Video Content Delivery over the Internet

Erik Kosonen

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 23.5.2016

Thesis supervisor:

Prof. Raimo Kantola

Thesis advisor:

M.Sc. Tomi Sarajisto

Aalto University
School of Electrical Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Erik Kosonen

Title: Video Content Delivery over the Internet

Date: 23.5.2016 Language: English Number of pages: 9+49

Department of Communications and Networking

Professorship: Networking Technology

Supervisor: Prof. Raimo Kantola

Advisor: M.Sc. Tomi Sarajisto

Popularity of multimedia streaming services has created great demand for reliable and effective content delivery over unreliable networks, such as the Internet. Currently, a significant part of the Internet data traffic is generated by video streaming applications. The multimedia streaming services are often bandwidth-heavy and are prone to delays or any other varying network conditions. In order to address high demands of real-time multimedia streaming applications, specialized solutions called content delivery networks, have emerged. A content delivery network consists of many geographically distributed replica servers, often deployed close to the end-users.

This study consists of two parts and a set of interviews. First part explores development of video technologies and their relation to network bandwidth requirements. Second part proceeds to present the content delivery mechanisms related to video distribution over the Internet. Lastly, the interviews of selected experts was used to gain more relevant and realistic insights for two first parts.

The results offer a wide overview of content delivery related findings ranging from streaming techniques to quality of experience. How the video related development progress would affect the future networks and what kind of content delivery models are mostly used in the modern Internet.

Keywords: Video, Content Delivery, Internet, Streaming, Video-on-Demand

AALTO-YLIOPISTO
SÄHKÖTEKNIIKAN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Erik Kosonen

Työn nimi: Videosisällön jakelu Internetin välityksellä

Päivämäärä: 23.5.2016 Kieli: Englanti Sivumäärä: 9+49

Tietoverkkotekniikanlaitos

Professuuri: Tietoverkkotekniikka

Työn valvoja: Prof. Raimo Kantola

Työn ohjaaja: Tietojenkäsittelytieteen FM Tomi Sarajisto

Multimediapalveluiden suosio on noussut huomattavasti viime vuosina. Videoliikenteen osuus kaikesta tiedonsiirrosta Internetissä on kasvanut merkittävästi. Tämä on luonut suuren tarpeen luotettaville ja tehokkaille videosisällön siirtämisen keinoille epäluotettavien verkkojen yli. Videon suoratoistopalvelut ovat herkkiä verkossa tapahtuville häiriöille ja lisäksi ne vaativat usein verkolta paljon tiedonsiirtokapasiteettia. Ratkaistaakseen multimedian reaaliaikaisen tiedonsiirron vaatimukset on kehitetty sisällönsiirtoon erikoistuneita verkkoja (eng. content deliver network - CDN). Nämä sisällönjakoon erikoistuneet verkot ovat fyysisesti hajautettuja kokonaisuuksia. Yleensä ne sijoitetaan mahdollimman lähelle kohdekäyttäjäryhmää.

Tämä työ koostuu kahdesta osasta ja asiantuntijahaastatteluista. Ensimmäinen osa keskittyy taustatietojen keräämiseen, videotekniikoiden kehitykseen ja sen siirtoon liittyviin haasteisiin. Toinen osa esittelee sisällönjaon toiminnot liittyen suoratoistopalveluiden toteukseen. Haastatteluiden tarkoitus on tuoda esille asiantuntijoiden näkemyksiä kirjallisuuskatsauksen tueksi.

Tulokset tarjoavat laajan katsauksen suoratoistopalveluiden sisällönjakotekniikoista, aina videon kehityksestä palvelun käyttökokemukseen saakka. Miten videon kuvanlaadun ja pakkaamisen kehitys voisi vaikuttaa tulevien verkkoteknologioiden kehitykseen Internet-pohjaisesssa sisällönjakelussa.

Avainsanat: Video, Sisällönjako, Internet, Suoratoisto, Tilausvideo

Preface

This Master's thesis has been written as partial fulfillment for the Master of Science degree in Aalto University, school of Electrical and Communications Engineering. My thesis was done at TeliaSonera Finland Corporation alongside of my work as IP-network Designer at Design & Implementation, Enterprise Services -unit.

I would like to thank my supervisor Professor Raimo Kantola and my instructor Tomi Sarajisto for their guidance. Both provided me with inspiring support and advises which proved to be very valuable throughout my entire thesis writing process.

I wish to express my gratitude to my family and friends as well, for their support and patience throughout my studies.

Otaniemi, 23.5.2016

Erik A. O. Kosonen

Table of Contents

List of Figures

List of Tables

Terminology

Abbreviations

4G	4th Generation Mobile Network
5G	5th Generation Mobile Network
AVC	Advanced Video Coding
CAPEX	Capital Expenses
CDN	Content Delivery Network
DNS	Domain Name System
DSL	Digital Subscriber Line
DTT	Digital Terrestrial Television
DVB-C	Digital Video Broadcasting, Cable Networks
DVB-T	Digital Video Broadcasting, Terrestrial Networks
DVB-S	Digital Video Broadcasting, Satellite Networks
FTP	File Transfer Protocol
FHD	Full High Definition
HD	High Definition
HEVC	High Efficiency Video Encoding
HTTP	Hypertext Transfer Protocol
ICMP	Internet Control Message Protocol
IETF	Internet Engineering Task Force
IP	Internet Protocol
IPTV	Internet Protocol Television
ISP	Internet Service Provider
OPEX	Operational Expenses
OTT	Over-the-Top
PoP	Point of Presence
RTP	Real Time Protocol
RTSP	Real Time Streaming Protocol
RTT	Round Trip Time
SD	Standard Definition
SLA	Service Level Agreement
SSH	Secure Shell
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
UHD	Ultra High Definition
VM	Virtual Machine
VOD	Video-on-Demand

1 Introduction

The Internet has become a solid platform for information exchange applications such as media content streaming, Internet protocol television (IPTV), large file downloading, network gaming, social networking and high definition (HD) television [1][2]. Interactive information sharing services are part of our everyday lives and all these applications are generating high amount of data traffic and have high demand for Quality of Service (QoS), which bring great challenges to current best-effort model of the Internet. Implementing these resource-hungry applications on top of the Internet Protocol (IP) layer cost-efficiently on a large scale has become a core challenge as the fundamental ideas of the Internet are simplicity and scalability.

Each day, there are more and more devices connected to the Internet, what in turn results in increasing amount of generated data traffic over heterogeneous access networks. This creates even bigger strain on the network resources. According to Cisco Inc. annual *Visual Networking Index* (VNI) forecast report, dated 27th May 2015 [3], the consumer Internet traffic would grow more than three fold over a 5 year span with overall estimated compound annual growth rate (CAGR) of 27%. The report includes traffic generated by both mobile and fixed network consumers. Similarly, the traffic generated by online Internet video, either streamed or downloaded, will grow with an estimated CAGR rate of 33% each year from 2014 to 2019. More details are shown in Table 1. The aim of this research is to investigate and give an overview of technological solutions that address growing demand for network capacity in the near future.

1.1 Research Problem

The research focuses on two core questions:

1. How multimedia content is distributed over the Internet?
2. What are the business relations in the content distribution ecosystem?

The purpose of the first question is to understand how multimedia content is actually distributed over the Internet and what technologies lie behind the process. Additionally, the second question aims to unveil the business relations and roles of involved stakeholders in the video content distribution ecosystem.

Finally there are three secondary questions, which by design would help to form a better understanding on the selected video content delivery topic, and to find the actual motivations behind the technologies revealed by the two primary questions.

- What are the usage trends in the Internet, especially in data traffic?
- How the Internet is interconnected?
- What is the role of content management and distribution rights?

1.2 Research Scope and Objectives

In this work, the big picture is formed from the technical perspective of the multimedia content delivery. Although, the research topic covered in this work is technology oriented, ecosystem business relations were also inspected. Thus the main focus of the thesis is in multimedia distribution techniques, especially in those which are related to video content. Distribution of such content would inevitably raise questions about content management and licensing. However because the in-depth analysis of digital rights management is out of the research scope, only a basic model is presented.

The recent trends in the Internet video content consumption and improvements in video quality are driving the development of more efficient ways to deliver heavy multimedia content to the end-users over the Internet. If these are not addressed correctly, the Internet Protocol (IP) networks would face serious congestion and general performance issues in the near future. The main goal of these technologies is to find a solution to the rapidly increasing demand for network bandwidth and define how these technologies actually work.

The main objectives are narrowed down to identifying key trends in video content consumption over the Internet, the content delivery ecosystem stakeholders and their respective roles in it. Additionally, the content delivery mechanics are selected for closer inspection. Also, a video technology is shortly presented in this work, since it plays a key role in the multimedia distribution ecosystem development.

1.3 Research Methods

The research is divided into three parts. First, a *literature review* is carried out to unveil underlying technologies and value networks behind the *video content delivery* topic.

Second, series of *interviews* with the selected experts are conducted, where each expert was chosen from the *video content delivery ecosystem*. Collected data is analyzed by comparing with the theories and data presented in the literature review. The value networks are generated based on the *value network configuration* method proposed in a research conducted by Casey et al. [4]. This enables better overview of current matters and what the future technological development directions might be.

Last, key findings and conclusions are provided based on the gathered information summarizing the research topic of this work.

1.4 Structure of Thesis

This work consists of several chapters, each logically segmented according to the division of used methods. Overall research process layout is presented in Figure 1.

The study begins with an introduction of the problem, scope and used methods and proceeds to literature overview part. Chapter 2 is designed to unveil background concepts, stakeholders and primitives behind the video content delivery platform, the Internet. Next, Chapter 3 proceeds to video related topics, such as image formats, compression and the types of distribution. As a last part of the literature

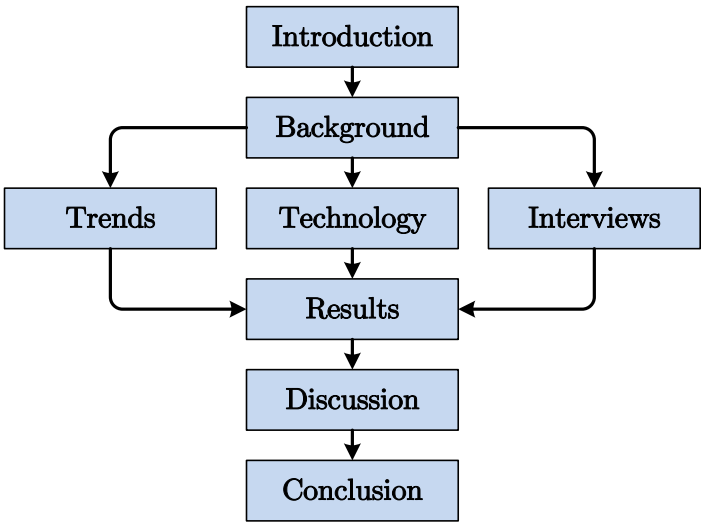

Figure 1: Structure of the Thesis

overview, the most common content delivery models and technologies are presented in Chapter 4.

In the second part, interview methods and results are presented in Chapter 5. All interviewees are summarized into a separate table for clear overview.

Last part is designed to gather the key findings and conclusions of the study in Chapters 6 and 7 respectively. Key findings include the relations of the video content delivery ecosystem and other observations related to the video and content delivery technologies.

2 Background

The purpose of this chapter is to introduce base theory and terminology within the defined scope of the thesis.

2.1 Stakeholders of Content Delivery Ecosystem

Several key roles can be identified in the content delivery market: a content provider, a data center provider, an internet service provider, a content delivery network (CDN) provider, an advertiser and an end-user. There might be even more primary or secondary roles, but in the scope of this work, the ecosystem model and relation of the actors are simplified to comprise the following roles:

A *content provider* (CP) manages and distributes content to the end-users via different content delivery methods. The CP either buys, makes its own content or provides a content distribution platform for actual content creators, for a certain fee or royalty.

A *data center provider* (DCP) offers storage and server capacity services for its customers. Also, in some cases might offer the cloud computing and virtualization services.

An *Internet service provider* (ISP) enables internetworking related services over the global Internet. There are mainly two types of ISPs [5]: an Internet access provider and Internet backbone provider. The former offers Internet access to CPs and end-users. The latter acts as backbone provider and has a significant global network reachability, often classified as Tier 1 ISP. However, to keep things as simple as possible, both are referred as one entity, the ISP.

A *CDN provider* is an actor who provides and manages content delivery services and infrastructure for CPs, and serves requests generated by the end-user. The CDN provider can either have its own CDN or rent service capacity and infrastructure from data center providers.

An *advertiser* is an alternative revenue source for the CPs, while the end-user is considered as the primary source. Sometimes the multimedia services, such as Google's YouTube, are free for the end-users, but then the advertisements are embedded into the content by the CPs for a compensation paid by the advertisers. However, the advertisers are not considered in the value network scenarios for simplicity.

As the last but not least, actor in content delivery ecosystem is the *end-user*, who buys services from CP and generates requests to retrieve multimedia content from a CDN node.

2.2 Defining Quality of Service and Quality of Experience

According to ITU-T recommendation E.800 [6], the *Quality of Service* (QoS) defines how well a service can satisfy the needs and requirements stated by the user of the service. For example the quality of VoIP service is measured by delay, jitter (a variation in end-to-end delay), bandwidth, and reliability.

What comes to the *Quality of Experience* (QoE), the definition of QoE may vary depending on the source. European Telecommunications Standards Institute (ETSI) defines QoE as "*a measure of user performance based on both objective and subjective psychological measures of using an ICT service or product*" [7]. However International Telecommunications Union (ITU-T) defines QoE with slightly different words: "*The overall acceptability of an application or service, as perceived subjectively by the end-user.*"[8].

Both ITU-T and ETSI include additional notes stating that the QoE is not a single measurable quality, but rather it is a set of distinctive end-to-end system effects, which may be influenced by the user expectations and the context [7][8]. When these qualities are combined and viewed as one big picture, we get the actual QoE. The QoE consists of at least three factors: the *user*, used *system* and *context of use*. Each user has his own distinctive opinions and experience of the surrounding environment. The system can be any product or service [7], which the user can utilize, such as mobile phone. Finally, the actual context where the system is used may add some variation to the final outcome, for example calm environment vs. stressful event with limited time.

According to the research conducted by Ericsson Ltd. and Neurons Inc., a smart-phone user would experience a significant increase in stress levels during video streaming over a network. One six-second delay resulted in a smaller stress level increase, but any additional delay after the first one would put a user through a similar strain as a horror movie or a complex mathematical task would do, as illustrated in Figure 2, especially when performing under time pressure [9].

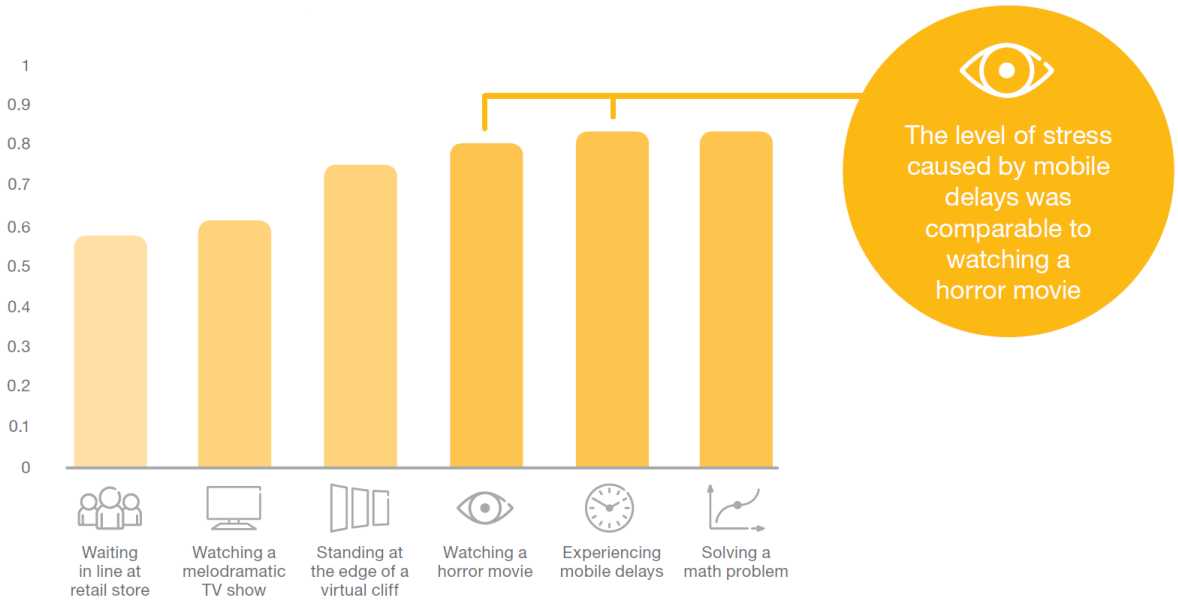

Figure 2: A comparison of stress levels in various situations

2.3 Trends in the Telecommunications Sector

The Internet has grown and expanded significantly during the past decade [3]. The amount of the Internet IP-traffic has been increasing nearly at exponential rate. By definition the Internet traffic is all Internet Protocol (IP) traffic that crosses an Internet backbone. The main reason for such growth has been the proliferation and increased popularity of video-on-demand (VoD) streaming services. A *VoD* system is a service, where video content can be watched by a user at any requested time.

