early deployment of such as technology, even without handset support.

The arguments that we have set out here for this unified unicast and PTM approach to content delivery apply equally to the fixed and the mobile network. We believe there is no reason why the functional components, the APIs and the logic cannot be exactly the same, irrespective of the network type. It is in this sense that we describe our approach as 'converged'.

Our observations of failure of network QoS to gain widespread use, outside of a network operator's own services, leads us to conclude that requirements such as session handover and hybrid access will preferentially be dealt with at the application layer with no network signalling. There will be some services whose demands on latency and packet loss are such that explicit network support is required. However, for many content services, seamless handover from one network to another, or the bonding of different access networks, can be handled entirely at the application layer. We therefore believe that application intelligence is preferred over explicit network signalling, despite the possibility that having the network handle handover or bonding could be more efficient.

In summary, our vision for content delivery consists of:

- CDNs for global reach, including the use of NFV for allowing caches to move ever closer to the edge.
- PTM, in the form of multicast or broadcast for efficient synchronous delivery at the edge of the network to improve efficiency of spectrum usage or last mile of the fixed network.
- Treating PTM as an internal optimisation capability for network operators which does not need to be exposed as an externally consumable network service in its own right. The use of a PTM capability should be dynamic, following consumption patterns and business rules.
- The use of local, or on-device storage to allow synchronous, pre-emptive delivery of content efficiently using a PTM capability,
- Application-layer intelligence is preferred over network features, particularly where the requirements on latency and packet loss are relatively relaxed, such as in typical media delivery scenarios.

In a subsequent deliverable (D5.2) we will add details to this approach and suggest ways to overcome some of the many challenges to realise it.

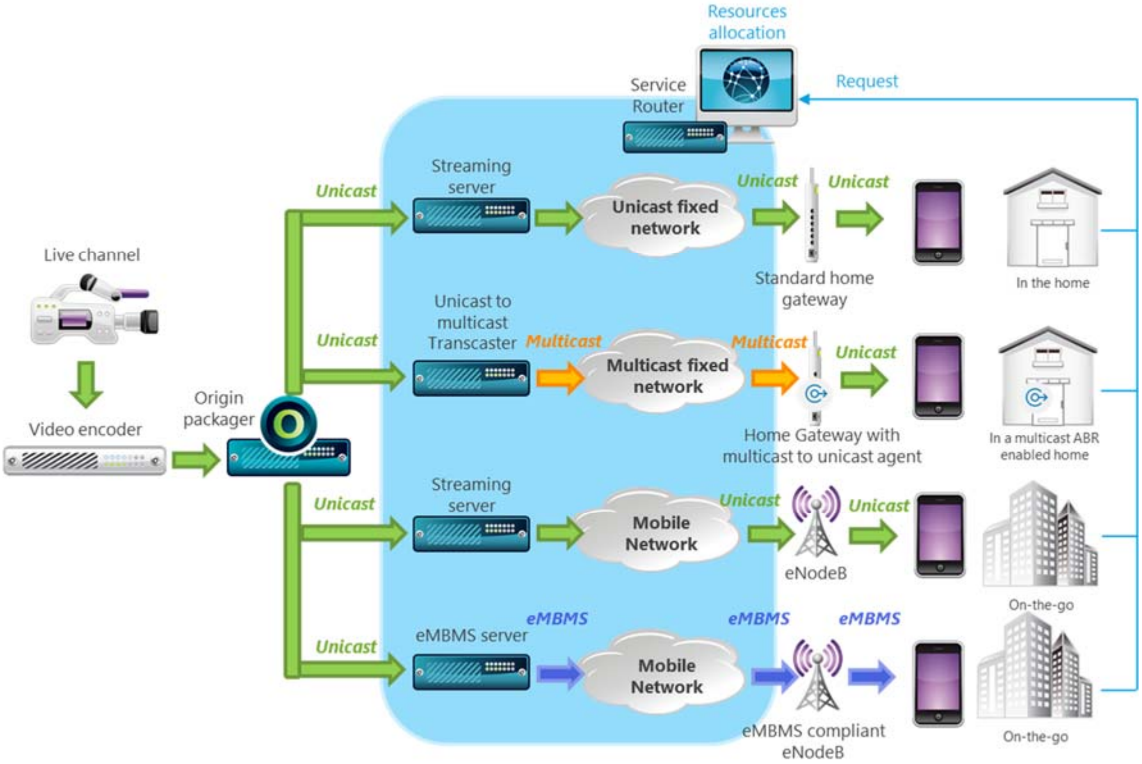

Figure 13. Unicast / Multicast / Broadcast convergent video delivery.

