CONVINcE

D1.1.4

Theoretical Models

Editor: Adrian Popescu (BTH)

Reviewer: Yannick Carlinet (Orange)

Authors: Adrian Popescu (BTH)
Markus Fiedler (BTH)
Yong Yao (BTH)
Patrick Arlos (BTH)
Dragos Ilie (BTH)
Jerome Bardon (Harmonic)
Raoul Monnier (Harmonic)
Yannick Carlinet (Orange)
Rickard Ljung (Sony)
Zhi Zhang (LU)
Erkki Harjula (University of Oulu)
Harri Hyväri (VTT)
Reza Farahbakhsh (IMT)

EXECUTIVE SUMMARY

Objective

Deliverable D1.1.4 "Theoretical Models" is the fourth document delivered in WP1 – Task 1.1. This Deliverable is a continuation of the Deliverable D1.1.3 "High-Level Architecture Design". The objectives of this document are, on one side, to provide the theoretical models for video flows in the CONVINcE system and, on the second side, to suggest the concepts of performance optimization in CONVINcE, which are in terms of minimum end-to-end energy consumption and best QoE at end user. Towards this goal, the document further develops on the model for end-to-end (e2e) performance previously developed in Delivery D1.1.3 in terms of power consumption and Quality of Experience (QoE).

Methodology

The methodology used in D1.1.4 is to first define the fundamentals for video distribution, video streaming and, associated with this, video standards. The video standards IPTV (Internet Tv) and OTT (Over The Top) are considered. Theoretical models for networking elements like Head End, Core Network, Content Distribution Network, Edge-Cloud, Radio Access Network and Terminal are considered. Furthermore, theoretical models for video flows and the associated bandwidth models and resource allocation are presented.

The next step is to present a new concept suggested for multi-objective multi-constraint performance optimization, which is in terms of minimum e2e energy consumption and best QoE performance at end-user. The process of performance optimization is considered at both network level (layers 1-3 in the TCP/IP protocol stack) and TCP level (layer 4).

Based on this, theoretical evaluation of the performance expected to be obtained in the CONVINcE testbed can be done and, associated with this, resource allocation in CONVINcE can be done. This is the focus of the next Delivery D1.1.5

Implementation

The first part of the document is dedicated to presenting an overview of basic concepts of video distribution (including video coding and compression), video delivery over the Internet and a short presentation of the CONVINcE high-level architecture. This is followed by a short presentation of the basic concepts of video streaming, also including the streaming process and strategies.

The second part of the document is dedicated to an detailed presentation of the most important theoretical models that can be used to describe the video signal and, associated with this, comparison of these models. Based on this, it has been decided that the Fluid-Flow model will be used in the performance evaluation of the theoretical model of CONVINcE system, to be done in the next Delivery D1.1.5.

Associated with this, networking and performance aspects relevant for the CONVINcE testbed like, e.g., definition of inter-provider link and use of percentiles, are presented as well.

The fundamental target of CONVINcE is to develop new solutions for energy-efficient Video Distribution Networks, under the requirement of good QoE performance at the end-user. This is a very difficult and complex effort. Accordingly, the Delivery describes a solution suggested by other industrial companies for QoE-oriented architecture. This is the starting point for our activity, where one of the main requirements is to avoid the extreme complexity existing in these architectures when developing our solutions.

Therefore, the last part of the document advances a new concept (CONVINcE) for the provision of expected performance, which is less complex. A global optimization mechanism is developed, according to which particular classes of e2e performance are defined in terms of end-user QoE and the associated e2e energy saving performance. Solutions for resource allocations and implementation strategy in CONVINcE are presented as well.

This concept is based on two elements, namely a new algorithm for multi-objective multi-constraint optimization developed for low-layers in the TCP/IP protocol stack (layers 1-3) and, associated with this, a new algorithm for performance provision at the TCP protocol level. By using the new algorithms suggested for CONVINcE, we expect to provide the end-user and system performance.

Results

The main results obtained in the deliverable D1.1.4 are in form of theoretical models as well as new algorithms for performance optimization at CONVINcE. These algorithms will be theoretically evaluated in the next Deliverable D1.1.5, and the expected results will be used to validate the CONVINcE testbed results expected to be obtained in Work Package WP5.

Future Work

The future work is in form of Deliverable D1.1.5, which is about CONVINcE theoretical models and performance evaluation. The expected theoretical results will be used to validate the CONVINcE testbed results, obtained by measurements.

Table of Contents

Table of Figures

Table of Tables

CONVINcE confidential

1 DOCUMENT HISTORY AND ABBREVIATIONS

1.1 Document history

Version	Date	Description of the modifications
0.1	01.02.2015	Draft of ToC (BTH)
0.2	15.03.2015	First version (BTH)
0.3	30.04.2016	Second version (BTH)
0.4	30.05.2016	Third Version (BTH)
0.5	30.06.2015	Fourth Version (BTH)
0.6	15.07.2016	Final version

1.2 Abbreviations

ASF	Advanced Streaming Format
ARPU	Average Revenue Per User
BS	Base Station
CRAC	Computer Room Air Conditioning
CBSN	Camera-Based Sensor Network
CDN	Content Distribution Network
CG	Cloud Gaming
CPU	Central Processing Unit
CU	Common Unit
CRN	Cognitive Radio Network
DASH	Dynamic Adaptive Streaming over HTTP
DC	Data Center
DE	Deployment efficiency
e2e	end-to-end
EE	Energy Efficiency
ESB	Enterprise Service Bus
FLV	Flash Video
FIFO	First-Input-First-Output
GR	Goodput Ratio
HD	High Definition
HE	Head End
HFC	Hybrid Fiber-Coaxial
HPC	Hardware Performance Counter
IaaS	Infrastructure as a Service
IP	Internet Protocol
IPTV	Internet Protocol TV
ISMA	Internet Streaming Media Alliance
LVS	Live Video Streaming
LLN	Low-Layer Network
MOOP	Multiple Objective Optimization Problem
NFV	Network Function Virtualization
MOS	Mean Opinion Score
MRN	Mobile Radio Network
NVE	Network Virtualization Environment
ODVS	On-Demand Video Streaming
OPEX	Operating Expenditures

OTT	Over The Top
PaaS	Platform as a Service
PCA	Principal Component Analysis
PDU	Power Distribution Unit
PDH	Provisioning Delivery Hysteresis
PMC	Performance Monitoring Counter
QoE	Quality of Experience
QoEW	Quality of Experience per Watt
QoS	Quality of Service
RAN	Radio Access Network
RBS	Radio Base Station
RTP	Real-Time Transport Protocol
RTT	Round Trip Time
SaaS	Software as a Service
SD	Standard Definition
SE	Spectrum Efficiency
SLA	Service Level Agreement
SDN	Software Defined Network
SOTA	State of the Art
UDP	User Datagram Protocol
UPS	Uninterruptible Power Supply
VDN	Video Distribution Network
VM	Virtual Machine
WAN	Wide Area Network

1 INTRODUCTION

The goal of the CONVINcE project is to develop new technological solutions to reduce the energy consumption in IP-based video networks with an end-to-end perspective, from the Head End (HE) through the network and finally to terminals. Four categories of network architectural solutions are considered in CONVINcE, as mentioned in (Popescu A. e. a., 2015). Furthermore, on top of the network architecture, CONVINcE considers the Content Distribution Network (CDN) as well as the Video Distribution Network (VDN) (Popescu A. e. a., 2015).