According to Cisco Inc. VNI forecast shown in Table 1, the consumer generated VoD streaming traffic is expected to more than double by 2019, and the majority of this traffic would be *High Definition* (HD) quality video content, which is not even the highest quality level available. Recently, the *Ultra High Definition* (UHD), or commonly known as the 4K video content has been launched by Netflix and more such services are going to follow. Multimedia streaming is already accomodating more than 70% of the Internet traffic during the peak hours in North American fixed-access networks [10], where the top two sources of significant video traffic are Netflix and Google's YouTube streaming services.

Table 1: Total Estimated Petabytes (PB) per Month 2014-2019

Year	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**CAGR**
Internet Traffic	33 595	41 338	52 110	67 021	86 520	111 592	27%
Internet Video	21 624	27 466	36 456	49 068	66 179	89 319	33%

In addition to the aforementioned facts, there is another tool for web-based estimation created by Ericsson Ltd. for drawing graphs of IP-traffic growth [11]. The data can be sorted by the source application as shown in Figure 3, where a noticeable exponential growth in IP-traffic can be observed, especially in traffic generated by various video streaming applications.

According to Streaming Media magazine [12], there has been a clear decline in linear broadcasting services and the trend is moving toward live and VoD streaming over the Internet. As a new trend, the linear television has been steadily shifted towards VoD streaming services during the recent years. Live gaming and electronic sports (e-sports) streaming services, such as Twitch, have been gaining increasing popularity among the younger generations. Additionally virtual reality (VR) is emerging as new market for video applications, along with new (live) video sharing applications. Sharing of live video material via Periscope application between peers has become a trend as well. All these applications are expected to generate intensive bandwidth consumption of the underlying network infrastructure.

2.4 The Internet Protocol Suite

There are two well-known reference models for networking technologies, which are used to break down communication system entities into smaller manageable abstraction layers. These abstraction layers are designed to serve as clear boundaries for protocols and function responsibilities.

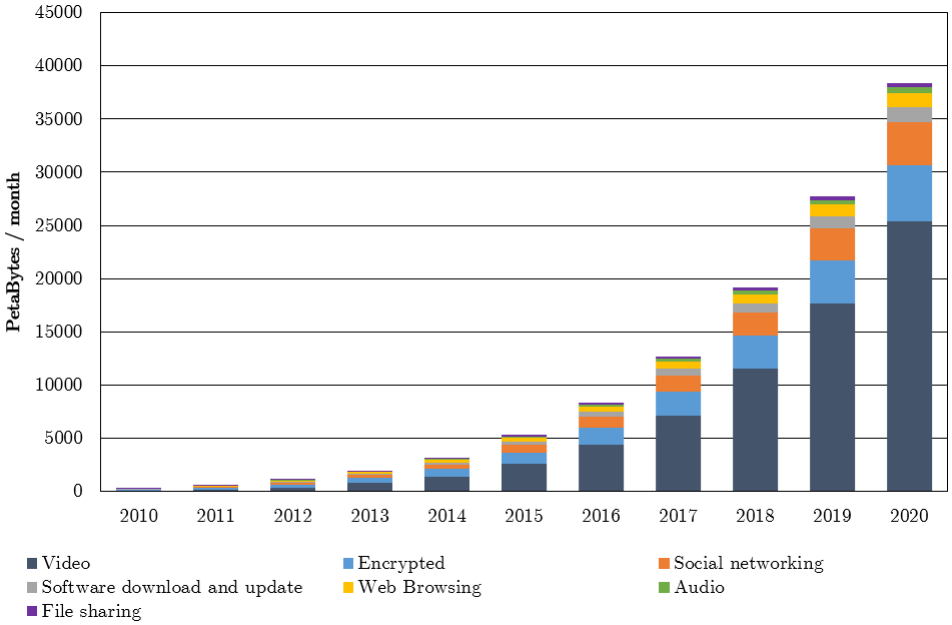

Figure 3: Estimation of IP-traffic generated by various applications (Ericsson)

One of the reference models is the open systems interconnection (OSI) model defined by ITU-T recommendation X.200 [13], where a communication system is partitioned into several abstraction layers as shown in Figure 4.

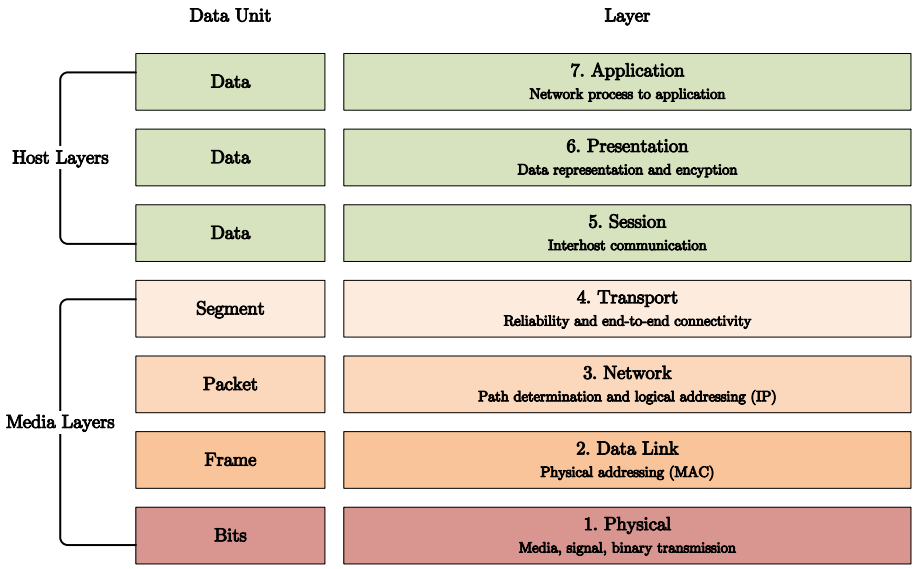

Figure 4: An illustration of the OSI reference model

Similarly, there is another more simplified and maybe more modernized reference model specially adopted for the current Internet architecture. The Internet Engineering Task Force (IETF) defines the Internet reference model, also known as TCP/IP model depicted in Figure 5, which combines one or more OSI layers with similar

functions into one entity [14]. There might be slight differences in terminology, but in general the core idea of both reference models is identical. Also the abstraction layers of the Internet reference model are defined more loosely.

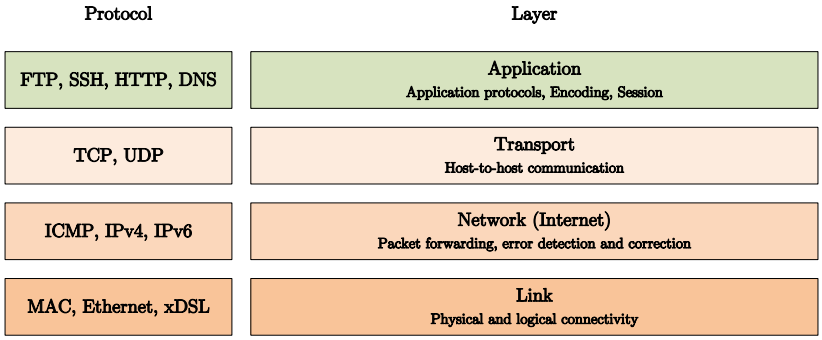

Figure 5: The Internet reference model by IETF

Both of these models are very important when discussing the structure of the Internet, since otherwise designing new services or protocols would require tremendous amount of effort. Defining clear responsibility boundaries makes designing of services more simple.

2.5 The Internet as a Platform

The Internet can be described as a network of networks, where transit peering and service agreements are used to define interconnection relations between involved parties [5]. This can be also referred to as a *global Internet peering ecosystem*. The global Internet peering ecosystem can be further segmented into a set of Internet regions, which can be defined approximately by the country borders. These Internet regions are operating a smaller Internet peering ecosystem within given boundaries.

The terms *transit* and *peering* are used to describe the type of provided connectivity service (Figure 6) in the Internet ecosystem. Transit agreement refers to a simple customer-supplier business relationship, where a customer buys access to all networks (routes) known by the ISP. In other words, transit is a gateway towards the Internet and money flows upstream as shown in Figure 6.

Peering is an arrangement, where two parties agree to share access to their networks for mutual benefit on more or less equal terms. This also includes access to any customer networks that the involved peering parties may have. It has to be noted that peering is a non-transitive relationship. Any networks learned via transit agreements are excluded. Typically, peering does not involve any fees for sharing networks with the involved parties. However, any inequality in a peering relationship might lead to a paid peering, where a compensation of some form is agreed upon. For example, if one party generates more traffic than the other party does, then the peering agreement has to be redefined.

In the modern Internet ecosystem, there are three distinguishable entities [15]. These can be classified as Tier 1, Tier 2 and content provider networks. Two first types

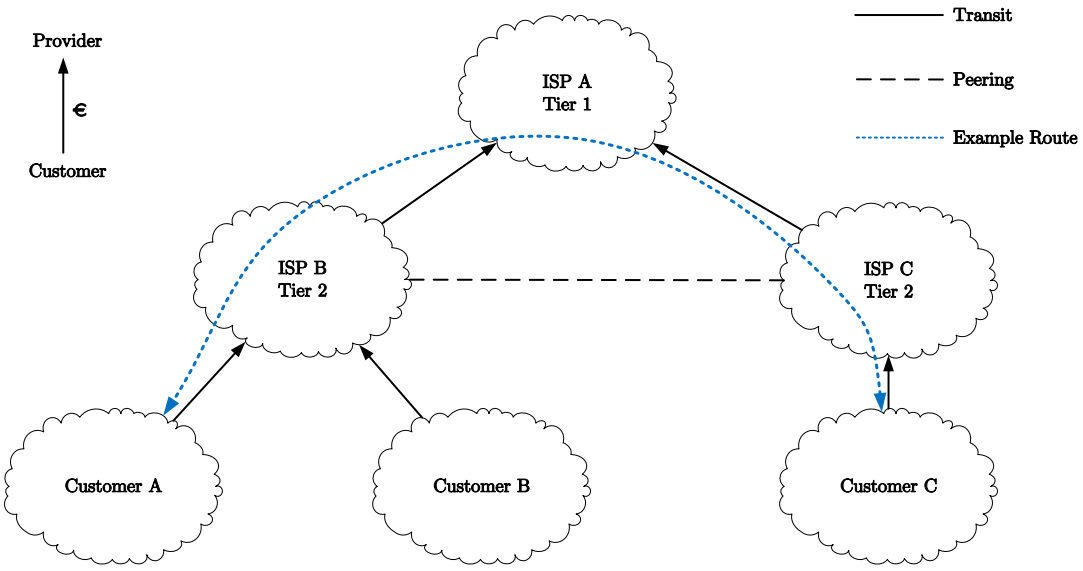

Figure 6: Simplified illustration of Internet structure relations

are often operated by an ISP, and the latter is an enterprise operated network, which sometimes can be referred to as a *Tier 3* network. There is no central authority that would define tiers or types of networks participating in the global Internet. However, the most common definition of a *Tier 1* network is that it can reach every network on the Internet via peering and without purchasing any transit services. Thus Tier 1 ISPs form the Internet backbone. A *Tier 2* network peers with other networks, but is still required to purchase transit to fully reach some portions of the Internet. Therefore, a Tier 2 network also resells transit to other networks. A *content provider* network can be referred to as a customer network that solely purchases transit from other networks to gain an access to the Internet with main focus around content creation and distribution.

Many services in the Internet have been built on top of the Internet protocol (IP) which relies on packet switching technology. In IP-networks the data is encapsulated within a packet, which in turn is transmitted over a network link. Unlike in circuit switched network, there is no end-to-end connectivity and reservation of network resources along the path. The responsibility for correct information routing is given to intermediate network devices, such as routers. The packet switched IP-networks have been designed with best-effort philosophy in mind. Any time sensitive application tends to struggle in the IP-network if proper measures are not taken into account. End-to-end delay and variation (jitter) combined with packet loss and re-transmission of packets cause unwanted disturbance for real-time applications, such video and voice services.

2.6 Logical Layers of a Network Topology

Similarly as with the OSI reference or the Internet protocol models, a network can be divided into three logical layers as shown in Figure 7. Each layer has a clearly

defined set of functions: core, distribution and access [16]. This way network design and management become simpler and more efficient. Any network related changes, such as routing information should not reflect and cause any issues on the higher layers of a network.

A *core* network layer, or commonly known as backbone, is a high speed transit area between interconnected network sites of one network service provider. The primary function of a core network is to forward packets as fast as possible towards the correct destination.

A *distribution* network layer, also known as middle-mile, is designed to aggregate traffic and summarize routes. These routes are called *prefixes* and there were over 600,000 prefixes during the writing process of this work [17]. The routers located on this layer must be able to handle a large amount of routes. Networks are connected to each other by distribution layer edge routers in the Internet, which are known as provider edge (PE) routers. Sometimes the distribution layer might be also referred as an *aggregate* network layer by ISPs and other network service providers.

An *access network layer*, or alternatively, last-mile performs network entry control, feeds traffic to higher network layers and provides edge and connectivity services to the end-users. A few examples of access network technologies could be the digital subscriber line (DSL) and mobile radio access network such as 4G.

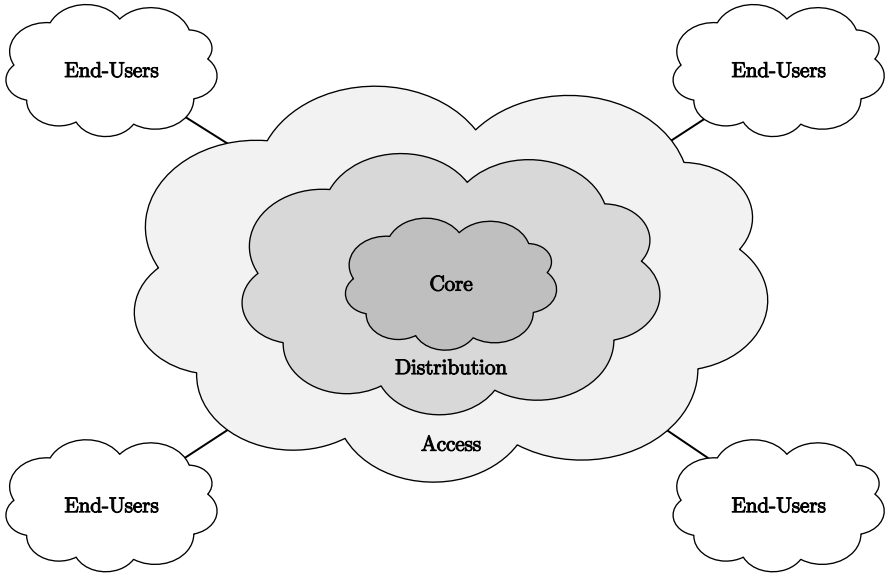

Figure 7: Illustration of topological segmentation of a network.