6 References

[1] Cisco, VNI Mobile Forecast Highlights 2016-2021, [online]
[2] Ericsson, "Ericsson Mobility Report," June 2017 [online].
[3] Rani Molla, "Verizon and AT&T customers are getting slower speeds because of unlimited data plans," Recode, Jul 2017 [online]
[4] Chris Welch, "Netflix speeds on Verizon Wireless appear to be capped for some customers," The Verge, July 2017 [online]
[5] Jon Brodkin, "Verizon offers unlimited data and won't throttle video," Ars Technica, Feb. 2017 [online]
[6] D. Gomez-Barquero (Ed.), "*Next Generation Mobile Broadcasting*," CRC Press, March 2013.
[7] S. Mason, "Mobile TV – Results from the DVB-H Trial in Oxford," EBU Technical Review, April 2006 [online].
[8] C. Sattler, "Mobile Broadcast Business Models – Generic Business Models and Country-specific Implementations," BMCO forum white paper, Sept. 2008 [online].
[9] C. Sattler, "Best Practice Regulatory Framework for Mobile TV," BMCO forum white paper, June 2008.
[10] L. Goode, "*Netflix CEO says he isn't worried that Trump administration will gut net neutrality*," The Verge, March 2016 [online]
[11] L. Downes, "*The Tangled Web of Net Neutrality and Regulation*," Harvard Business Review, March 2017, [online]
[12] BEREC, "All you need to know about Net Neutrality rules in the EU," [online]
[13] BEREC, "Guidelines on the Implementation by National Regulators of European Net Neutrality Rules," August 2016 [online].
[14] G. Faria, J. A. Henriksson, E. Stare, and P. Talmola, "DVB-H: Digital Broadcast Services to Handheld Devices," *Proceedings of the IEEE*, vol. 94, no. 1, pp. 194-209, Jan. 2006.
[15] M. R. Chari, *et al*., "FLO Physical Layer: An Overview," *IEEE Transactions on Broadcasting*, vol. 52, no. 1, pp. 145-160, March 2007.
[16] G. Bedicks and C. Akamine, "Overview of ISDB-T: One-Segment Reception," Handbook of Mobile Broadcasting: DVB-H, DMB, ISDB-T and MediaFLO, CRC Press, 2008.
[17] S. Cho, *et al*., "System and Services of Terrestrial Digital Multimedia Broadcasting (TDMB)," *IEEE Transactions on Broadcasting*, vol. 52., no. 1, pp. 171-178, March 2007.
[18] R. Pearce, "*Telstra bets big on LTE-Broadcast*", Computerworld, Feb. 2017 [online]; and Expway press release, "*Telstra and Expway Partner to Demonstrate Intelligent Switching of LTE Broadcast and Unicast On-Demand*," [online]
[19] ATSC, "ATSC 3.0 Standards," [online]
[20] IETF RFC 6726, "FLUTE - File Delivery over Unidirectional Transport," [online]
[21] G. K. Walker, T. Stockhammer, G. Mandyam, Y.-K. Wang, and C. Lo, "*ROUTE/DASH IP Streaming-Based System for Delivery of Broadcast, Broadband, and Hybrid Services*," IEEE Transactions on Broadcasting, vol. 62, no. 1, pp. 328-337, March 2017.
[22] DASH-IF, "Guidelines for Implementation: DASH-IF Interoperability Points for ATSC3.0", Jan 2016, [online]
[23] ETSI GS MEC 003 v1.1.1, "Mobile Edge Computing (MEC); Framework and Reference Architecture," March 2016 [online].