A set of adequate system requirements has accordingly been defined for CONVINcE, as mentioned in (Popescu A. e. a., 2015). These are requirements that refer to functional aspects, non-functional aspects as well as users and system interfaces. Based on this, the document CONVINcE D1.1.3 has defined the high level architecture and, related to this, the service models for CONVINcE as well as some implementation details and the associated performance optimization.

The document CONVINcE D1.1.4 goes one step forward and reports theoretical models associated with CONVINcE and, related to this, the multi-objective multi-constraint optimization algorithm for performance provision. It is reminded that the target of CONVINcE is to develop new solutions for minimum e2e energy consumption associated with best end-user Quality of Experience (QoE) obtainable in video distribution networks of category IPTV and OTT.

1.1 Scope

The goal of the CONVINcE research project is to develop new solutions for reducing the power consumption in IP-based video networks with an end-to-end (e2e) approach, from the Head End to terminals, also including the core and the access networks. With reference to the Deliverables D1.1.1 “Application Scenarios”, D1.1.2 “System Requirements” and D1.1.3 “High-Level Architecture Design”, the purpose of Deliverable D1.1.4 “Theoretical Models” is to go forward and to provide the theoretical models necessary for the performance evaluation of CONVINcE testbed. At the same time, this Deliverable provides the fundamental concepts for performance evaluation and provision of the testbed. The ultimate goal of CONVINcE is to develop new solutions for minimizing the e2e energy consumption in Video Distribution Networks (VDN) under the condition of good Quality of Experience (QoE) obtained at the end-user. Towards this goal, multi-objective multi-constraint optimization algorithm is suggested to provide the expected performance. Both network perspective and TCP perspective are considered to provide the expected performance for CONVINcE.

It is also reminded that the e2e system includes several networking elements, like:

- Home Network
- Radio Access Networks (RANs)
- Wide Area Networks (WANs) / Internet
- Edge-cloud
- Data Centers (DCs)
- Video Distribution Networks (VDNs)

It is further reminded that three categories of architectural solutions are considered as well as four particular application scenarios. The architectural solutions are

- Non-cloud based architecture
- Edge-cloud based architecture
- SDN/NFV

Whereas the application scenarios considered in our project are

- On-Demand Video Streaming
- Live Video Streaming
- Camera-Based Sensor Networks
- Cloud Gaming

It is also important to mention that the results of Deliverable D1.1.4 are used in defining the Milestones 1.1.2 and 1.1.3 as well as to design the CONVINcE architecture and to evaluate the performance.

1.2 Document Structure

The document structure is as follows:

- Section 2 presents the basic concepts of video distribution, including video coding and compression, video delivery over the Internet as well as short presentation of the CONVINcE high-level architecture
- Section 3 presents the basic concepts of video streaming, including short presentation of the streaming process and strategies
- Section 4 presents video traffic models used to theoretically describe the video signal. A set of popular video models are described, among them Autoregressive models, Markov-Based models, Self-Similar models, Wavelet-Based models and Fluid-Flow models
- Section 5 presents networking and performance aspects relevant for the CONVINcE testbed like, e.g., definition of inter-provider link and use of percentiles.
- Section 6 describes the complexity of developing a QoE-oriented architecture, which is exemplified by a particular example
- Section 7 describes the solution suggested for performance optimization at CONVINcE, which is in terms of minimum e2e energy consumption and best end-user QoE
- Section 8 summarizes the contents and provides concluding remarks

2 VIDEO DISTRIBUTION

The creation, distribution and delivery of video content is a sophisticated process that contains elements related to video acquisition, pre-processing and encoding, content production and packaging as well as distribution to customers. IP networks are usually used for the transfer of video signals.

The treatment of video content is also very complex, and we have a multi-dimensional process with elements like content acquisition, content exchange and content distribution (Monnier R., 2016).

The Internet is undergoing an adaptation process to provide large demands for bandwidth increase, where one of the most important contributors is the video traffic. The appearance of new bandwidth-demanding Over-The-Top (OTT) video streaming services like Netflix, and Skype-like video communications in combination with the growing number of multimedia users has accelerated this process. Further complication is because the Internet has democratized the process of creation, distribution and sharing of video like for instance in the case of YouTube. A very important element is the adoption of more advanced video formats like the Ultra High Definition (UHD), defined and approved by the International Telecommunication Union (ITU), which needs even more bandwidth. UHD is intended to be used for displaying with an aspect ratio of 16:9 and at least one digital input is capable of carrying and presenting native video at a minimum resolution of 3840×2160 pixels (Monnier R., 2016).

Video distribution networks are therefore distribution networks including elements like encoders, transcoders, multiplexers and decoders as well as networking elements used to support professional video distribution and delivery. The basic operations are capturing and initial processing of the video content, initial transportation prior to distribution, primary and secondary distribution and finally delivery to end users.

These networks can be of category terrestrial or satellite or cable television networks. The first category, also known as broadband networks, is based on using the IP technology. These networks are targeting individual users although multicast and broadcast mechanisms can turn them into multicast networks as well.

Basically, the video signal is collected at, e.g., a football stadium and transported in compressed form to some broadcast facility that may be placed in another location. The video is then forwarded to a number of secondary broadcast entities for ultimate transmission to end-users. Examples of distributions networks are IP based TV networks (IPTV) and Over-The-Top (OTT) networks. Other networks like terrestrial or satellite networks can be used as well.

Some of the main characteristics of video streams are the number of frames per second (also known as frame rate), interlaced (e.g., NTSC, PAL, SECAM formats) or progressive video (e.g., LCD TV), aspect ratio (describes the dimensions of video screens and video picture elements), colour space and bits per pixel, video compression method.

An important observation is that today we have a multi-dimensional process for the treatment of video content, where the most important elements are content acquisition, content exchange and content distribution. Furthermore, basic operations done on the video signals along the video distribution chain are video coding and compression, encapsulation, forward error correction, transmission, reception and decapsulation, error correction and decompression. Like in the case of other categories of service provider networks, video over IP distribution systems are expected to provide services with a good mix of simplicity, scalability, security, manageability and cost effectiveness. Service Level Agreement (SLA) requirements for video are used to define the service requirements. These requirements refer to parameters like network delay, network jitter, packet loss, availability, loss recovery.