2.7 Connectivity and Forwarding Primitives

There are four well known transmission models for data packets, each with unique role and purpose. The traffic flow can be categorized into unicast, multicast, broadcast and anycast.

Unicast, also known as *point-to-point*, is a connection between two endpoints, sending host and receiving host [18]. The traffic flow can be both uni- and bi-directional. Each unicast connection is treated as separate, even when receiving the same content from the server as demonstrated in Figure 8, where the traffic is linearly scaling per connection.

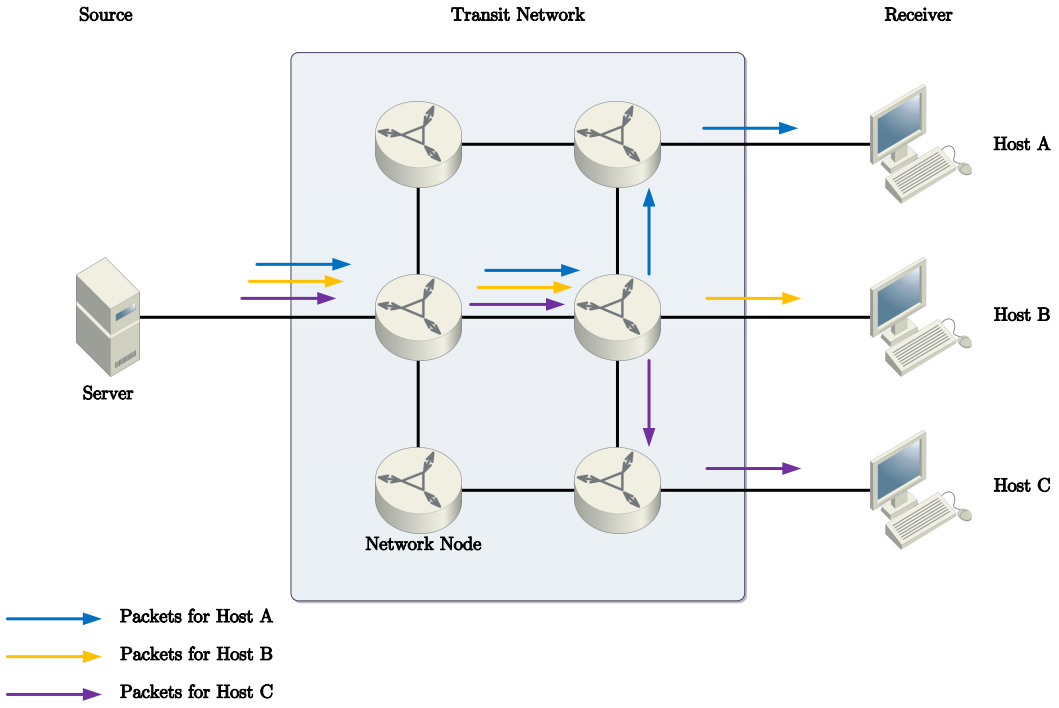

Figure 8: A visualization of unicast (point-to-point) communication.

The term *broadcast* or *point-to-anypoint* describes connectivity between one sending host and multiple receiving hosts. The packets are sent to all nodes on the network and the traffic flow is unidirectional [18]. The drawback of this method is that all hosts have to process the incoming broadcast traffic.

Multicast uses a *point-to-multipoint* connectivity between one ingress point and one or more egress points *of a selected group*. In other words, packets are sent from one or more source hosts and they are transmitted to one or more receiving hosts on different networks [18]. The traffic flow is unidirectional in most cases. The idea of the multicast technique is to deliver the same information to a group of hosts simultaneously. This minimizes the capacity loads on the network links. A good example of this kind of use is video content delivery over a network. See figure 9, where hosts A and B have joined a multicast group. In terms of traffic scalability, this is a lighter solution than unicast. As the downside, the multicast mode works well only within a closed network environment due to lack of access control, since multicast addresses can be used freely by any entity. Also the multicast traffic is not accepted at the network operator peering points, which makes traversing between

two ISPs impossible.

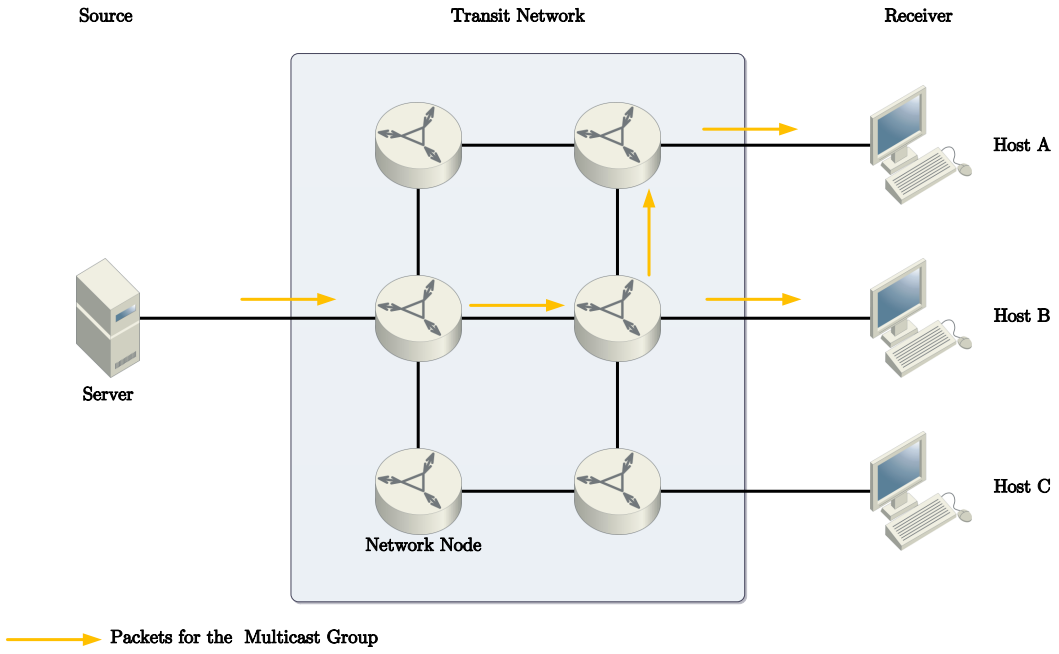

Figure 9: Illustration of multicast (point-to-multipoint) communication.

Anycast is the term used to describe the connectivity between one ingress point and one egress point in a defined set based on certain metrics, such as distance, location and so on. The information is transmitted by the sending host to a single member of a group of potential receiving hosts, which would be identified by the same destination address [19]. Anycast can be also referred to as one-to-nearest association.

3 Video Technology

3.1 Development of Digital Video Formats

A digital image is a series of small elements known as pixels. A video is a continuous sequence of images or frames, which creates an illusion of motion. Standardization of a digital video format addresses how these pixels are presented on a display. There are numerous standards for display size and pixel numbers (resolution). Of all these, four resolutions are chosen for further inspection, as they are de facto standards in television and computer displays. Most of the video services follow this de facto standard providing video formats supporting widescreen displays with 16:9 aspect ratio. The *aspect ratio* of a display image defines height-to-width proportion of pixels. The screen resolutions used in this work are presented in Table 2.

Table 2: Widescreen Digital Video Formats (16:9 ratio)

Format	Shortname	Width (pixels)	Height (pixels)
Standard Definition	SD	720	567 or 480
Full High Definition	FHD	1080	1920
Ultra High Definition	UHD-1	2160	3840
Ultra High Definition	UHD-2	4320	7680

The ITU.R BT.709 [20] defines resolutions for HD television and the ITU-R BT.2020 [21] defines UHD resolutions, commonly known as UHD-1 at 3840*2160 pixels (4K) and UHD-2 (8K) at 7680*4320 pixels. Figure 10 demonstrates the difference in pixel count of various video formats.

3.2 Compression and Decompression

Before a video can be transmitted efficiently over a network, it has to be compressed. In most cases the compression, and the opposite action decompression, are both computationally resource heavy processes where a video file is packed into a more compact, smaller size file. This not only saves storage space, but also reduces the number of bits to be transmitted. Constant improvements over time in both storage and computational technologies enable and drive opportunities for development of even better quality digital video format standards such UHD-1 (4K) and UHD-2 (8K). However these improvements also have certain drawbacks, since better quality image format often results in a larger file. Storage space requirement for raw uncompressed video files tends to be rather high, let alone the requirement of network bandwidth for streaming purposes. Driven by this, more efficient video compression and decompression algorithms, also known as codecs, are being developed.

ITU-T H.264 or *Advanced Video Coding* (AVC) is widely adopted in various multimedia applications [22]. But since parts of H.264 codec are proprietary and involve licencing fees, Google has developed an open source alternative VP9 [23] for its Youtube service. A further development in video compression techniques

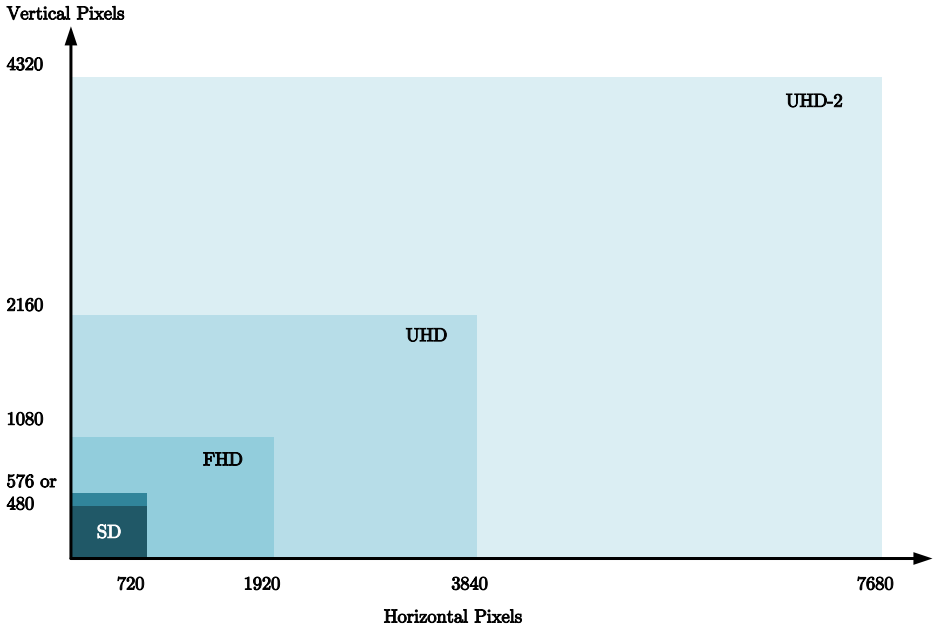

Figure 10: A comparison of digital video formats (16:9 ratio).

enables even more savings in storage and network capacity at the cost of a heavier computational process. H.265, also known as *high efficiency video coding* (HEVC) is designed to replace the H.264 standard in the near future. It is developed by JCT-VC organization, a joint collaboration between the ISO/IEC MPEG and ITU-T video coding experts group (VCEG). Former refers to HEVC as MPEG-H part 2 and latter to H.265 [24]. In this thesis the ITU-T notation is preferred.

The new H.265 video standard has the same fundamental problem concerning the commercial usage as with the H.264 video standard; where licensing fees and royalties to patent owners are involved in a similar way. As a workaround for the issue, a development of a new more efficient open-source VP10 video codec standard has been initiated by Google [25].

There is very little reliable data concerning network bandwidth requirements for Google's VP9 and especially VP10 video standards, unlike in the case of H.264 and H.265 video standards. This is so, even though the H.265 is considered as a rather fresh technology. In this thesis, both H.264 and H.265 are chosen for a closer inspection. Adopted from research conducted by Analysys Mason [26], Figure 11 demonstrates the difference between two well-known video compression algorithms. The quality of streams can be reduced to adopt to a slower connection speeds, which explains the "high" and "low" notation in the Figure 11. An UHD-2 stream endcoded with H.264 and high quality can require a remarkable high bandwidth of 48 Mbit/s, and similarly with lower quality the result is 32 Mbit/s. There is clearly a noticeable progress in video encoding, as with next-generation H.265 codec the bandwidth requirement is reduced approximately to half of it's predecessor.

As comparison to the information stated in Figure 11 [26], the Netflix Help Center article [27] defines the minimum requirements for *stable* Internet access speeds for SD,

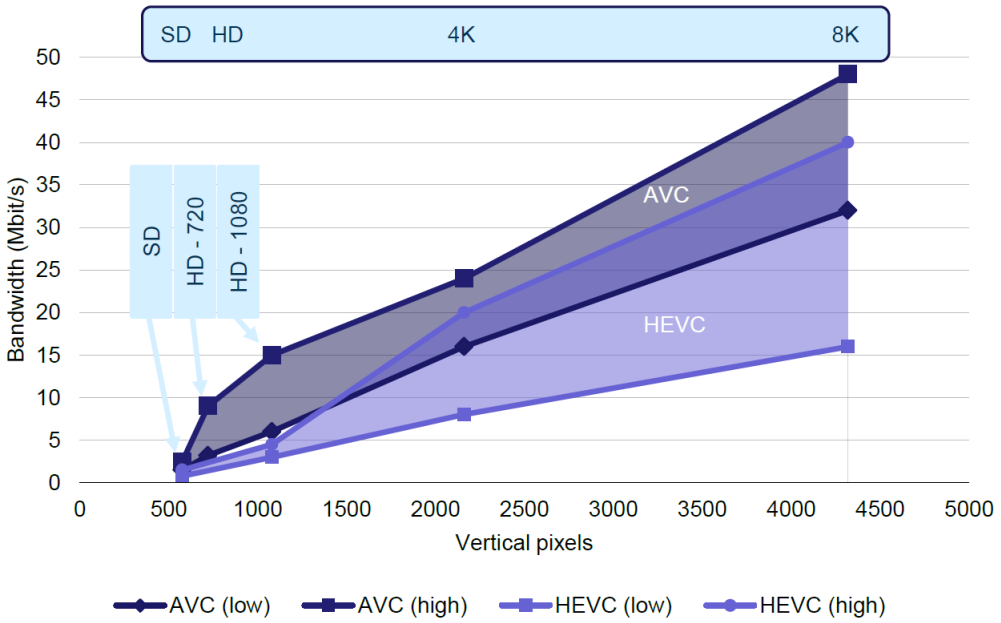

Figure 11: Comparison of AVC (H.264) and HEVC (H.265) encoded video network bandwidth requirements.

HD and UHD as 3 Mbps, 5 Mbps and 25 Mbps, respectively. Unfortunately, there is no statement about any relations to video codecs, thus no clear conclusions can be drawn between streaming bit rates and video codecs. There is a brief description about the H.264 and H.265 video codecs in Netflix tech blog [28], but no further linkage to stream bit rates is provided.

Internet connection bandwidth requirements from Figure 11 are further simplified and summarized in Table 3. These values are used later in this document.

Table 3: A summary of bandwidth requirements for the Internet connection.

Video Quality	**Mbit/s**
SD	3
HD	5
UHD	25
UHD-2	48

3.3 Streaming over an IP-network

The advances in computer, video compression, network and storage technologies have enabled the possibility of near real-time video delivery over the Internet. Video streaming refers to (near) real-time transmission of a live or stored video over a network [29]. There are multiple transmission modes for video content delivery: download, progressive download, streaming and adaptive streaming.

Download mode is simply a process where an entire video file has to be fully delivered before it can be viewed. In a *progressive download* mode parts of a video file may be played out of the buffer during the downloading process. *Streaming* mode behaves in a similar way to progressive download, however the video can be played out almost immediately with minimal buffering in real-time. In other words, the video can be viewed as soon as first bits are received. Lastly, an *adaptive streaming* mode was specially designed to adapt to dynamic network conditions of unmanaged networks, such as the best-effort nature of the Internet demonstrated in Figure 12 [30].