Today, the IP-based video distribution networks are facing the challenge of introduction of Ultra High Definition (UHD) services (Monnier R., 2016). UHD does not mean the sole increase of the picture resolution: “4K”, which is 4 time the number of pixels of a High Definition (HD) video, or “8K” which is the format Japan intends to introduce in 2020 with 16 times the number of pixels of an HD video. The UHD project also advances improvements in other dimensions. The first one is the temporal dimension with Higher Frame Rates (HFR): 100 or 120 frames per second. A better rendering of the luminance (darker black and brighter white) is also offered as well as a better representation of the colours: Wide Colour Gamut (WCG). As a result of the introduction of UHD services, there is a significant increase of the bitrates, especially for high-resolution videos. Fortunately, progress in video coding tempers this increase. For instance, High Efficient Video Coding (HEVC) allows, at a constant picture quality, in the order of 50% improvement of the bitrate versus the State-Of-The-Art Advanced Video coding (AVC), even more for very high resolutions. Additional gains are further expected in the area of video coding as well (Monnier R., 2016).

Several fundamental aspects regarding the video compression, network transport, error resilience, video quality assessment and a short overview of important video standards like H.264, VC-1 and VP8 are presented in (Bardon, 2016; Bing, 2010). For instance, the H.264 standard overcomes many limitations existent in MPEG-2 motion estimation with improved inter-prediction via fine-grained motion estimation, multiple reference frames, unrestricted motion search and motion vector prediction. Furthermore, H.264 offers improved intra-coding in the spatial domain as well.

Regarding the common display resolutions, they cover a large domain, which depend upon the particular application. For instance, a 15-inch laptop monitor is more suited to 720p than to 1080p or Quarter Common Intermediate Format (QCIF) videos (Bing, 2010). On the other hand, a CIF video coded with a fine quantization level can achieve good video quality on a smartphone. Also, the number of pixels in a video frame is lagging behind digital image resolutions.

Another important aspect is that the large storage and streaming bandwidths involved in video processing dictate the need for compression. For instance, raw videos require massive storage space in the order of hundreds of Gbytes. This in turn demands for the presence of strong compression algorithms, able to reduce such memory spaces, such as reported in (Bing, 2010). This in turn must be evaluated in terms of eventual deterioration of the video quality. Furthermore, given that LCD/plasma screen are rapidly replacing CRT screen, interlaced content must be de-interlaced at playback time. The risk in this case is in form of jittery image in the case of poor interlacing. Thus, de-interlacing is typically performed before video coding (Bing, 2010). The interlaced video format is popular for broadcast and payTV services whereas progressive format is widely adopted by online video portals.

2.1 Video Coding and Compression

Image and video coding and compression have been the area of intensive research activity and development of coding standards to achieve low bit rate for data storage and transmission, while maintaining acceptable distorsion. The general structure of an image coding process contains several elements: image partitioning, transformation (to decorrelate the signal), quantization (to reduce the amount of information required to be stored or transmitted) and entropy encoding. Also, movie files of videos are usually combined with associated audio information and encapsulated in the so-called video containers (Bing, 2010). The problem here is to keep the average percentage increase low.

Major coding predecessors are H.262/MPEG-2, and H.263 (Ohm J-R., 2012). Today, some of the most popular standards that support efficient video delivery storage and provide higher coding efficiency than legacy standards are H.264/MPEG-4 AVC (Advanced Video Coding) used by nearly all services, VC-1 and VP8 as well as the latest international video standard High Efficiency Video Coding (HEVC) (Ohm J-R., 2012; VCODEX, 2016).

H.264 is standardized by the International Telecommunications Union (ITU), whereas VC-1 is standardized by SMPTE 421M and it was initially implemented by Microsoft as Windows Media Video (WMV) 9 for supporting online video streaming. Finally, VP8 is an open source video codec formerly owned by Google but released later as part of the WebM project. Detailed description of key features of the MPEG, H.26x, VC-1, HEVC and other popular video coding standards is provided in (VCODEX, 2016; Ohm J-R., 2012). Among others, it has for instance been observed that HEVC encoders can achieve equivalent subjective reproduction quality like H.264/MPEG-4 AVC encoders using 50% less bit rate on average (Ohm J-R., 2012).

Detailed description of standards for video coding and compression is provided in (Bing, 2010; Ohm J-R., 2012; VCODEX, 2016). Main characteristics of these standards are as follows (VCODEX, 2016).

Standard	Key features
MPEG-1	Developed for audio/video storage on CD-ROM; motion vectors coded losslessly
MPEG-2	Supports video on DVDs, standard/high definition TV; supports scalable extensions
MPEG-4	Supports video on low-bit rate multimedia on mobile platforms and Internet and also object-based coding
H.261	Developed for video conferencing over ISDN; support for CIF and QCIF resolutions
H.262	Standardized as MPEG-2 Part-2
H.263	Improved quality compared to H.261 at lower bit rate
H.264 AVC	Significantly improved picture compared to H.263, at low bit rates, but increased computational complexity
H.265/HEVC	Support for ultra HD video; greater flexibility in prediction modes; support for parallel processing
VC-1	Adaptive block transform for easier implementation in hardware devices
VP8	Developed to operate in low bandwidth environment such as Web video
VP9	Basic structure simpler than VP8

Table 1: Features main standards for video coding and compression (VCODEX, 2016)

2.2 Recent Developments

There has been a phenomenal growth over the last years in video applications and transport, e.g., an ever-increasing number of users upload and download video signals using sites like YouTube. Applications like HD DVD and video streaming (Netflix, Hulu) are increasing in popularity as well. Video calling over the Internet (by using, e.g., Skype, Facetime) is also increasing in popularity. Video conferencing systems like Cisco TelePresence and WebEx are popular today in different business and organizations. The consequence of this growth in combination with the growing number of multimedia users is that recent forecasting studies done by Cisco predict that all forms of video will increase to 80-90% of global consumer traffic by 2017 (CISCO).

However, the design of robust and reliable networks and services to provide expected performance in terms of minimum e2e energy consumption and best end-user QoE is becoming increasingly difficult. For doing this, one needs to first have a detailed understanding of the network and traffic characteristics. Furthermore, from the viewpoint of a network service provider, the demands on the network are not easily predictable. This aspect must be also correlated to the demands to accurately estimate the e2e network performance. This in turns demands for accurate traffic models able to capture the statistical characteristics of traffic in the network as well as good understanding of the process of video streaming.

2.3 CONVINcE High-Level Architecture

High-level architecture of a system is defined to be a general-purpose architecture that describes the system considered for study. The diversity of video streaming network solutions and components indicates the complexity existing in the design and development of architectural solutions for CONVINcE and the associated optimization solutions. A three-layers high-level architecture is used in CONVINcE, which is shown in the figure below: Video Distribution Network (VDN), Content Distribution Network (CDN) and low-layer network (LLN). LLN refers to networking elements at layers L2 (Data Link Layer) and L3 (Network Layer) in the TCP/IP protocol stack (Figure 1) (Monnier R., 2016).