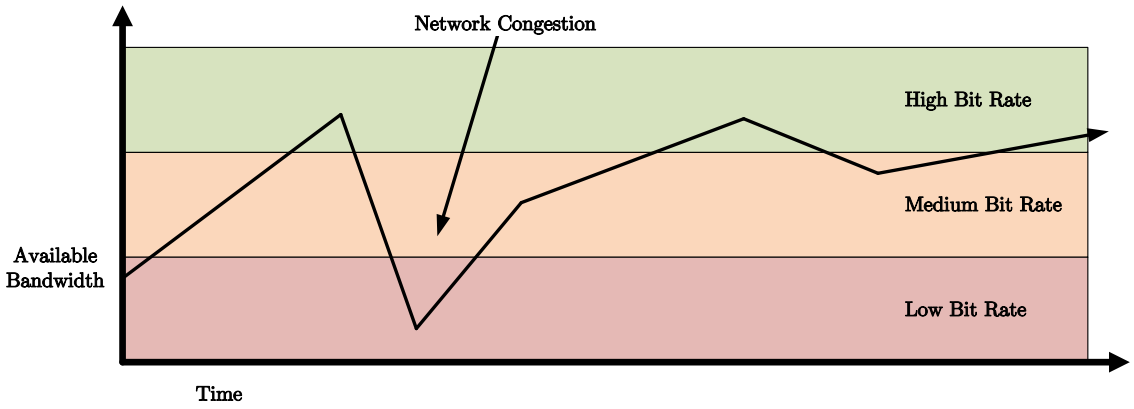

Figure 12: A practical illustration of adaptive streaming.

When comparing streaming mode to progressive download mode, the key difference appears to be at linearity of playback. A user cannot jump ahead without downloading a video file entirely, up until the selected point in the progressive download mode. However, in the streaming mode it is possible to skip directly to a desired point on the video timeline.

A good example of the adaptive streaming is the Netflix video service, where videos are being encoded at scale as shown in Figure 13. This means that a video stream is split into multiple chunks, where each instance of the same chunk is encoded at various bit rates. This way a video stream generated by Netflix video service can adapt to variations in network conditions of per end-user basis.

A video can be streamed by using an application-level protocol, the *hypertext transfer protocol* (HTTP) [31]. Most of the modern streaming services have been adopted to use HTTP for streaming purposes [32]. Originally the HTTP was designed to transmit web page content using the progressive download mode, which unfortunately does not work well for streaming purposes by default. However with the help of Microsoft Smooth Streaming (MSS), Adobe HTTP Dynamic Streaming (HDS), and Apple HTTP Live Streaming (HLS) proprietary media container technologies the HTTP has been modified to support adaptive streaming.

A *media container* is a format used to describe how multimedia data such as digital audio and video are stored within a file. It is also designed to handle partitioning of a file before it can be transmitted over a network.

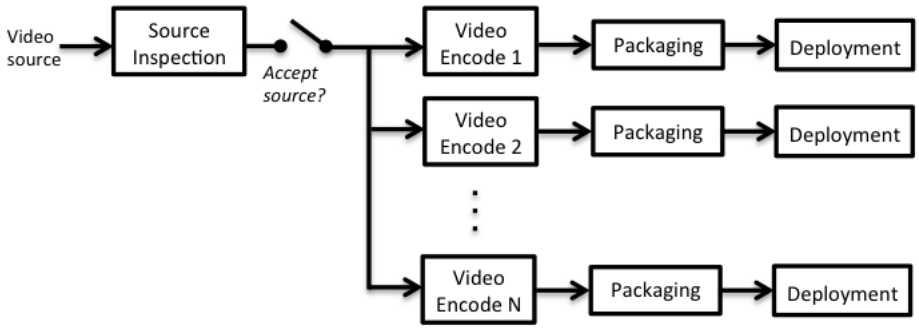

Figure 13: Video pipeline for adaptive streaming demonstrated by Netflix.

There are also speculations about similar royalty-free media container technology known as *Dynamic Adaptive Streaming over HTTP* (MPEG-DASH) ISO Standard ISO/IEC 23009-1 [33], that has been planned to replace most of the aforementioned media container formats. Unfortunately according to MPEG License Administration (MPEG LA) the MPEG-DASH is going to be a licensed technology [34].

3.4 Distribution Process

Video files are stored in a content asset library, usually, owned by a content provider. Then the video content is multiplexed within a broadcasting stream and sent to distributors, who distribute and deliver compressed videos to the end-users. Modern video distribution process involves mainly three types of delivery methods as shown in Figure 14 [26].

Digital terrestrial television (DTT), satellite television and cable television are categorized as traditional broadcast video distribution techniques, which rely on variants of the digital video broadcasting (DVB-X) standards. The 'X' in the term DVB indicates the type of the underlying broadcast medium. However, this thesis explores video distribution over the Internet, thus traditional broadcasting is out of scope of this thesis. Two other delivery methods are designed to operate on top of the Internet architecture, over managed and unmanaged IP-networks.

A managed network is usually operated by a certain entity such as the ISP, which might also provide Internet Protocol Television (IPTV) services. In the managed network, all QoS parameters are actively monitored and enough bandwidth is reserved for video distribution purposes. On the contrary an unmanaged network is not managed by any central entity. The video traffic is streamed based on best-effort model over multiple networks, in other words over the Internet. Therefore, there is no QoS guarantee for the end-users. This type of video service over unmanaged networks is often called as Over-the-Top Television. Good examples of such video service providers are Google's Youtube and Netflix.

There are several distribution models for VoD based content, and these can be classified into three different categories according to payment models: subscription VoD (SVoD) transactional VoD (TVoD) and advertising VoD (AVoD). A perfect

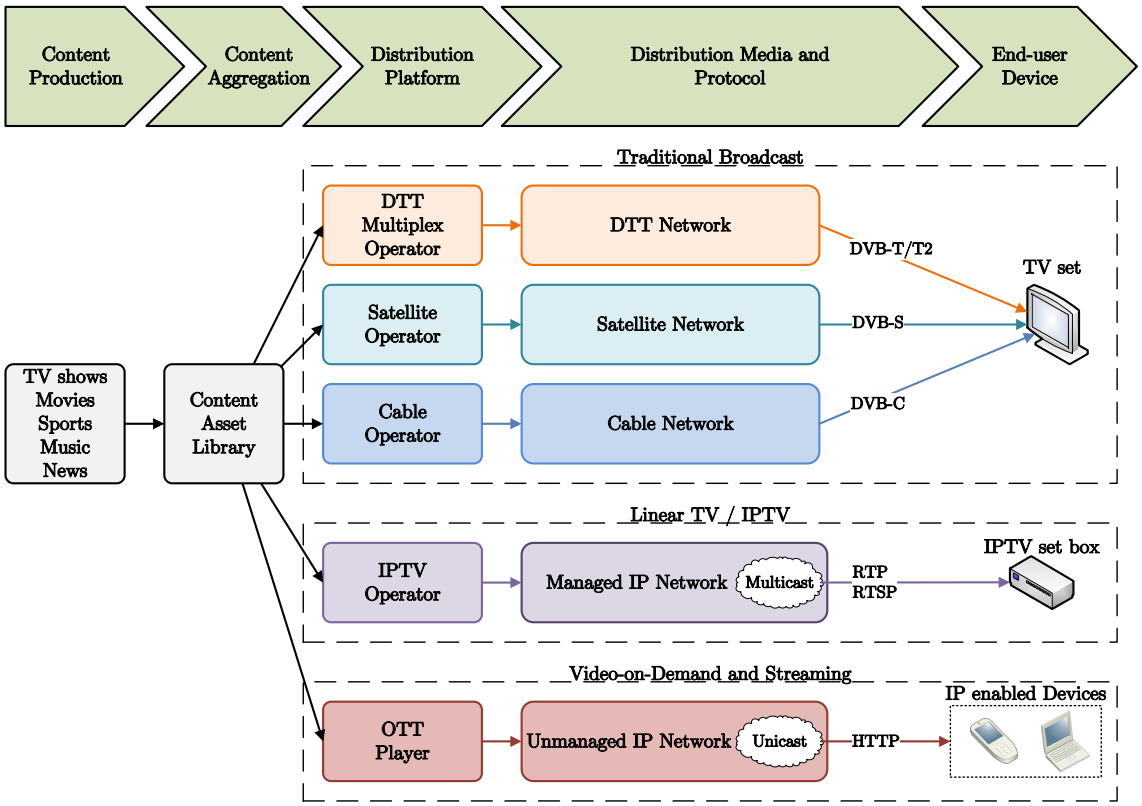

Figure 14: Summary of video distribution process.

example of SVoD would be Netflix, where a user can gain an unlimited access to video service platform against a fee for certain period of time. On the contrary, the TVoD enables pay per use access. The AVoD service is in most cases free for the users, however, advertisements are mixed in with the content. A good example for the AVoD case is the Google's Youtube video service, which is mostly funded by the advertisements embedded into the video stream.

4 Content Delivery Network Technology

The definition "*content delivery network*" seems to constantly pop up when reading and searching material about performance of the Internet and especially delivery of bandwidth intensive (video) content over the Internet. Literature often refers to multimedia distribution technologies over the Internet as a *content delivery network*. There are also some variations such as a*content distribution network*. However, in this work the content delivery network term is preferred, or shortly CDN. The goal of CDN is to minimize or reduce content delivery latency, which is the time taken for requesting device to receive a response and jitter, the unpredictable fluctuations in latency and maximize available network bandwidth. In other words the CDN is an effective measure to improve perceived QoE of the video service.

According to a forecast estimation concluded by Cisco Inc., over half of the Internet video traffic will be served by content delivery networks by 2019 [3]. CDN is designed to deliver often requested content more efficiently from content provider to an end-user, while off-loading data traffic from backbone and core links towards the edges of a network closer to the end-users. The content request can sometimes pile up randomly at certain time of the day or during a popular event causing *Flash Crowd* [35] and *SlashDot Effect* [36] phenomena. Similarly acts a *Distributed Denial of Service* attack, which a CDN can also absorb or mitigate.

A CDN relies on a rather heavy server and router infrastructure, requiring also fast network interconnections. The CDN services rely on application layer protocols [37], which might be slightly modified or improved for certain purposes.

4.1 Terminology

When referring to CDN, *content delivery* is a chain of events triggered by a request, created by an end-user. The *content* itself is persistent or transient digital data resource stored on a server, most often pre-recorded or retrieved from live sources. It consists of two main components, which are the *encoded media* and *metadata* [38]. The encoded media is a stream of encoded static, dynamic or continuous data, anything from audio, video, documents, images and even web pages. Metadata is used to identify, discover and manage multimedia data [38].

Within a CDN, there are three main distinguishable entities, each with their own role: *content provider*, *CDN provider* and *end-user* [39][40]. The first two are not to be confused with each other. The difference between the two is that a content provider is a customer to a CDN provider, much like an end-user is to a content provider. The content provider stores Web objects on an *origin server* and delegates Uniform Resource Locator (URL) names space of these objects to the CDN provider for distribution purposes [40].

The content stored in the origin server is replicated by the CDN provider to geographically distributed *replica servers* [41]. A large concentration of replica servers is called a *Web cluster*. Depending on a source, a replica server can be also referred to as a *cache*, an *edge server* or even a *surrogate* [41] [40]. These are good to know alternatives, however, the term replica server is used through the rest of the document.

Although a *cache* could be one of the replica server functionalities, just as some of reviewed literature agrees with this argument where a replica server can act as cache server additionally to Web server and media server functionalities [39].

Requests generated by end-users are redirected to the most optimal replica server by using certain metrics. These metrics can vary from server with the least load, shortest logical distance or geographical location [41]. A simple example scenario is where an end-user wants to watch a video provided by a Netflix video service: When a Netflix user clicks on a particular video, he or she creates a request to Netflix video service. Then the request is redirected to the nearest replica server [41].

4.2 Brief History

In order to get a broader view of the current situation and how it evolved, this section offers a brief look at the early stages of development. The term *content delivery network* dates back to late 1990s [40], when the first cooperation of distributed servers farms across the Internet appeared. The service was designed to be fully transparent for end-users, meaning, that there would be no visible interactions with it. Also the new design provided increased reliability and scalability which was quickly noticed by content providers. Inspired by these actions, students at MIT began to develop the current solution even further. One of these research projects was addressing *the flash crowd* [35] problem, which later on resulted in Akamai Technologies Inc.[42].

Early attempts to improve Web-page performance were quite simple and straightforward. Whenever a server hosting a certain web-content was having performance issues, the problem was usually solved by upgrading the web server hardware components; such as installing additional memory, better processor, more storage disk space. Also the server network link speeds were scaled up as the load increased. However, these options are only effective for short time spans, not very scalable and they tend to turn out quite expensive on the long run [40]. Later on Internet Service Providers (ISP) started to deploy caching proxy servers for narrow-band network users. This scaled up to a hierarchically chained proxies, that eventually formed server farms.

As the Web-content evolved, it become more and more complex. Serving numerous end-user generated multimedia content requests has become challenging from one location, besides servers at one locations are prone to denial of service (DOS) attacks. Similarly, as with mobile networks or any other technology, a CDN has distinguishable technology life-cycle pattern shown in Figure 15 [40]. The next step in the technological progression of the CDN technology was to address scaling of the content delivery by using distributed methods, which would improve overall perceived user QoE [40].

4.3 Infrastructure Components

The CDN infrastructure can be modeled with four basic components: content delivery, distribution, request routing and accounting [41]. Each of these components has been designed to fulfill certain role and functionality within a CDN as shown in Figure 16.

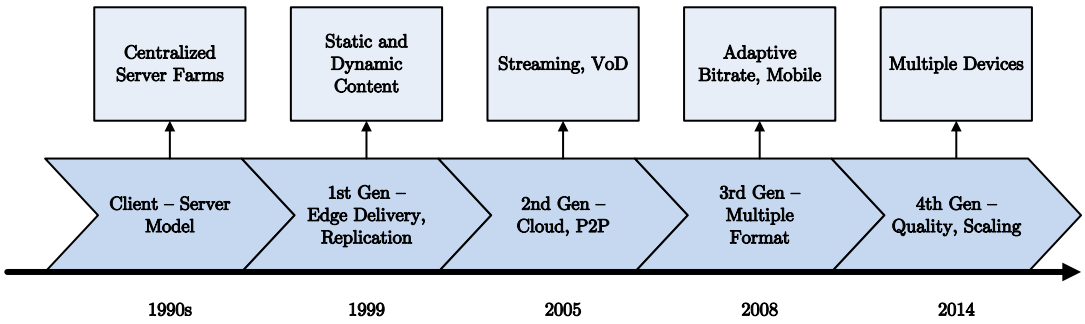

Figure 15: The evolution of a CDN technology

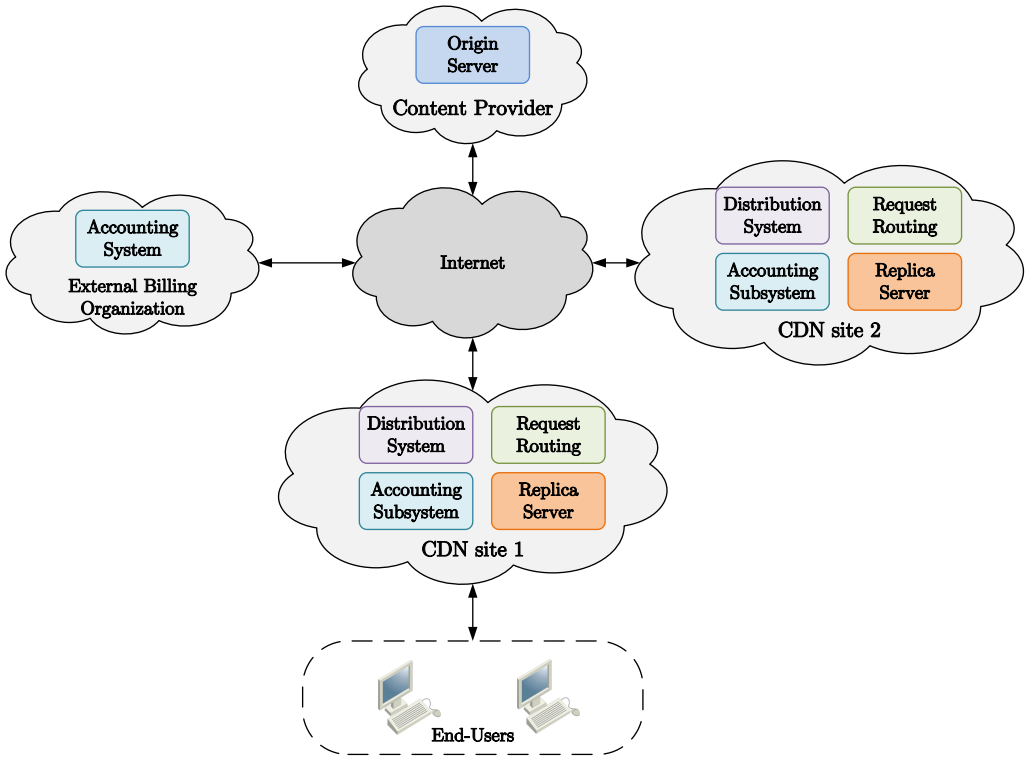

Figure 16: The functional components of a CDN

The *content delivery* involves the origin server and a number of replica servers that are particularly designed to handle content delivery to the end-users. *The distribution system* is responsible for moving content from the origin server to the replica servers, as well for maintaining the integrity of stored cache at these nodes. The *request routing system* interacts with end-users by redirecting requests to appropriate replica servers and also communicates with the distribution system in order to maintain updated view of the content stored at CDN nodes. The purpose of the *accounting system* is to control client authentication and logging functions that are used to measure the CDN usage levels for reporting and billing purposes. The billing systems can be either internal within a CDN or external third party system. In the Figure

16 the billing system has been drawn as an external third party component.

4.4 Request-Routing

In a CDN the available content has to be served efficiently to end-users on request. In most cases each incoming request is forwarded to a replica server geographically closest to the end-user [39]. Sometimes however, the closest replica server might be chosen based on different metrics such as server and network load levels or smallest hop count [43]. In computer networking, the *hop count* is used to describe the number of traversed intermediate device interfaces along the chosen path towards the destination network or host by the packet. Larger hop count indicates that a packet has to pass many intermediate devices, which in turn would result in increased delay due to processing and forwarding of the sent packet.

There are multiple implementation methods for request-routing mechanisms used within CDNs. Request-routing is also commonly known as either content routing or content redirection, however the term request routing is preferred in this document. Many of these techniques are listed in RFC 3568 [43]. Request-routing mechanisms can be categorized based on IETF reference model layers (earlier presented in Section 2.4) starting from top: Application-layer request routing and Transport-layer request-routing.

Most of the request-routing happens on the *Application-layer* in a modern CDN [2], where systems provide finer per object request-routing due to deeper inspection of end-user requests allowing more precise controls over the request-routing process. Application-layer request-routing includes mainly three techniques, which are DNS request-routing, HTTP redirection and URL rewriting.

The *DNS request-routing* is relying on the universality of the Internet DNS system, which makes DNS based request-routing solutions very common. In such solution a DNS server can be specialized to handle the DNS resolution process and the request-routing at the same time. The DNS request-routing is used at least by two large-scale but radically different CDN providers, namely Akamai Inc. and Limelight Networks [2]. The former relies on a hierarchical DNS routing solution and the latter employs IP anycast assisted DNS routing.

In *hierarchical DNS* approach, there are multiple levels of DNS servers. Top level DNS servers are used to process incoming user generated requests and pass those deeper into the CDN infrastructure towards the DNS server closest to the end-user. These second level DNS servers are responding directly to end-users. Unlike Akamai Inc., the Limelight Networks DNS infrastructure is designed to utilize *IP anycast based DNS*. The idea behind it is quite simple. One IP address is mapped to multiple locations, fastest location to respond will be serving the queries. It does not matter from which geographic location the DNS queries are made, the reply seems always to originate from the same IP address.

The *HTTP redirection*, also known as URL redirection, is a process where request-routing decision is based on the contents of a HTTP packet, where the address of the requested object is described as the URL. The forwarding is done based on decisions by the request-routing system, which redirects HTTP packets to a better

replica server in a CDN system. This method is similar to HTTP proxy, where an intermediate server is used to forward HTTP packets to correct destinations based on various parameters.

Similarly to aforementioned HTTP redirection, the *URL rewriting* request-routing method is designed to go a step further in HTTP packet inspection. Modifications may incur to the contents of an URL contained within the HTTP packet. The idea behind this method is that the content provider server may communicate with the end-user device to provide a better replica server based on selected parameters. URL rewriting enables better performance and more scalability when delivering dynamic content, but it has been deemed as a rather overhead heavy method for request-routing [1].

In transport-layer request-routing approach, the request redirection can be achieved by various network traffic engineering methods listed in RFC 3272 [44]. However, these methods are out of scope of this research. Generally describing, the requests generated by the user are inspected on packet level, where the redirection decision is made based on IP address of the user (source), port and transport layer protocol. Transport-layer request routing may include heavier processing overhead and may suite better situations where protocol sessions are long-lived, for example RTSP [45].

Determining the best replica server can be achieved by various *metrics*. These metrics can be categorized into *passive* or *active* measurements. The difference between active and passive measurements is that in active mode a probe can be sent periodically via a network link in order to measure for example latency. In passive mode, the measurements are done out-of-band by inspecting packets that are traversing the network link. The most common used metrics are *latency* between source and destination network device, *packet loss*, *hop count* or even information provided by a network routing protocol.

4.5 Use Cases

There are many use cases for CDN technology, but most notable of them are various multimedia content distribution situations [40]. These use cases include, but are not limited to; web page hosting, audio and video streaming services, electronic services such as e-docs and e-commerce, e-learning and many more as shown in Figure 17. Also a CDN can be used to enhance security by absorbing DDoS attacks and limiting the effects caused by Flash Crowd phenomenon [46].

There are mainly two types of content, static and dynamic. The *static* content does not change often and it does not require additional processing, such as images web page code and etc. Unlike the *dynamic* content, which might change during delivery process because it is generated on the fly by the content server. Generally, a CDN is used to bring static content closer to end-users and accelerate dynamic content. However, with the recent trends, content *streaming* has emerged as new third content type, where video and audio content are played via a web browser in real-time. A perfect example of such streaming case would be Netflix and Google's Youtube web browser based video streaming services.

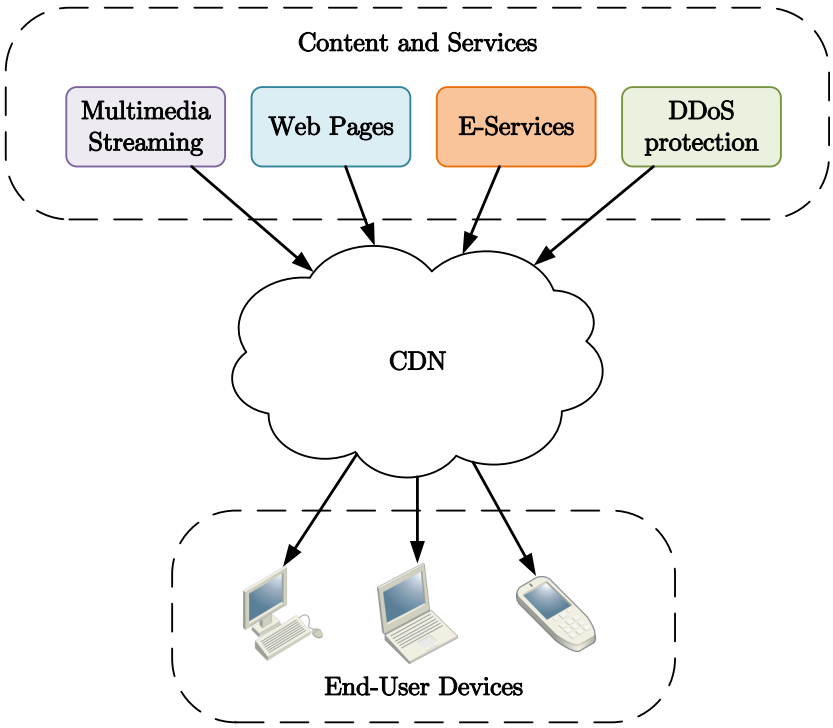

Figure 17: Service possibilities of a CDN

4.6 Content Distribution Models

This sections aims to provide a comprehensive and at the same time a compact description of various CDN technology models. Each of these models is specialized to address certain issues, whether these are related to management, maintenance performance, scalability, server location and reachability [40]. The idea is to present basic well-known CDN models, which have been present in various CDN related literature and research papers without diving too deep into details. There might be even more models available, but those are most likely a combination of technologies presented in this research.

Everything begins with a basic client-server model, which has been developed into two currently dominant and distinct, but radically different architectural solutions: highly distributed co-location and large data center network-core [47]. The co-location architecture attempts to maximize presence near edges of networks and network locations. On the other hand, the network-core architecture focuses to have presence near large data centers and the main network backbones. Additionally there are three mode models, where the benefits provided by other technologies such as cloud, peer-to-peer or even multiple CDN are harnessed to complement the weaknesses of the aforementioned pure CDN technologies.

4.6.1 Client-Server

Arguably the simplest and perhaps the oldest method to distribute multimedia content over the Internet is the *client-server* model, where one centralized server (farm) provides services to multiple end-users [40]. The management of such approach is quite straight forward and simple. However, this solution has a single point of failure and it does not scale well when there are thousands (or more) users generating numerous requests to the single server at one site. In other words a single server could become a *hotspot* [39]. Adding more servers, upgrading hardware and the underlying network has its limits and does not scale well. Cost wise this approach becomes quickly very expensive.

For example, one stream of a HD video at 5 Mbit/s equals to one customer, then the sum of 100 customer streams would be equal to 500 Mbit/s. One server would be serving 100 HD video streams at 500 Mbit/s, which is not much. The server network connection should be able to scale up to 10 Gbit/s on one link, which means up to 2 000 video streams. However, when the amount of data traffic is measured in hundreds of Gbit/s, 20 000 or more customers, then the scalability of one server at one data center site becomes quickly a bottleneck in the centralized client-server model and therefore, the load has to be distributed across multiple servers.

4.6.2 Highly Distributed

Highly distributed CDN, or alternatively *co-location*[47] concept was initially developed by Akamai Technologies Inc. to efficiently serve small-sized files over the Internet [39][40]. The basic idea is to deploy a CDN inside as many ISP PoPs as possible. This method brings content closer to a user in order to improve perceived performance, which in turn offers smaller delays in content delivery. Smaller delay often mean better throughput. The highly distributed CDN approach is ideal for addressing the *last mile problem* [1], where the Internet access speeds of the end-users are rather limited.

Large number of server clusters scattered around the globe might become challenging from management and maintenance perspective. However, with automated fault detection and proper arrangements these should not be an issue. What comes to content storage location, static content is stored at the network edges and dynamic content is served directly from the origin server. At least this is the case for Akamai [1].

4.6.3 Large Data Center

Large data center CDN, also known as *network-core* [47], solution relies on much smaller global dispersion with servers being deployed only at key locations, for example near main network backbones. This approach has been adopted by Limelight Networks [1] and it aims at providing a solution to the *middle-mile* problem [48], since over the years, the increased speed in the Internet access has enabled richer content.

When comparing to highly distributed CDN, end-users of the large data center CDN solutions might experience higher delays. However, according to one research highly distributed CDN could be operated with less servers than initially planned [47]. Smaller global footprint results in reduced management and maintenance overhead. These would translate into smaller OPEX, because smaller amount of servers and other infrastructure.

4.6.4 Peer-to-Peer Assisted Content Delivery

A *Peer-to-peer* or shortly P2P, is an overlay service build on top of the Internet. The network components are similar to client-server model, but in case of P2P each client can also act as a server. The P2P model could be seen as a competing platform to CDN technologies, which is not quite true. Both are designed to improve content delivery to the end-users, but there are few fundamental differences in the two approaches.

There is no centralized management entity controlling communication sessions in the P2P model, unlike in the CDN architecture. Also, P2P does not involve any infrastructure cost. It relies on existing network infrastructure, such as the Internet. Also the supply of resources grows with demand, since each new joining client starts to share the workload of the peering sessions [49]. However, there is no QoS guarantees and P2P is often just a simple client based application. On the contrary, building a new CDN architecture would require heavy initial investments, which later on would provide better network efficiency with QoS support. A summary of comparison between CDN and P2P models is presented in Table 4.

Table 4: A short comparison of CDN and P2P technologies

Feature	**CDN**	**P2P**
Capability	Limited	Increases by Each Peering Node
Scalability	High Cost	Low Cost
Reliability	High	Low
Stability	Stable	Dynamic
ISP Friendly	Yes	No, ISP independent
QoS	Yes	Best-Effort
User Management	Centralized	Unmanaged
Service Node Authentication	Centralized	Distributed or None
Content Source Monitoring	Possible	Difficult
Content Copyright and Security	Controlled	Uncontrolled

A next-generation P2P-CDN hybrid has been already proposed by researchers [50], where content delivery would be off-loaded at least partially to end-users. When combined with a CDN, the P2P is used to leverage the resources of participating peers to relieve load levels on content servers [1]. In other words the P2P and the CDN architectures could complement each other. For example Akamai Inc. and other CDN providers have been actively acquiring P2P related technologies [51].

4.6.5 Cloud and Virtual CDN

Cloud services could be described shortly as pools of virtualized resources, which are allocated dynamically according to customer needs and requirements. Service cost is often defined by pay per use model and service level agreements (SLA). The cloud provides mainly three types of services: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS).

Cloud CDN hybrid, or a CCDN, employs functionality of a CDN over cloud technologies and combines them into one service platform. Typically a CDN provides request redirection, content delivery, content distribution and management services. On other hand, a cloud infrastructure concentrates on providing computing, storage, virtualization, security and etc. services [52]. However, traditional cloud services often struggle to provide a fast and reliable content service to end-users over the Internet. Reliable and scalable service provisioning requires significant CAPEX and OPEX investments for cloud service providers [52]. Besides they are not specialized in designing efficient solutions on top of the Internet architecture. This is where the CDN technology steps in.

A virtual CDN (VCDN) takes a step further, where CDN software would be running within a cloud virtual machine (VM) [53]. Resource utilization and service provisioning is often inefficient in cloud or CDN services, especially in scenarios where the resource demand might fluctuate between high and low values. Virtualization technology enables more efficient resource allocation leaving none to waste, since a number of CDN services running on a VM would compete for these resources. This leaves little room for under-utilized expensive processing power of the VM host.

4.6.6 Multi-CDN

A multi-CDN could be described as a combination of multiple (third-party) CDNs used for multimedia content delivery over the Internet. Often a content provider does not own any CDN infrastructure, since it would be a very expensive investment. Netflix is a perfect example of such multi-CDN outsourcing strategy adoption as presented in Figure 18 [54].

The performance of a certain CDN provider may not be consistent within different geographical locations. Since the request-routing systems are designed to find the best replica server based on various metrics for content delivery within a CDN, why not do the same on higher level where the selection is based on best available CDN provider. There is no need to switch content providers manually, as the process is automated and managed by the content providers origin server as shown in Figure 18.