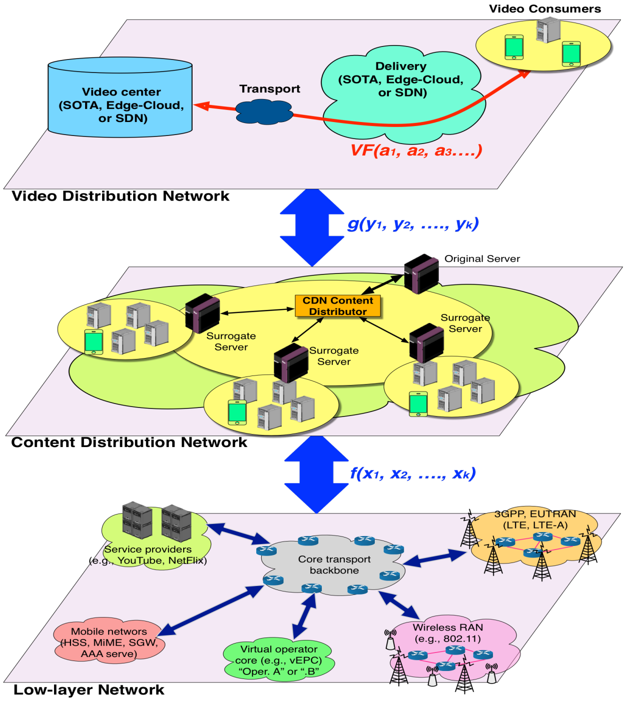

Figure 1: CONVINcE high-level architecture (Monnier R., 2016)

Three categories of low-layer network architectural solutions are considered in CONVINcE. These are the non-cloud based architecture, the edge-cloud based architecture and the SDN/NFV-based architecture. The non-cloud based architecture is the State-Of-The Art (SOTA) architecture and it is used as the reference to measure improvements in energy saving It relies on a classical headend making use of dedicated equipment for encoding and transcoding. The focus in the project is laid on the edge-cloud based architecture.

VDN is highly related to the formal representation of video flows considered in different CONVINcE application scenarios. A video flow is defined to be an end-to-end (e2e) data transmission from the video streaming source side (e.g., headend) to the video consumer side (e.g., terminal). This process is highly related to the particular application scenario and it can be described as a function dependent on elements like the application (e.g. On-Demand Video Streaming, Live Video Streaming), the network solution (e.g. edge-cloud based, SDN) and video flow elements (e.g. Video Data Center, consumer) (Monnier R., 2016).

CDN is basically a large system of geographically distributed specialized cache servers deployed near access providers in Data Centers across the Internet, which are combined with the use of a special routing code to redirect the media requests to the closest servers. High-level network intelligence is used to improve the performance of delivering media content over the Internet. Requests for content are typically directed to servers that are optimal in some way. Media delivery can accordingly be done with reference to different parameters used in the process of optimization like highest possible throughput, minimum power, minimum cost.

The concept of distributing the content to cache servers located close to end users results in better performance with regard to maximizing the bandwidth, minimizing the latency and jitter, improving accessibility as well as better balance between the cost for content providers and the QoE for customers. The benefits of using CDN are substantial to end-users, content and application owners as well as network to service providers.

Finally, another important architectural element is the coordinating middleware, which coordinates the operation of individual architectural elements towards minimizing the e2e power consumption combined with the best possible QoE performance obtained at the end user, like for instance the OpenStack API and the ApacheServiceMix [8, 9].

2.4 Video Delivery over the Internet

In the following, we consider the process of video delivery at the VDN level. There are two categories of Internet-based video delivery services. These are the “Internet Protocol TV (IPTV)” and the “Over the Top (OTT)” services. The fundamental difference between them is that, whereas the OTT distribution service is carried out over an un-managed network by using intelligent technology available at end-points to provide the best possible quality offered by the IP best effort networking approach, the IPTV dissemination occurs over a managed proprietary network by using a variety of protocols like, e.g., Internet Group Management Protocol (IGMP) and Real Time Streaming Protocol (RTSP) to a user Set Top Box (STB) or computer. These two categories of services are quite different with regard to many aspects like, e.g., business model, quality of service, quality of content, ownership, cost, and others. The table below provides several key differences between OTT and IPTV (Mediaentertainmentinfo).

From the service provider point of view, the main challenge in an IPTV system is to provide high quality service at a minimum cost by using the existent delivery networks. Bandwidth reservation and admission control need to be available in this case to guarantee the QoS requirements for video. This networking architecture started initially with a central architecture, where a single server was used for content delivery. Later on, the server was replaced by a server farm or cluster. The next phase was the introduction of a hierarchical architecture, which distributes the contents from a set of cache servers. The content was therefore replicated into a set of servers distributed at different geographic places. Finally, the latest development is the cloud-based architecture, where the content is placed in the cloud. Important questions are in this case regarding storage, broadcasting, buffering, QoS and QoE for both real and scheduled IPTV services.

On the other hand, the OTT content delivery has enabled many new actors to enter, establish and compete with traditional media players. There is no bandwidth reservation in this case with the consequence of no QoS provisioning. The service is able to deliver video by adapting the video to meet the available bandwidth limitations as well as the best effort nature of the IP network. These aspects, combined with the last mile problem has led to the development of adaptive streaming protocols (e.g., Microsoft Smooth Streaming, Apple HTTP Adaptive Bitrate Streaming, Adobe HTTP Dynamic Streaming) that dynamically monitor the end user bandwidth and associated video performance and subsequently optimizes the video quality by switching between lower or higher quality streams (Blair A., 2011). OTT has relatively less technical and financial requirements. However, it is still limited by quality limitations and catalog content such as today this is only considered as complementary service to regular cable. Furthermore, broadcast OTT providers work today closely with operator CDNs to provide QoS guarantees across well-managed IP networks.

OTT has the advantage of providing cheaper service model without the need for heavy investments. OTT has also been instrumental in driving multi-screen, multi-platform content convergence. All in all, OTT is today not yet in a position to disrupt or to replace the service providers but future perspectives show great potential. A comparative list of OTT and IPTV characteristics is shown in table 2.