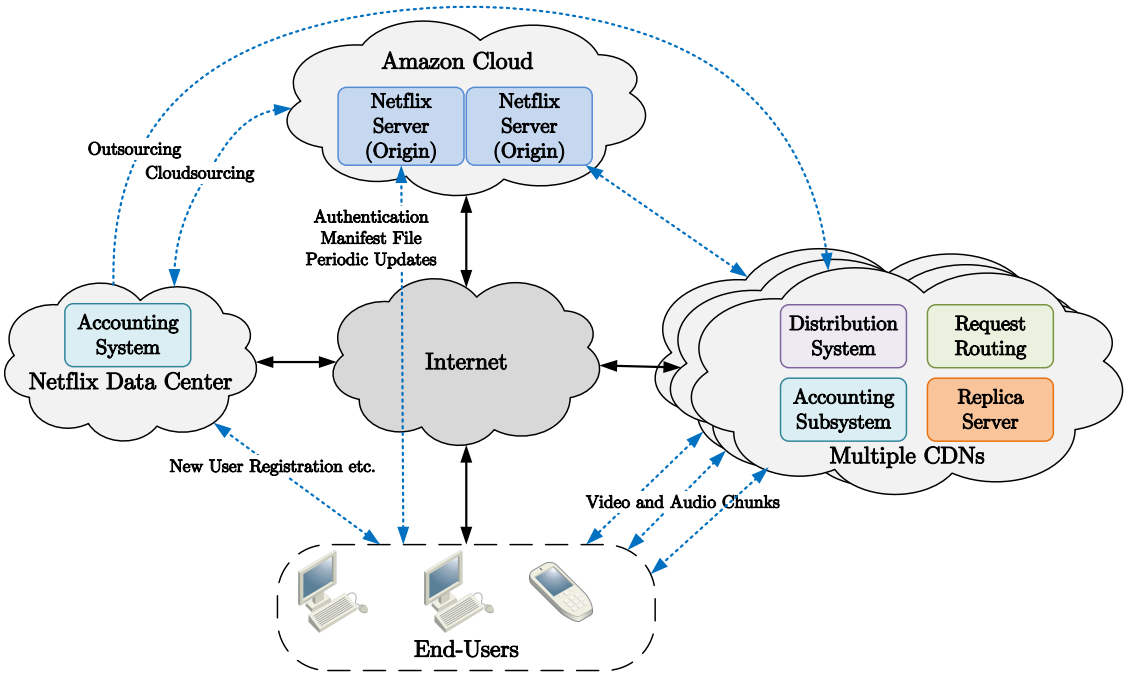

Figure 18: Netflix Multi-CDN architecture

5 Interviews

In order to gain better understanding of the content delivery ecosystem a group of experts were interviewed from different technology domains. This section introduces the method of the interviews and the obtained results.

5.1 Methods

There are multiple methods for conducting an interview. These can be classified into four main categories [55]: structured, semi-structured, unstructured and informal. In a *structured interview*, the set of questions is clearly predefined with carefully chosen wording. The structured method produces consistent data among interviewees, since there is little room for variation in the responses. A *semi-structured* interview acts as a base guideline for selected topics, where the structure may be slightly altered, leaving room for adjustment during the interview process. In an *unstructured interview*, there is a clear plan regarding the focus, but the questions tend to be open ended. Lastly, an *informal* interview is mostly based on the observation of participants without any general guidelines to follow.

In this research, the semi-structured approach was chosen for the interviews. A list of the original questions is presented in Appendix A. These questions are divided into five categories, where each category was designed to cover certain topic from the selected domains presented in Table 5. A total of five experts were chosen for interviewing, one for each specified domain. Additionally, each expert was given a shortened name for the referencing purposes as shown in Table 5.

Table 5: Summary of interviewed experts.

Domain	**Expert Role**	**Reference**
Trends	Technology and Development Manager	TDM
Video	Video Architect	VA
Content Distribution	Distribution Manager	DM
Content Management	Global Sourcing Manager	GS
Network	Network Architect	NA

The interviews were conducted over a one-week period in April 2016. An estimated 45 minutes long time slot was reserved for each session and the questions were submitted beforehand to the interviewees. At the beginning of each interview session. a short presentation about the thesis research topic was given. These methods enabled for the interviewees to prepare more relevant answers around the selected interview domains.

5.2 Outcome

Due to the nature of chosen semi-structured interview method, it was possible to change the order of questions freely according to the conversation flow. The

interviewees had many interesting and inspiring aspects to consider for each domain, but some of the results were omitted due to being out of the scope. All relevant results were gathered as notes written on a paper. In reality, the duration of each interview session varied between 30 to 60 minutes.

5.2.1 Trends in Video Distribution over the Internet

The very first step was to identify trends and development patterns behind Internet video consumption. The interviewee [TDM] is responsible for technology and development management of an OTT Internet based video service in a large broadcasting company, which also employs digital television broadcasting (linear TV) service.

If compared to traditional broadcasting, most of the video content (about 80%) is also published for internet distribution. There might be exceptions where video content is being published solely for the Internet distribution purposes. What is notable about the future development plans and current growth is that they are heavily in favor of the IP based distribution.

Videos can be watched on various devices and these have been categorized as desktop computers and small mobile devices, such as smartphones and tablets, according to interviewee [TDM]. Laptops are considered as desktops for simplicity. Desktop computer in terms of the video content consumption is clearly in slow decline, while the share of mobile devices is increasing with high rate. Two years ago the situation was rather even with the share values of approximately 50% for both categories.

The user base is steadily growing as the video service platform is reaching its mature point. The amount of users within segment above age of 45 years is noticeably increasing each year. This means that the customers are becoming familiar with possibilities of Internet based video streaming services as the video service matures. Generally, the user statistics indicate that the amount of users in a certain age segment strongly correlates with the age of the user. Internet video streaming services are typically preferred by younger generations, making them early adopters.

An average video screen time varies from 30 minutes to 40 minutes depending on available published content and an average user would visit the video service at least three times a week. In most cases, the user watches the whole session entirely at once. VoD is an active decision, therefore full attention is on the video playback. This is unlike in the broadcast TV where the content might be considered as background noise up until something interesting appears on the screen.

The video content library might contain 30 000 – 40 000 assets at once. To ease the selection for the users, a recommendation system similar to Google's YouTube has been implemented.

5.2.2 Video Technology

Video content distribution has become a vital part of the Internet. In order to find out useful information, an interview with a video architect was scheduled [VA]. The discussion was related to video standards, encoding and distribution.

Most of the video production is done in HD (1080) quality. There are some exceptions such as video content coming from old archives and third parties, which most likely are in SD video quality. UHD or higher screen formats are not yet mainstream and those are not yet widely supported. That is the main reason why these are not yet used in production. Another reason is the high storage costs of UHD quality raw uncompressed video involved in the production process. Despite everything, the main content production units would like to produce content in UHD quality and internet distribution is ready for UHD.

Video material has been compressed using H.264 codec standard since 2008. It was selected out of many codecs due to its wide support and applicability among different devices. Also, saved storage space and transcoding efficiency were important. *Transcoding* is a process where a video encoding is converted from one encoding standard to another. Other video codec candidates were Google's VP8 and VP9.

In the near future a change towards H.265 is being considered, but that depends on several factors. There must be a wide range of devices supporting this codec. Moreover, the H.265 has royalty fees. Whether or not the H.265 will be the final choice is not certain yet. At the moment H.265 is chosen as the best candidate.

Currently the video streaming services are based on Adobe's HTTP dynamic streaming (HDS) media container protocol, commonly known as Flash. However, the Flash based streaming system is going to be replaced with something else in the near future, most likely MPEG-DASH or HLS streaming technologies within a year. Nevertheless, MPEG-LA organization has announced a call for patents, which most likely will end up into creating royalty pools for the MPEG-DASH related technologies. This might greatly influence the final choice.

5.2.3 Content Distribution

The third interview was planned to address the content distribution for an OTT video streaming service. The interviewee is specialized in video distribution management, especially in using modern Internet-based techniques.

According to the source [DM], they use a mix of their own data center and a third party CDN. The CDN is used both for load balancing and distribution purposes. The data center acts as the origin server for the used CDN. There have been considerations for P2P-assisted CDN, but this seems to have certain limitations from legal and digital rights perspective. It is unsure whether the digital rights would be violated or not in P2P assisted distribution model. On other hand, how much of the end-user owned network capacity would P2P be allowed to be used for distribution purposes. Additionally, multi-CDN strategy is under consideration, where a CDN could be selected by various criteria, such as best performance, closest location to end-user, redundancy and many more.

The most important features of a CDN are scalability and reliability. A CDN must be able to adapt quickly to changing conditions. Any other features are counted as value adding services, which include but are not limited to, adaptive streaming and different protocol support. In addition, statistics collection is seen as a very important feature, where the server collects data and end-users media player sends

back the playback related information enabling an extensive view of the end-to-end performance.

Distributions costs are based on the amount of transmitted bits. Unicast mode is most expensive due to nature of point-to-point connections, where each user is treated as a separate data flow. Multicast mode would be most desirable for the live events, which would most likely attract many simultaneous viewers.

5.2.4 Content Management

The interviewee [GS] is responsible for video content related global sourcing management in a content provider company specialized in terrestrial broadcasting and OTT IPTV service.

Most of the content is purchased and obtained from international sources. Roughly 1000 agreements are processed each year, which results in more than 4000 hours of video content. For such large scale processing a database is used to manage and store everything related to content agreements. These include the digital rights, prices and detailed agreements. The stored information is linked to the actual content. The content distribution rights are stored in a metadata file, which in turn is used to identify the distribution rules related to the corresponding content. These rules may vary from geographic location (international and national level) limits to content delivery type such as broadcasting and live events.

At the moment the content coming from international sources is limited to domestic distribution only. Similarly, the domestic content is mostly available in domestic area. The identification of individual internet users and their location is based on the IP-address, since each public IP-address space has been assigned to a specific country. The traffic can be routed over the Internet only with public IP addresses.

When speaking about distribution rights, there are several types such as linear broadcast television (TV), online and VoD. Sometimes the linear broadcast TV and online are linked together in a way, where the latter can be distributed online only after the linear broadcast TV session has been aired.

Content may also have certain restrictions, such as how many series can be published at a time. This dictates the price for certain material, especially video series. Content maker often defines and scales the price by evaluation the target country by asking for statistics of local viewers.

5.2.5 Network

Lastly, a contact from an ISP was interviewed. As a network architect, he plays a key role in overseeing development of the ISP networks [NA]. Due to the nature of the interviewee's role within his organization, he has a broad view of the network infrastructure planning.

As stated earlier in Section 2.6, the network topology can be segmented into three logical layers. These layers are access, distribution and core networks. According to interviewee [NA], the ISPs prefer to use term aggregation instead of distribution. Nonetheless, both terms refer to the same logical network layer. All of these layers

can be either fixed or mobile networks related, although mobile network components are interconnected by underlying fixed network topology.

At the end of 2015, the ratio between fixed and mobile traffic in the ISP network was 75% (58000 TB/month) and 25% (17400 TB/month), respectively. The amount of traffic within fixed networks grows linearly each year, unlike the mobile, networks where traffic growth is nearly exponential. Over 50% of foreign data traffic is originating from Netflix, Akamai and Google (Youtube) each with shares of 23%, 14% and 13% respectively. All these are classified as video streaming applications.

The greatest challenge lies within the mobile networks due to massive data traffic growth each year. Mobile access networks are clearly identified as future bottlenecks. More potential bottlenecks can be found deeper within the mobile core network, if mobile data traffic will continue to grow with the current pace. Many upgrades are planned for all mobile network layers in the near future. The new mobile technologies, such as the 5G, promise large mobile overall network capacity improvements. Mobile access networks can be upgraded within a two-week period, unless there are some special requirements, such as additional site construction related work. On the contrary, the mobile core is much slower to deploy, since it requires careful network planning along with limited maintenance breaks and other hardware installation related matters. The general idea is that the components deeper towards the network core are harder to upgrade.

Fixed networks are staying way ahead of current traffic growth, especially on distribution and core network layers. Access networks are very location dependent, which often might limit the maximum access speeds for used data connections.

Estimations for the network capacity are based on thresholds, which monitor the amount of traffic on the links. These thresholds are defined by a group specialized in network upgrading and planning, such as the place where the interviewee works [NA]. The upgrades within a network are planned specifically for each network element based on predefined time intervals with preferably large capacity increments at a time once per year. The frequent changes and small capacity increments would reduce usability of the network and therefore the QoE for the end-users. Furthermore, frequent maintenance breaks involve higher costs. Redundancy is most important aspect when planning network upgrades. Capacity over links, which are designed to be redundant must never exceed the total sum of one link in the redundant set. In another words, if one link goes down the traffic for the first link will be re-routed on the secondary link and therefore possible network congestion is avoided.

From the ISP perspective it does not matter what kind of distribution model is used for video streaming. What matters is the distance between the end-user and the CDN node. The closer the content is to the end-users the less video content would traverse over the core and the distribution network layers.

ISPs cooperate with CDN providers to some extent, at least the in-house IPTV service has been built on top of a third party CDN provider platform. Currently, the ISP in question is providing its own CDN services and more expansions to CDN market are planned in the near future. For example, a new data center for cloud-based services is under construction, where CDN capabilities are also taken into account.

6 Key Findings

In this section, key findings are presented for further discussion. First, the content delivery ecosystem value networks that are created according to gathered information from various sources are discussed. Next, the study proceeds to analyze QoS and QoE relation, followed by video and streaming related observations. Lastly, the benefits of the CDN are summarized along with the most popular request-routing mechanisms.

6.1 Value Network Configuration

Value networks were constructed based on information gathered from both the interviews and various literature sources used for this study. Method and illustration framework used to construct the value networks is derived from the research conducted by Casey et al. [4].

Firstly, basic relations of the content delivery ecosystem are presented. Secondly, there are mainly two distinguishable video distribution models in the Internet. These are the over the top (OTT) television and internet protocol television (IPTV). Although there might be even more models available, those are deemed as variations of the two previously mentioned models, therefore are not considered here. More detailed descriptions of the OTT television and the IPTV are provided in the next three subsections.

6.1.1 Business Relations of the Stakeholders

Ideally, the value network would be configured as shown in Figure 19, where the stakeholders and their relations are as presented earlier in Section 2.1. As a result the ISP is located in the middle of the Figure 19 interconnecting all stakeholders when inspecting from technical perspective. It should be noted that this scenario is only accurate if all stakeholders are from the same geographic area, where one ISP should be able to connect everything. Therefore, the value network presented in Figure 19 is a simplified scenario.

From the business perspective, an end-user buys access to multimedia content separately from a content provider and the Internet access from an ISP. The content provider on the other hand rents servers from data center provider or buys CDN services directly from a CDN provider. Lastly, the network connectivity is provided by the ISP for all stakeholders in the ecosystem.

6.1.2 Over the Top Television

Term *OTT television* is used when a video service is provided without specialized QoS over an unmanaged network, such as the Internet. Unlike in the traditional linear broadcast television, where content is published within a set of television channels at a given time slot, the OTT model is relying on VoD streaming, where the content is accessible at any time as long as the video content is available for public distribution. There is no compensation for traversing the underlying ISP owned

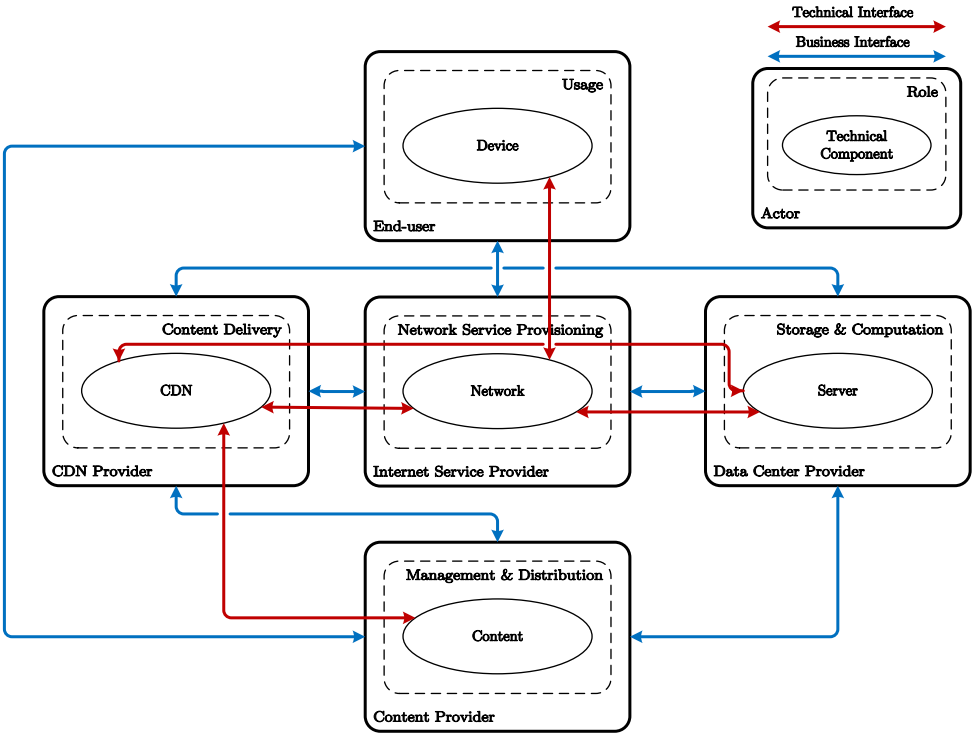

Figure 19: Content delivery base relations

network, when using the OTT video services. Revenues generated by the OTT video service are going straight to the third party content provider as presented in Figure 20. Often, this kind of situation is not approved by the ISPs, and naturally they would like to have a compensation due to additional traffic load in the network. The content distributed by the OTT video services can be accessed from any network connected to the Internet. However, geographical restrictions for the content may apply due to the limitations set by the digital rights and the publication related license agreements.