Category	OTT	IPTV
Content delivery	Streams; Uses open Internet, un-managed network	Broadband channels; Uses dedicated, managed network
Network type	Delivered from content provider/aggregator to the viewer using open network; use of CDN	Closed, proprietary network, accessed via specific ISP
Network relationship	Without need for intervening carriage negotiations or infrastructure investments	Services are delivered on optimized and custom high bandwidth network
QoS	Net guaranteed, best effort conditions	High quality, reliable network with QoS control
Service examples	Downloads; YouTube, Netflix, Hulu	VoD, IPTV services, e.g., U-Verse (AT&T)
Delivery protocol	Delivered over HTTP/TCP	IPTV uses Transport Stream (TS) technology, RTP over UDP
Content catalogue	Widely used for VoD	Primarily used for VoD and real time content delivery
Content type	Not premium in nature	Premium content

Routing topology	Unicast (HTTP), Simulated Multicast (UDP/TCP)	Multicast
Service category	Complementary Service	Main Service, similar to cable/satellite TV services
Major platform players	Online Video Platforms (OVP), e.g., Akamai, L3, Kaltura	Telecom Service Providers (TSP) and IPTV platform vendors (Cisco, Ericsson)
Key challenges	Low quality of service, absence of live broadcast, unicast delivery model	Expensive, heavy investment in bandwidth and infrastructure
Key benefits	Low cost, flexible model, easy to manage and operate	High QoS and QoE, monitoring and control

Table 2: OTT and IPTV characteristics (Mediaentertainmentinfo)

3 VIDEO STREAMING

Basically, streaming means that people are listening to music or watching video in real time, instead of first downloading a file to computer and listening or watching it later. The client starts the content playback a few moments after it begins receiving the content from the server. This method is different from the classical method of normal file download where the entire file is first downloaded before starting up the playback.

Streaming media refers to multimedia content (video or audio) sent in compressed form over the Internet and played in real time instead of first being saved to the hard drive. That means, a Web user does not need to wait to download a file and to play it. The data stream is played as it arrives.

There are several different delivery methods: traditional streaming; progressive download; adaptive video streaming (Zambelli A., 2009). A good example of traditional streaming is the Real-time Streaming Protocol (RTSP), which is a stateful protocol where the users interact with the media content provider. RTSP is used together with the Real-time Transport Protocol (RTP). Today the move is however towards HTTP-based streaming where the media delivery is adapted to the Internet instead of trying to adapt the entire Internet to streaming protocols.

The progressive download streaming method is a method based on downloading from a HTTP Web server. The player client allows here the media to be played back while the file download is still in progress. This means the user can, e.g., pause the streaming, waiting for the download to finish, allowing so a smooth playback when the user decides to play the media. The drawback of this method is that it is not perfect and not very flexible in selecting parameters for downloading.

A better method for video streaming is the HTTP-based adaptive video streaming, which is a hybrid method of progressive download and streaming. Here the video is cut into short segments and encoded at the specific delivery format and rate. The segment length may vary from implementation to implementation, but they are typically couple seconds long. It is the adaptive properties of this streaming solution that raises the popularity of the method. Adaptive facilities like available network resources (e.g., bandwidth), device facilities (e.g., display resolution, available CPU) and current streaming conditions (e.g., playback buffer size) may be used in the streaming process. With these adaptive facilities a major benefit is that the video streaming client over time can adjust its selection of encoded segments, based on e.g. communication link throughput capability. A good example of HTTP-based adaptive video streaming standard is MPEG-DASH, which stands for MPEG Dynamic Adaptive Streaming over HTTP, also known as ISO/IEC 23009-1 (Sodagar I., 2011). The International Organization for Standardization (ISO) finally published MPEG-DASH in April 2012.

Basically, live streaming requires some form of source media (e.g., video camera, audio interface), an encoder to digitize the content, a media publisher and a CDN to distribute and deliver the content. Furthermore, a user needs a player entity, which is software that uncompresses and sends video data to display and audio data to speakers. The associated storage size is calculated from the product of streaming bandwidth and media length. Examples of streaming technologies are Adobe Flash, Apple QuickTime, Microsoft Windows Media and Silverlight. Streaming technologies use compression to shrink the sizes of audio and video files so they can be retrieved and played by remote viewers in real time. The codecs are usually independent of the streaming technology that implements them.

There are several transport protocols that can be used in video streaming. These protocols are Hypertext Transport Protocol (HTTP, this is however not a streaming protocol, but used to distribute small files), Real-time Streaming Protocol (RTSP), Real-time Transport Protocol (RTP), Real-time

Transport Control Protocol (RTCP), Real Time Messaging Protocol (RTMP) and Real Data Transport (RDT). Details about these protocols as well as other relevant information are provided in different books on computer networking.

A broadband speed of at least 2 Mbps is recommended for streaming standard definition video without experiencing buffering, especially live video, e.g., to Apple TV, Google TV or Sony TV Blue-ray Disc Player. Under such circumstances, streaming media storage sizes of about 128 MB is needed for one hour of video encoded at 300 kbps (Storage, 2016). In the case of live video with 3000 viewers, then the media storage sizes increases to almost $2*10^6$ MB (Storage, 2016).

3.1 Streaming Process

Figure **2** shows an example of video streaming process encountered at video download (Rao A., 2011).

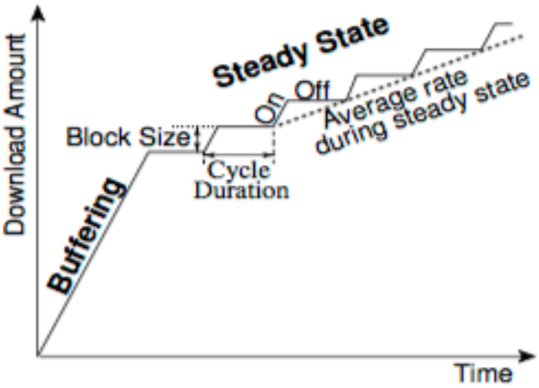

Figure 2: The process of video download (Rao A., 2011)

As it is observed in the figure, there are two phases in a video streaming process. These are the buffering phase and the steady state phase. In the first phase, the data transfer rate is limited by the end-to-end available bandwidth. The slope of the line gives this during the buffering phase. The video playback starts up as soon as sufficient data is available in the buffer. In other words, the video playback process does not wait till the end of buffering phase.

The steady state phase is a state where the average download rate is (little bit) larger than the video encoding rate. The ratio of average download rate (during the steady state phase) to video encoding rate is denominated as accumulation ratio (Rao A., 2011). This means that an accumulation ratio of at least one is desirable. This is because an accumulation ratio of less than one would mean that the video playback might get interrupted due to empty buffers. Further, an accumulation ratio larger than one means also that the amount of video data present in the video buffers increases during the steady state phase, which in fact improves the resilience to transient network congestion.

Data downloading in the steady state phase is done by periodically transferring blocks of video content, with the consequence of ON-OFF cycles observed in the figure. During the ON period a block of data is transferred under the condition of an end-to-end bandwidth limited by, e.g., TCP. On the other hand, the TCP connection is closed during the OFF periods. In other words, the slope of data downloading during the steady state indicates the end-to-end existing bandwidth. The authors of (Rao A., 2011) call the amount of data transferred in one cycle as the block size.

The buffering phase ensures that the player has a sufficient amount of data to compensate for the variance in the end-to-end available bandwidth during video playback. The reduced transfer data in the steady state phase ensures that the amount of video content does not overwhelm the video player while keeping the amount of buffered data during the buffering phase constant or increasing.