As shown in Figure 20, the end-user buys video services directly from the content provider. Therefore, the ISP is going to lose some revenues by acting merely as a bit-pipe. Also in this scenario, the CDN provider has its own server infrastructure. The content might traverse several networks along the path before reaching the end-user. The content provider is also a customer to the ISP.

6.1.3 Internet Protocol Television

IPTV is a video service, where the video content is distributed within a managed IP-network, often managed by an ISP. In managed IP-networks, the QoS parameters are carefully monitored and enforced. The IPTV service is quite similar to the traditional linear broadcast television with the difference of content distribution happening over the IP-networks, but also the video content might be made available in the form of the VoD streaming as an exception. In this case the customer subscribes for a fee

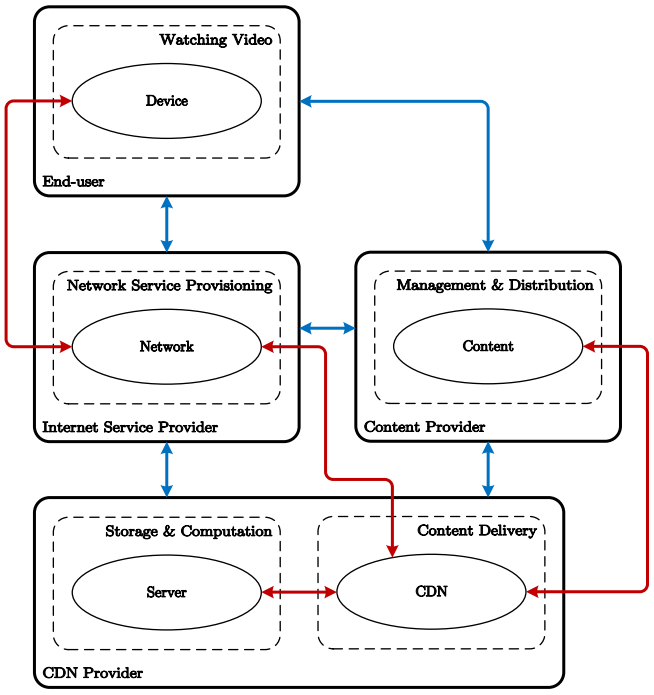

Figure 20: Value network of OTT television

to both, the Internet access and the video service provided by the ISP. The IPTV service can be considered as value added service to the Internet access. Available content may be restricted to the customers of this particular managed network only in most cases.

Similarly as in previous value networks, a scenario is demonstrated in the Figure 21, where the network, the CDN and the server related services are provided by the ISP. Access to the content is bought from the content provider. In this scenario, the ISP can be seen as the middle-man. This kind of setup would be ideally preferred by the ISPs, since there would be guaranteed compensation for the video traffic.

6.2 Relation of Quality of Service and Quality of Experience

As previously defined in Section 2.2, the QoS is defined by a set of agreed technical parameters for operating a service at acceptable levels. The QoS parameters and mechanisms related to transport of packets at the network layer in the protocol stack enable the network operator for example to design, build and manage their networks. This is often invisible to end-users. In other words QoS could be described as the ability to satisfy the requirements of a user, but from a user perspective the QoS does not take into account what a user could possibly experience personally nor does it provide any specifications to measure the actual experience. Shortly QoS is seen as technology focused, whereas the QoE is seen as more customer focused. Views provided by the QoS are solely from within a technology, rather than from outside. The user is rarely interested in small details deep within the service.

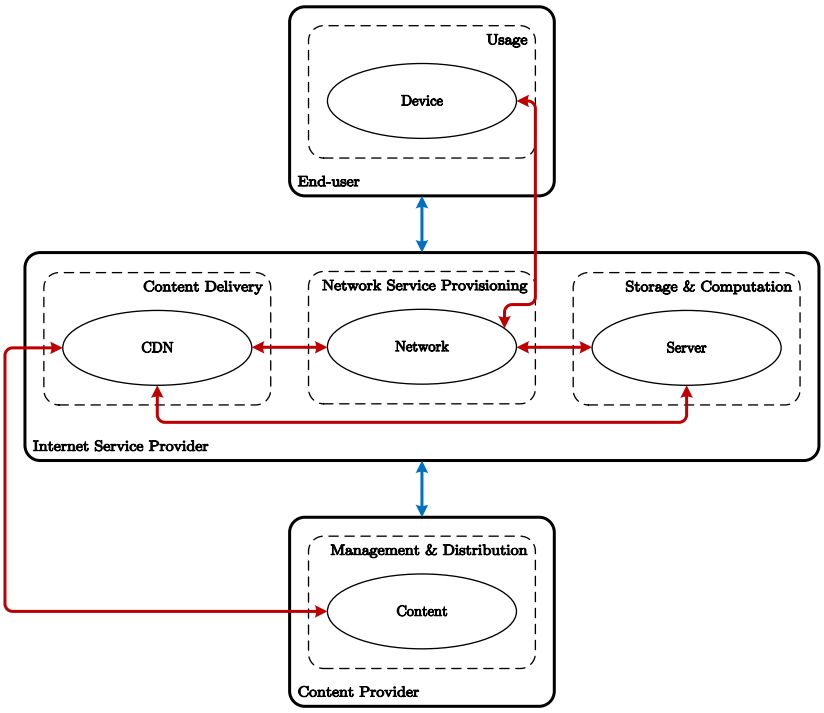

Figure 21: Value network case for IPTV

What end-users personally experience is the QoE during the use of services. The QoE is designed to address user's personal expectations in a measurable way. However, measuring QoE is often complex and challenging due to possibly subjective views of the user. The metrics are hard to define or they might be not perfect. Also, it is important to note that QoE should not be mixed with *quality of service* (QoS). According to the QoE definition, the QoS of a service is only a property of a system. However, the actual user and the context of use are not taken into account in QoS.

As conclusion, the QoS could be defined as a part of a bigger outward picture, the QoE. The QoS and QoE are complementing each other. The user has certain needs and these needs are translated into technical requirements measured by QoS. Expectations are generated by the user during the usage of a service and these are measured often by rather creative cross-disciplinary research methods as was done in collaborative research conducted by Ericsson Ltd. and Neurons Inc. [9].

6.3 Observations Related to Video

There are two aspects which are most relevant for this work. Firstly, how the video codecs and formats are linked together and how they affect the evolution of the CDN technologies. Secondly, what kind of impact video distribution modes could have?

6.3.1 Codecs and Formats

The development cycle of digital video image format standards tends to trigger the development of more efficient video encoding algorithms due to increased file sizes. Smaller and compact size video files would be preferred by both storage and network applications, since larger files require more storage space and more network bandwidth for streaming services. As it can be noticed, while the UHD-1 (4K) and UHD-2 (8K) formats have not yet become mainstream even though first compatible hardware is out, the development of new more efficient video encoding standards such as H.265, VP9 and VP10 is already in full progress.

Also, the CDN evolution is mostly driven by video technology development cycles. Technology of the CDN must stay well ahead of video related trends, since scalability and efficiency of bandwidth heavy content is critical for the core businesses of the CDN, especially on streaming services domain. The new codecs such as H.265 (or VP9 and VP10) will buy time for the CDNs as well for ISPs to get ready for the future of the bandwidth heavy multimedia applications.

6.3.2 Distribution Modes

What comes to video distribution over the IP-networks, the most efficient method would be to mimic broadcast distribution from the linear DTT networks. However, due to the Internet being a shared platform where other services are also present the broadcast distribution is out of the question. As presented in Section 2.7 broadcast is a connectivity primitive where the same data is sent to all nodes on a network whether a node is destined or not to receive the traffic. For example, there is a person in a room trying to concentrate on a demanding task and there is also another person speaking out loud. The studying person will hear the speaking person even though he would not like to.

In IP-networks the problem created by broadcast reception could be solved with multicast, which is a special case of broadcast. The multicast mode is designed to work similarly to broadcast mode with an exception that the participating receivers could join a multicast group, which is similar to a channel on a TV network.

Unicast tends to be the most expensive solution for the OTT streaming services, since according to an interview results, CDN charges according to transferred bits. In unicast mode each user is treated as a separate stream, but the problem of unicast and multicast is not that simple. Unicast works best for VoD streaming, since users are less likely to watch the same content at exactly the same time, which in turn renders the idea of the multicast obsolete. This would explain why CDN providers might be reluctant to provide multicast services, since that would result in great drop in revenues.

6.4 Importance of HTTP Adaptive Streaming

Traditionally a video has been streamed over a network by using application-level *real time protocol* (RTP)[56] and *real time streaming protocol* (RTSP) [45]. There might be certain issues involved with the usage of dedicated streaming protocols such

as RTP and RSTP when traversing certain intermediate network elements, since these could cause unwanted interference in the video stream. A dedicated port was required for these streaming capable protocols, which is often blocked at a router or a firewall. In host to host communications an application port is comparable to an apartment number in the box of flats for example. This way each application residing on a host can be easily differentiated and identified. Additionally a dedicated streaming protocol might be proprietary, which would involve royalty and licensing fees.

The most flexible streaming services are achieved by utilizing a generic protocol such a HTTP. Streaming on top of the HTTP has become the de-facto standard for video content delivery over the Internet. Unlike the HTTP, other streaming protocols require special arrangements and support from devices. HTTP is widely supported by many devices with IP-network connectivity out of the box. Additionally HTTP is much easier to cache at standard HTTP proxy enabled servers, which are used to redirect HTTP traffic.

An interesting feature of HTTP is the possibility of intelligent streaming, where back-channel communications between server and client can be used for analyzing the quality of running streaming session on the fly.

As a platform, the Internet is rather unpredictable due to many reasons related to lack of the proper QoS mechanisms. There might be congested network links along the path between the content server and the end-user causing fluctuations in end-to-end connection speeds. Even though Internet routing should be able to detect these anomalies, i.e. if the underlying network has been properly configured, the re-routing process takes some time to stabilize. In some cases there is no re-routing possibility at all. Additionally the quality of access network connectivity on link level has the greatest impact on connection speeds. As a consequence, the video stream packets are delivered at uneven rate to the receiver.

The HTTP has been adopted to address both buffering and network anomalies. Higher resistance to network variations and lesser buffering would translate into improved QoE for the end-users.

6.5 Benefits of a Content Delivery Network

There are two main selling points for a CDN service. The first point is the on-demand capacity provisioning for content providers and the second point is the improved content delivery performance for the end-users. Any other factors tend to be value adding features.

The CDN has been particularly designed to address so called last-mile and middle-mile network performance issues, which are alternatively called access and distribution network layers. Location is what matters most in video content delivery. Traditionally content has been served from one location resulting in greater delay, if the user happens to be far away from the server with the content due to the law of physics. This would translate into lesser user experience. Having global reach and being closer to the actual end-users, CDN attempts to achieve lesser content delivery latency with more reliable distributed content delivery architecture.

On the contrary, the CDN benefits Internet networks by off-loading traffic from certain parts of the network freeing capacity for other use. Similarly from content provider perspective a CDN would off-load traffic destined to the origin server, since a replica server could also handle the requests.

Furthermore, CDN technology can be used to further enhance cloud services. Traditional Cloud services have often trouble with providing fast and reliable content service to end-users over the Internet. Scalable and reliable service requires significant operational and maintenance cost, but with the help of CDN technology it should not become a problem.

Another hot topic in CDN front is the CDN-P2P hybrid technology which can significantly reduce content delivery related costs especially for last-mile content delivery. Therefore P2P technology acquisitions have become a new trend.

Before a CDN can be deployed for service production purposes, several factors have to be considered [1]. Placement of the replica servers plays a key role in performance on geographical location scale, because greater physical distance between the two would strongly correlate with increased content delivery delay. Organization and structure of deployed data centers must be able to cope with the load generated at the point of presence of the CDN, otherwise excessive content requests are going to be redirected to another node with less load. Additionally, there should be a scheme for content distribution mechanisms, such as replica server cache integrity and other content management related matters. Another important CDN component is the request-routing, which defines how effectively the content will be served for customers from different locations.

Lastly the entire CDN system management must be well planned and executed, since the management of large scale systems tends to be tedious and very prone to errors if no proper task automation has been implemented.

6.6 Request-routing

The Application layer request-routing is often used by the CDN providers, where solutions rely on some form of DNS infrastructure. The other solutions are not so effective as the DNS at large scale. For example, the HTTP URL rewriting at the origin server is deemed overhead heavy, since it creates an additional need for computational resources and request processing time as stated in the research conducted by Yin et al.[1].

There are two well known variants for the DNS based request-routing implementations as presented earlier in Section 4.4. The Hierarchical DNS structure used by Akamai Inc. also attempts to conceal the content source from the end-users, while redirecting the requests. Another simple but effective approach is used by Limelight Networks, where single level DNS structure is combined with IP anycast. The combination of DNS and IP anycast has been proven to be quite robust in the research conducted by Huang et al. [2].

7 Conclusions

This section presents discussion and summary based on the findings of the previous section.

7.1 Discussion

Overall constant improvements in consumer devices and internet access connectivity are making live streaming more appealing each year, while physical media has been almost abandoned by the current generation. Most of the modern multimedia content is being primarily published for distribution over the Internet. Multimedia content delivery capable services are built on top of the Internet architecture. This means that the Internet protocol suite could be expanded by one more abstraction layer on top of the application layer, a CDN-layer.

Many CDN providers advertise that a CDN would increase the speed of a service. This argument is based on a quite vague logic. Technically a CDN does increase service responsiveness, which seems to create an illusion of increased speed. It would be wrong to think that a CDN would improve existing network capacity over certain links. However, it is true that CDN does offer a better utilization of network resources. The development of video distribution techniques is clearly driven by upcoming bandwidth intensive UHD video formats. The base criteria for CDN are service functionality, performance and cost.

Characteristics of the Internet Protocol (IP) networks are originally not designed to be in line with real-time requirements of multimedia services. Each packet is individually routed and therefore packets may arrive out of order or be lost along the path due to transmission errors. Link loads may vary along the route and there might be also temporary congestion, which cause additional delays to the delivery of individual packets. Many protocols are inefficient and are not designed for heavy content delivery and protocol timeout settings may vary. As a result, there have been attempts to the to improve reliability at higher levels of the Internet Reference model, mostly at application and transport layers.

CDN infrastructure operates mainly on transport and application layers dealing with the routing and forwarding of requests and responses for content. Well-tuned knowledge of the requested objects provide application layer request-routing systems with better control over best replica selection process.

In addition CDN-hybrid technologies are rather intriguing. When combined with cloud services, the CDN nodes could be effectively provisioned on the fly per need basis. This would enable tiered performance and pricing plans with much more fine-tuned resource allocation. Another interesting direction is the combination of P2P- and CDN-technologies, which results in nearly perfect symbiosis with certain drawbacks. However, the P2P is not carrier-grade service, because it does not meet high availability standards meant for telecommunications purposes due to lack of reliability. Also P2P causes considerable security issues. For example, not many users would be happy to expose their network connected devices to the Internet.

For content providers, the largest expenses come from the amount of transferred

bits due to streaming in unicast mode. They would prefer to switch over to multicast based distribution, but CDNs are reluctant for a reason. Multicast mode might result in large revenue losses for CDN providers. In addition, the multicast is not well supported in the Internet core and it would require a well-coordinated agreement between large backbone network providers, such as Tier 1 ISP, and also with network equipment manufacturers. Multicast is only effective for live video transmission, which increases chances that there would be multiple users viewing exactly same content at exactly the same time. In other words, multicast is a niche service and it would only work within a closed managed network. Another application for multicast could be the replication process, where origin server content is synchronized across replica servers within a CDN. However, this would most likely require the content provider to specify whether to use multicast or unicast for the replication process.

ISPs might have chances to gain strong position in CDN market, due to precise knowledge of their network capabilities. Furthermore the largest costs associated with a CDN are controlled by ISPs. These include the network connectivity, capacity and point-of-presence (PoP) at key locations. This speculation is only on theoretical level though, but in reality everything might be different.

7.2 Summary

During last decade, steady improvement of the Internet access speeds has enabled extensive use of various real-time based multimedia applications. Also, improvements in video image quality have been following the technological progress. Greater strain on the distribution and core network layers is created because of these two factors. Nonetheless, according to the interview results, the fixed networks are well ahead of the traffic growth. The true problem lies within mobile networks, where data traffic amount has been growing at exponential rate.

Overall, the Internet as a platform is quite challenging for real-time delay sensitive multimedia applications. End-to-end QoS cannot be guaranteed due to the best-effort nature of the Internet, as there is no central entity controlling the heterogeneous interconnection of networks. Service level agreements (SLA) are only valid between two involved parties and that does not apply for the whole Internet.

The motivation behind various CDN technologies is to address the dynamic conditions of the Internet by enabling global reach with more efficient and scalable content delivery. The content is brought closer to the end-user for reducing delivery related delays of any time sensitive content, such as video, in order to improve end-user's perceived QoE. In other words the CDN is an attempt to bring carrier grade service level into content distribution market. CDN infrastructure mainly consists of content server network and DNS server network and the performance can be evaluated by analyzing content and DNS servers' responsiveness.

As a summary the CDN is the core of rich multimedia distribution over the Internet, where it is perfectly suitable for streaming purposes by relying on application layer protocols such as HTTP for transport of a video stream.

7.3 Future Research

There are several topics that raised many questions. Firstly, how HTTP streaming protocols operate on packet level. Another notable feature is the adaptive streaming, which could be studied more.

Secondly, there is lack of reliable comparison between video codecs, namely H.264 and H.265, and their equivalents VP8 and VP9. An extensive research could be conducted to gain an insight on, for example, how much network bandwidth each codec would actually consume when streaming for example HD and UHD video content.

Thirdly, there are rather many documents about CDN hybrids, but there is no extensive performance comparison done between traditional CDN and hybrid solutions such as P2P assisted CDN and Cloud based CDN. Also, one of the interviewees presented an interesting question concerning the Internet P2P assisted distribution model: Would customers allow their network bandwidth to be used for P2P assisted content delivery?

Lastly, one of the interviewees claimed that IP based video content distribution is very cost effective. The broadcast television is expected to make at least a partial transition towards a fully IP-based video content distribution in the far future. In the end, cost effectiveness, compatibility and scalability factors are driving decisions when selecting technology from multiple available variants.

References

[1] H. Yin, X. Liu, G. Min, and C. Lin, "Content delivery networks: A bridge between emerging applications and future IP networks", *IEEE Network*, vol. 24, no. 4, pp. 52–56, 2010, ISSN: 0890-8044.

[2] C. Huang, A. Wang, J. Li, and K. W. Ross, "Measuring and evaluating large-scale CDNs", in *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement*, 2008, pp. 15–29, ISBN: 978-1-60558-334-1.

[3] Cisco Systems Inc., "Forecast and methodology 2014 – 2019", *Cisco Visual Networking Index*, May 2015. [Online]. Available: `http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf` (visited on 02/11/2016).

[4] T. Casey, T. Smura, and A. Sorri, "Value network configurations in wireless local area access", in *Telecommunications Internet and Media Techno Economics (CTTE) 9th Conference*, 2010, pp. 1–9.

[5] W. B. Norton. (2010). Internet service providers and peering, [Online]. Available: `http://drpeering.net/white-papers/Internet-Service-Providers-And-Peering.html` (visited on 03/04/2016).

[6] Recommendation E.800, *Definitions of terms related to quality of service*, ITU-T, 2008.

[7] ETSI, "Human factors (HF); quality of experience (QoE) requirements for real-time communication services", *ETSI Technical Report*, 2010.

[8] Recommendation P.10/G.100, *Vocabulary for performance and quality of service. amendment 2: new definitions for inclusion in recommendation itu-t p.10/g.100*, ITU-T, 2008.

[9] Ericsson Ltd., "Streaming delays mentally taxing for smartphone users: Ericsson mobility report", *Press Releases*, 2016. [Online]. Available: `http://www.ericsson.com/news/1986667` (visited on 02/25/2016).

[10] J. D'Onfro, "More than 70 percent of internet traffic during peak hours now comes from video and music streaming", *Business Insider UK Tech News*, 2015. [Online]. Available: `http://uk.businessinsider.com/sandvine-bandwidth-data-shows-70-of-internet-traffic-is-video-and-music-streaming-2015-12` (visited on 04/18/2016).

[11] Ericsson Ltd., "Data traffic - application", *Traffic Exploration Tool*, 2015. [Online]. Available: `http://www.ericsson.com/TET/trafficView/loadBasicEditor.ericsson` (visited on 02/18/2016).

[12] T. Drier, "The state of live video 2016", *Streaming Media Magazine: March 2016 Sourcebook*, 2016. [Online]. Available: `http://www.streamingmedia.com/Articles/Editorial/Featured-Articles/The-State-of-Live-Video-2016-110041.aspx` (visited on 04/18/2016).

[13] Recommendation X.200, *Information technology - open systems interconnection - basic reference model: the basic model*, ITU-T, 1994.

[14] F. Baker, "Core protocols in the internet protocol suite", *IETF Netork Working Group - Internet Draft*, 2009. [Online]. Available: `https://tools.ietf.org/id/draft-baker-ietf-core-04.html` (visited on 04/14/2016).

[15] W. B. Norton. (2003). The evolution of the u.s. internet peerin ecosystem, [Online]. Available: `https://www.nanog.org/meetings/nanog31/presentations/norton.pdf` (visited on 03/04/2016).

[16] A. Retana, D. Slice, and R. White, *CCIE Professional Development: Advanced IP Network Design*. 1999, ch. 1, pp. 5–17, ISBN: 1-57870-097-3.

[17] T. Bates, P. Smith, and G. Huston. (2016). Cidr report for 16 may 16, [Online]. Available: `http://www.cidr-report.org/as2.0/` (visited on 05/16/2016).

[18] T. Lammle, *CCNA: Cisco Certified Network Associate Study Guide*, 4th ed. 2004, pp. 89–90, ISBN: 0-7821-4311-3.

[19] RFC 4786, *Operation of anycast services*, IETF, 2006.

[20] Recommendation BT.709, *Parameter values for the hdtv standards for production and international programme exchange*, ITU-R, 2015.

[21] Recommendation BT.2020, *Parameter values for ultra-high definition television systems for production and international programme exchange*, ITU-R, 2015.

[22] Recommendation H.264, *Advanced video coding for generic audiovisual services*, ITU-T, 2003.

[23] A. Grange and H. Alvestrand, "A VP9 bitstream overview", *VP9 draft*, 2013. [Online]. Available: `https://tools.ietf.org/html/draft-grange-vp9-bitstream-00` (visited on 03/15/2016).

[24] Recommendation H.265, *High efficiency video coding*, ITU-T, 2015.

[25] J. Ozer, "Amazon, google, and more working on royalty-free codec", *Streamingmedia.com News*, 2015. [Online]. Available: `http://www.streamingmedia.com/Articles/News/Online-Video-News/Amazon-Google-and-More-Working-on-Royalty-Free-Codec-106091.aspx` (visited on 04/15/2016).

[26] M. Yardley, C. Jones, and S. Montakhab, "New service developments in the broadcast sector and their implications for network infrastructure", *Analysys Mason*, 2014, Ref: 2001575-494. [Online]. Available: `http://stakeholders.ofcom.org.uk/binaries/research/infrastructure/2014/broadcast-dev.pdf` (visited on 02/09/2016).

[27] Netflix, "Internet connection speed recommendations", *Netflix Help Center*, 2016. [Online]. Available: `https://help.netflix.com/en/node/306` (visited on 04/15/2016).

[28] A. Aaron and D. Ronca, "High quality video encoding at scale", *The Netflix Tech Blog*, 2015. [Online]. Available: `http://techblog.netflix.com/2015/12/high-quality-video-encoding-at-scale.html` (visited on 04/15/2016).

[29] D. Wu, Y. T. Hou, W. Zhu, Y.-Q. Zhang, and J. M. Peha, "Streaming video over the internet: Approaches and directions", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 3, pp. 282–300, 2001, ISSN: 1051-8215.

[30] S. Akhshabi, A. C. Begen, and C. Dovrolis, "An experimental evaluation of rate-adaptation algorithms in adaptive streaming over http", in *Proceedings of the Second Annual ACM Conference on Multimedia Systems*, 2011, pp. 157–168, ISBN: 978-1-4503-0518-1.

[31] RFC 2616, *Hypertext transfer protocol – http/1.1*, IETF, 1999.

[32] T. Siglin, "Http streaming: What you need to know", *Streaming Media magazine*, 2010. [Online]. Available: `http://www.streamingmedia.com/Articles/Editorial/Featured-Articles/HTTP-Streaming-What-You-Need-to-Know-65749.aspx` (visited on 04/15/2016).

[33] Standard 23009-1:2014, *Information technology - dynamic adaptive streaming over HTTP (DASH) - part 1: Media presentation description and segment formats*, ISO/IEC, 2014.

[34] MPEG Licensing Administration, "MPEG LA announces call for patents to organize joint license for MPEG -DASH", *MPEG LA News Release*, 2015. [Online]. Available: `http://www.mpegla.com/Lists/MPEG%20LA%20News%20List/Attachments/96/n-15-07-27.pdf` (visited on 04/21/2016).

[35] M. Arlitt and T. Jin, "A workload characterization study of the 1998 world cup web site", *IEEE Network*, vol. 14, no. 3, pp. 30–37, 2000, ISSN: 0890-8044.

[36] S. Adler, "The slashdot effect: An analysis of three internet publications", *Linux Gazette*, no. 38, 1999.

[37] RFC 3466, *A model for content internetworking (cdi)*, IETF, 2003.

[38] T. Plagemann, V. Goebel, A. Mauthe, L. Mathy, T. Turletti, and G. Urvoy-Keller, "From content distribution networks to content networks - issues and challenges", *Computer Communications*, vol. 29, no. 5, pp. 551–562, 2006, ISSN: 0140-3664.

[39] M. K. Pathan and R. Buyya, "A taxonomy and survey of content delivery networks", *Grid Computing and Distributed Systems Laboratory, University of Melbourne, Technical Report*, 2007.

[40] R. Buyya, M. Pathan, and A. Vakali, *Content delivery networks*. Springer, 2008, vol. 9, ISBN: 978-3-540-77886-8.

[41] N. Bartolini, E. Casalicchio, and S. Tucci, "A walk through content delivery networks", in *Performance Tools and Applications to Networked Systems: Revised Tutorial Lectures*. Springer, 2004, pp. 1–25, ISBN: 978-3-540-24663-3.

[42] J. Dilley, B. Maggs, J. Parikh, H. Prokop, R. Sitaraman, and B. Weihl, "Globally distributed content delivery", *IEEE Internet Computing*, vol. 6, no. 5, pp. 50–58, 2002, ISSN: 1089-7801.

[43] RFC 3568, *Known content network (cn) request-routing mechanisms*, IETF, 2003.

[44] RFC 3272, *Overview and principles of internet traffic engineering*, IETF, 2002.

[45] RFC 2326, *Real time streaming protocol (rtsp)*, IETF, 1998.

[46] J. Jung, B. Krishnamurthy, and M. Rabinovich, "Flash crowds and denial of service attacks: Characterization and implications for CDns and web sites", in *Proceedings of the 11th International Conference on World Wide Web*, ACM, 2002, pp. 293–304, ISBN: 1-58113-449-5.

[47] S. Triukose, Z. Wen, and M. Rabinovich, "Content delivery networks: How big is big enough?", *SIGMETRICS Perform. Eval. Rev.*, vol. 37, no. 2, pp. 59–60, Oct. 2009, ISSN: 0163-5999.

[48] T. Leighton, "Improving performance on the internet", *Communications of the ACM*, vol. 52, no. 2, pp. 44–51, 2009, ISSN: 0001-0782.

[49] Z. Lu, Y. Wang, and Y. R. Yang, "An analysis and comparison of CDN-P2P-hybrid content delivery system and model", *Journal of Communications*, vol. 7, no. 3, pp. 232–245, 2012.

[50] J. Wu, Z. lu, B. Liu, and S. Zhang, "Peercdn: A novel p2p network assisted streaming content delivery network scheme", *Computer and Information Technology, 2008. CIT 2008. 8th IEEE International Conference*, pp. 601–606, 2008.

[51] C. Huang, A. Wang, J. Li, and K. W. Ross, "Understanding hybrid cdn-p2p: Why limelight needs its own red swoosh", in *Proceedings of the 18th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, 2008, pp. 75–80, ISBN: 978-1-60558-157-6.

[52] C. F. Lin, M. C. Leu, C. W. Chang, and S. M. Yuan, "The study and methods for cloud based cdn", *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2011 International Conference*, pp. 469–475, 2011.

[53] T.-W. Um, H. Lee, W. Ryu, and J. K. Choi, "Dynamic resource allocation and scheduling for cloud-based virtual content delivery networks", *ETRI Journal*, vol. 36, no. 2, pp. 197–205, 2014.

[54] V. K. Adhikari, Y. Guo, F. Hao, M. Varvello, V. Hilt, M. Steiner, and Z. L. Zhang, "Unreeling netflix: Understanding and improving multi-cdn movie delivery", *INFOCOM, 2012 Proceedings IEEE*, pp. 1620–1628, 2012, ISSN: 0743-166X.

[55] D. Cohen and B. Crabtree, "Qualitative research guidelines project", 2006. [Online]. Available: `http://www.qualres.org/HomeInte-3595.html` (visited on 04/24/2016).

[56] RFC 3550, *RTP: A transport protocol for real-time applications*, IETF, 2003.

A Appendix - Interview Questions

Trends

- What devices are used for viewing a video over the Internet?
- What type of video content is the most viewed?
- Are there any estimations about average video length?
- How often the Internet video service platform is used on a weekly basis?
- Where is the future of video distribution heading?
- Would it be possible to make a transition from terrestrial TV broadcasting to fully IP based solution?
- How much of the video content is also published for Internet based distribution?

Content Distribution

- What content distribution model do you use for Internet video and why? (Cliet-Server, Distributed etc.)
- How this particular model was selected?
- Have you considered to change the current distribution model?
- What do you consider to be the most important feature of a CDN?
- Are there any features that available CDN providers lack or cannot offer?

Video

- Which video format is used for video production and storing; SD, HD or UHD?
- Are there any plans to use higher quality video format in the near future?
- Is there any down or up scaling of video quality for different purposes?
- What standard is used to decompress and compress the video content?
- How this particular codec was chosen?
- What is the protocol used for video streaming?
- What video container used to control streaming?
- What kind of expenses are involved in video production?

Content Management

- How content and digital copyright related matters are managed?
- Are there any restrictions in viewing the video content over the Internet? (country, licenses)
- Who defines the prices for the video content and how?
- Does the video caching cause any issues from digital copyrights perspective?

Network

- On which network level the demand for network capacity is growing fastest? (Access, Distribution or Core)?
- Where could be possible bottlenecks in a network topology?
- What kind of applications are generating most data traffic?
- Are these particular sources of data traffic identifiable?
- Where the need for network capacity is growing with fastest rate, in mobile or fixed networks?
- How the network capacity upgrades are planned?
- Does it matter which kind of distribution model is used from the ISP perspective? (CDN, P2P and client-server)
- Are there any plans to cooperate with a CDN provider?