The buffer size needed in the steaming process for a single user and file is given by the product of streaming bandwidth (in bit/s) and media length (in seconds) (Gelman A.D., 1991).

$$Storage\_size \ \ (MBytes) = \frac{Media\_length \ \ (\sec) \ \ * \ \ Bit\_rate \ \ (bit/\sec)}{8 * 1024 * 1024}$$

CONVINcE confidential

Moreover, in the case of using unicast transport protocol, an increased number of receivers for the same document increase accordingly the storage size.

3.2 Streaming Strategies

As mentioned above, data streaming involves two phases: buffering phase and steady state phase. There are three strategies that can be used, e.g., in Netflix and YouTube (Rao A., 2011). These are as follows:

- No ON-OFF cycles: all data is transferred in this case during the buffering phase, with the consequence that there is no steady state phase; that means, we simply have a file transfer session in this case, but the risk does exist for overwhelming the player. The advantage is that this is simple, but measures must be taken to avoid the risk of overwhelming the player.

- Short ON-OFF cycles: this is a simple strategy, where measures must be taken such that the client is not overwhelmed by the data sent by server. The goal in this case is to maintain an accumulation ratio slightly larger than one.

- Long ON-OFF cycles: large amounts of data are transferred in a cycle, which demands for careful buffer dimensioning.

The table below shows a comparison of the three streaming strategies (Rao A., 2011).

Strategy	No ON-OFF	Long ON-OFF	Short ON-OFF
Engineering complexity	Not required	Explicit support at application layer	
Receive buffer occupancy	Large	Moderate	Small
Unused bytes on user interruption	Large amount	Moderate amount	Small amount

Table 3: Comparison of streaming strategies (Rao A., 2011)

A general observation is that the three streaming strategies may have different impacts on the network loss rate. This is am important element, which must be considered in the development of streaming strategies. This is because the wasted bandwidth associated with a particular streaming strategy may be different, depending upon the particular streaming and networking conditions.

4 VIDEO TRAFFIC MODELS

A good video traffic model is expected to be able to capture characteristics of video sequences and predict the network performance. A large number of video traffic models have been suggested over the last 20 years to capture and to describe different statistical properties of video traffic like encoding formats, creating synthetic loads and others. A short description of the main characteristics of the most popular methods for video encoding and compression is presented above, in Section 2.1. Related to this, a general comment is that the standards have been designed such as to provide the developers of encoders and decoders as much freedom as possible to customize their implementations. This is essential in order to let the particular standards be adapted to a wide variety of platform architectures, resource constraints and application environments.

There are two main characteristics of video traffic that are relevant for predicting the network performance for video transfer (Tanwir S. and Perros H., A Survey of VBR Video Traffic Models, 2013). These are the distribution of frame sizes and the Autocorrelation Function (ACF) that captures common dependencies among frame sizes in VBR video traffic. The problem of capturing the ACF structure of VBR video traffic is challenging because this traffic shows both Short-Range Dependent (SRD) and Long-Range Dependent (LRD) statistical properties. Detailed description of these statistical properties is presented in (Popescu A., Traffic Analysis and Control in Computer Communications Networks, 2008).

Mathematically, the difference between SRD and LRD processes can be stated as follows (Cox D.R., 1984). In the case of SRD process with the degree of aggregation m, the following properties are representative:

- The mean value $E(X^m)$ approaches second order pure noise as $m \to \infty$
- $Var(X^m)$ is asymptotically of form $\dfrac{Var(X)}{m}$ for large m
- $\sum Cov(X_n, X_{n+k})$ is convergent
- Spectrum $S(w)$ is finite at $w = 0$

On the contrary, LRD processes with the degree of aggregation m are characterized by the following properties:

- The mean value $E(X^m)$ does not approach second order pure noise as $m \to \infty$
- $Var(X^m)$ is asymptotically of form $m^{-\beta}$ for large m
- $\sum Cov(X_n, X_{n+k})$ is divergent
- Spectrum $S(w)$ is singular at $w = 0$

Good video traffic models for real video sequences should satisfy several criteria (Tanwir S. and Perros H., A Survey of VBR Video Traffic Models, 2013). First, the particular model should match basic statistical characteristics like, e.g., probability distribution function pdf, mean value, variance, peak, autocorrelation and coefficient of variation of the bit rate (Alheraish A.A., 2004). Second, the synthetic video sequence must be similar to the real video sequence. The third criterion is that the model must be simple and the synthetic trace generation must have low computational complexity. Finally, the last criterion is that the model must be able to characterize a wide range of video sources with low to high motion activity.

As a general comment, the existent video traffic models have been developed independent of each other, and today there exist six general classes of Variable Bit Rate (VBR) traffic models (Tanwir S. and Perros H., A Survey of VBR Video Traffic Models, 2013):

- Autoregressive (AR) models
- Markov process (MP) models
- Self-similar (SS) models
- Wavelet (WV) models
- Fluid flow models
- Other models

Short description of these models and their properties is as follows (Tanwir S. and Perros H., A Survey of VBR Video Traffic Models, 2013; Popescu A., Traffic Analysis and Control in Computer Communications Networks, 2008).

4.1 Autoregressive (AR) Models

The Autoregressive (AR) process is a process where the current value is the function of a weighted linear combination of past values

$$x(n) = \sum_{i=1}^{p} a_i x(n-i) + e(n)$$

where $x(n)$ are prescribed correlated random numbers, the parameters $a_1, a_2, ... a_p$ are real numbers (AR coefficients) and p is the order of the AR process. The sequence $e(n)$ is a white noise that consists of independent and identical distributed (i.i.d) random variables and provides the AR the stochastic nature. The residuals are uncorrelated and often assumed to be normally distributed with zero mean and variance σ^2.

There exist minor changes in the way predictions are computed according to which several model variations can be developed. Basically, when the model depends only on the previous outputs of the system, it is referred to as an auto-regressive model, also known as a Moving Average Model (MAM). These are models that depend upon both inputs and outputs, with the aim of predicting the current output.

The main advantage of AR is that it is simple and it requires only a few parameters. Specifically, $x(n)$ represents the bit rate or the size of the coded video during the n-th frame. While, $e(n)$ is the Gaussian process with zero mean and variance σ^2. Finally, the parameters a_i represent the lag i autocorrelation of the successive frame rates.

An important advantage of the AR models is that they capture the autocorrelation behaviour of the compressed video sources, which is essential in modeling compressed video sources. The coefficients of these models can be simply estimated from empirical data.

The drawback is related to the existing difficulties to find appropriate AR models able to capture different statistical characteristics of video signals (Tanwir S. and Perros H., A Survey of VBR Video Traffic Models, 2013). In other words, no single AR video model is suitable for all video sequences and purposes.

4.2 Markov-Based Models

Markov models are very popular models that model the activities of a process by using a finite number of states. The model accuracy increases linearly with the number of states used in the model. The drawback is that the model complexity also increases proportionally with increasing the number of states.

The basic property of the Markov model is the so-called Markov property, which states that the next (future) state $X_n + 1$ depends on the current state X_n only, and not any other additional previous state X_i, where $i < 1$. In other words

$$P\left[X_n \middle| X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, ...X_1 = i_1\right] = P\left[X_n = i_n \middle| X_{n-1} = i_{n-1}\right]$$

The set of random variables referring to different states $\{X_n\}$ is called the Markov chain. There are two categories of Markov chains, namely continuous and discrete. The state transitions of the system under study happen in continuous time and in discrete time, respectively.

Markov processes and chains are typically used to model other processes like Poisson, Bernoulli, Gamma and AR. Video models based on a Markov process use states to represent ranges of bit rates of a video sequence or ranges of frames or Group Of Pictures (GOP) sizes of a video sequence.

Today, there is a wide variety of Markov-based models suggested for modeling the VBR video traffic (the so-called Markov-modulated models) like, e.g., Markov-modulated Poisson Process (MMPP), Markov-modulated Bernoulli process (MMBP), Markov-modulated renewal process (MRP), Markov-modulated AR model, Markov-modulated Gamma model (Tanwir S. and Perros H., A Survey of VBR Video Traffic Models, 2013).

An important class of models is the MMPP models. Short description of MMPP is as follows.

A Markov-modulated Poisson process (MMPP) is a doubly stochastic Poisson process, which is a Poisson process with an intensity that is changing in time according to another random process. Figure 3 shows an example of MMPP process of order 2 - MMPP(2). There are two states involved in this process, and the transitions between the two states are not associated with events. The events occur with rate λ_1 in state 1 and with rate λ_2 in state 2. For instance, a MMPP(2) process can be a job arrival process alternating between high arrival period (state 1) and low arrival period (state 2).

An example of queueing system used to serve a MMPP(2) process is presented in Figure 4, which shows the state diagram of an MMPP(2)/M/1 queue. Here the jobs arrive according to a MMPP(2) process and their service demand has an exponential distribution with rate μ.

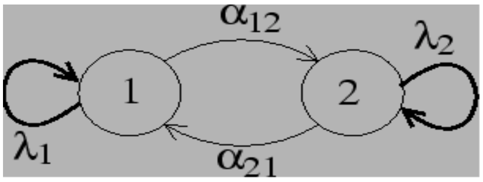

Figure 3: MMPP process

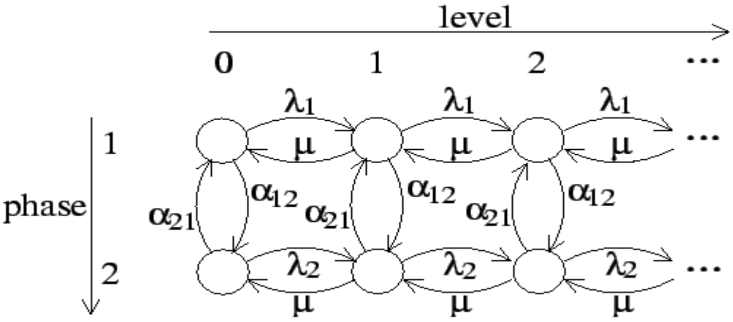

Figure 4: MMPP(2)/M/1 process

The Markov models seem definitely to be more accurate than the AR models (Popescu A., Traffic Analysis and Control in Computer Communications Networks, 2008). Further, there are two important factors that distinguish these models from each other (Tanwir S. and Perros H., A Survey of VBR Video Traffic Models, 2013). The first factor is regarding the modulated process, which is dependent on the frame size distribution of the video traffic. The second factor is regarding how the state space of the Markov process is determined and how the transition probabilities are calculated.

4.3 Markov-Modulated Rate Processes

Another important class of models is the Markov-Modulated Rate Processes (MMRP) (Arlos P., 2003)). This process was initially suggested by (Mannersalo P., 2002). This model can exhibit multi-fractal properties. In its simplest form, the model is based on the multiplication of independent rescaled stochastic processes, which are piecewise constant. That means, we have a multiplication rather than adding re-scaled versions of a mother process, with the result of a process with novel properties, whose properties are best understood in a multiplicative analysis and not additive one.
The MMRP are very popular in fluid flow analysis. This is because of simplicity in analysis, with straightforward results (Arlos P., 2003; Fiedler M, 2016).

4.4 Self-Similar Models

A self-similar process is associated with “fractals”, which are objects with unchanged appearances over different scales. The concept of fractals includes, besides the geometrical meaning, statistics as well as dynamics. This means, there are fractal processes in different dimensions, e.g., geometrical, statistical and dynamical (Popescu A., Traffic Analysis and Control in Computer Communications Networks, 2008). Examples of geometrical fractal processes are the Cantor set, Sierpinski triangle and Koch curve. In the case of statistical fractals it is the probability density that repeats itself on several scales like, e.g., Pareto’s law (economics), Zipf’s law (linguistics) and Lotka’s law (sociology).
Self-similar processes can be defined in continuous-time and in discrete-time. Also, there are two classes of self-similar processes in discrete-time, which are the exactly self-similar processes and the

asymptotically self-similar processes. The definitions of these processes are as follows (Popescu A., Traffic Analysis and Control in Computer Communications Networks, 2008).

4.4.1 Exactly self-similar processes

A process X is said to be exactly (second-order) self-similar process with parameter β where $0 < \beta < 1$ if the following conditions are fulfilled:

- Variance $Var\left[X^{(m)}\right] = \dfrac{Var\left[X\right]}{m^{\beta}}$
- Autocorrelation $R(k, X^{(m)}) = R(k, X)$

The parameter β is related to the Hurst parameter H by $\beta = 2(1-H)$ and m is the degree of aggregation. For stationary ergodic processes $\beta = 1$ and for self-similar processes β may take very low values.

The Hurst parameter H is a parameter that indicates the degree of self-similarity in a particular process, i.e., the degree of persistence of the statistical phenomenon. A value of $H = 0.5$ indicates the lack of self-similarity, whereas large value for H (close to 1.0) indicates a large degree of self-similarity or Long-Range Dependence (LRD) in the process. This means that, for a LRD process, an increasing (or decreasing) trend in the past or the presence of a process implies, with a large probability, an increasing (or decreasing) trend in the future.

4.4.2 Asymptotically self-similar processes

A process X is said to be asymptotically (second-order) self-similar process if, for k large enough

- Variance $Var\left[X^{(m)}\right] \sim \dfrac{Var\left[X\right]}{m^{\beta}}$
- Autocorrelation $R(k, X^{(m)}) \rightarrow R(k, X)$

as $m \rightarrow \infty$.

It is observed that, for both classes of self-similar processes, the variance of X^{m} decreases more slowly than $\dfrac{1}{m}$ for $m \rightarrow \infty$ and the aggregated processes are indistinguishable from the process itself with respect to the first and second order statistics. This is to compare with the case of stochastic processes where the variance decreases proportional to $\dfrac{1}{m}$ and approaches zero for $m \rightarrow \infty$ (consistent with white noise, which has uniform power spectrum).

4.4.3 Popular self-similar traffic models

Self-similar traffic models can be used to model both single-source models and aggregate traffic models. Several popular models are, for single-source models, the Pareto distribution and the lognormal distribution and, for aggregate traffic models, the ON-OFF model, fractional Brownian model, fractional Gaussian model, Fractional Autoregressive Integrated Moving Average (FARIMA) model, and Chaotic Deterministic Map (CDM). Out of these, the most popular model for video is perhaps the FARIMA model, in spite of some limitations like difficulties to capture the SRD structure of video traffic. On the other hand, the FARIMA model has been proved to be able to capture very well the LRD of Autocorrelation Function (ACF) as well as the marginal Cumulative Distribution Function (CDF) of the frame sizes. Detailed description of these models is provided in (Popescu A., Traffic

Analysis and Control in Computer Communications Networks, 2008; Tanwir S. and Perros H., A Survey of VBR Video Traffic Models, 2013).

4.5 Wavelet-Based Models

Techniques for modeling and generation of video traffic based on using the wavelet transform have been quite popular (Tanwir S. and Perros H., A Survey of VBR Video Traffic Models, 2013; Popescu A., Traffic Analysis and Control in Computer Communications Networks, 2008). This method is based on using wavelet analysis, which is based on a decomposition of the signal by using a family of basic functions. This includes both a high-pass wavelet function (that generates the detailed coefficients) and a low-pass scaling filter that produces the approximation coefficients of the original signal.

Basically, the wavelet-based method is similar to the variance-time analysis but allows analysis in both time and frequency domains. The wavelet estimator is based on the Discrete Wavelet Transform (DWT), which has the advantage of both aggregation-based and Maximum Likehood Estimation (MLE) and also avoids their drawbacks.

The underlying concept of the DWT is the so-called Multi-Resolution Analysis (MRA), which consists of splitting a sequence $X(t)$ into a (low-pass) coarse approximation $A(t)$ and a number of (high-pass) details $D_j(t)$. These are associated with the set of scaling coefficients and wavelet coefficients, respectively (Abry P. and Veitch D., 1998).

$$X(t) = A_j(t) + \sum_{j=1}^{J} D_j(t)$$

The focus is placed on details that are described by the wavelet coefficients. When going from high resolution to lower resolution, the MRA gives rise to details at larger time scales. This can be interpreted in the frequency domain as band-pass filtering, going from high to low frequencies with constant relative bandwidth. The wavelet constant relative bandwidth manages to provide a perfect match (Popescu A., Traffic Analysis and Control in Computer Communications Networks, 2008).

The wavelet basic functions absorb the long-range and short-range dependencies by differencing the averages at all time scales, hence the wavelet coefficients are short-range dependent. This makes it possible to model the wavelet coefficients as independent random variables without losing much information (Tanwir S. and Perros H., A Survey of VBR Video Traffic Models, 2013).

As a general comment, the wavelet-based models are good at modeling the co-existence of SRD and LRD behaviour in video traffic. This is because of the key advantage (when using wavelets) in form of ability to significantly reduce the complex temporal dependence such that the wavelet coefficients only possess short-range dependence. On the other hand, a serious drawback is that this requires several trials to determine the optimal number of levels of decomposition for a particular type of video.

Besides of describing SRD and LRD properties, the feasibility of using wavelet coefficients for detecting video delivery problems (e.g. short interruptions of traffic flows, implying the risk of freezes during video playback) has been shown in (Shaikh J. F. M., 2012; Shaikh J., 2015). Indeed, irregularities in the Wavelet spectrum indicate time scales that are affected by disturbances. If, for instance, the 1 s-scale shows to be disturbed, and the video was buffered for 10 sec, such disturbances are likely to have no visible impact on the video playout. However, if the disturbances showed up on the 8 sec scale, the 10 sec buffer would not be sufficient to avoid freezes.

4.6 Fluid-Flow Models

The fluid flow model is a theoretical model used for performance evaluation, which operates on flows. By flow we mean a stream of bits or packets that belong to the same application session. This means the focus in this case is not laid on individual packets or bits but on data flows or, in our case, flows of streams (e.g., video flows). In other words, a fluid flow cannot handle individual bits or packets but streams of bits or packets.

This model has been proven to be a very powerful modeling paradigm in a wide range of applications like, e.g., performance of a network switch, router, IEEE 802.11 protocol, P2P file sharing and others. The main advantages of this model are simplicity and ability to capture the network dynamics. The model captures the key characteristics that determine the performance of communication systems and, on the other hand, they remain mathematically tractable. This model is today very useful for

modeling streaming media in the presence of time-varying bandwidth. The problem in this case is that the Internet provides only best-effort service, which means that the packet streams generated by streaming media applications get distorted by fluctuations in throughput in the Internet and other networking entities. These fluctuations may become significant over the duration of a typical streaming application. To cope with these distortions, play-out buffers are used to temporarily store packets such as to reproduce the initial data stream, but with a fixed delay offset, i.e., the so-called video streaming procedure. Further, one should also arrange/dimension things such as the play-out buffer does not get empty and avoid so performance deterioration due to stalling process.
A simple theoretical model of fluid flow is as follows (Bosman J.W., 2012). A fluid queue can be considered to be a large tank, typically assumed to have infinite capacity. This tank is connected to other tanks, such as the fluid is poured from tank to tank. Further, an operator controls the pipes and pumps, controlling so the rate at which the fluid pours into the buffer and the rate at which the fluid leaves the buffer, as it is observed in Figure 5.

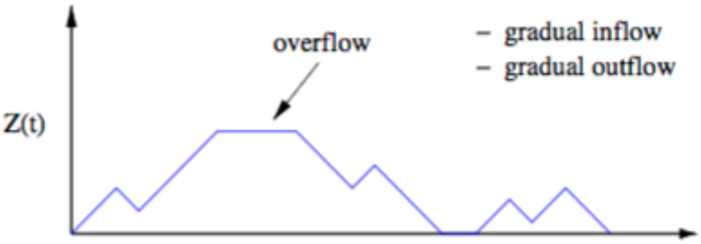

Figure 5: Typical fluid queue behaviour (Bosman J.W., 2012)

The fluid-flow queueing system can be modelled with the help of a time-varying fluid model $X(t)$

$$\frac{dX(t)}{dt} = \begin{cases} r_i \quad if \quad X(t) \rangle 0 \\ \max(r_i, 0) \quad if \quad X(t) = 0 \end{cases}$$

The operator is a continuous-time Markov chain and it is usually called the environment or background process. Given that the process $X(t)$ represents the fluid level, it can take only non-negative values. This model is in fact a particular type of piecewise deterministic Markov process (Bosman J.W., 2012).
A very interesting study on fluid-flow models is reported in (Bosman J.W., 2012). The objective in this case is to understand the performance implications of the play-out buffer settings for streaming applications over unreliable networks like the Internet. Figure 6 shows the tandem model of fluid buffers for video streaming at rate R_{play} through an IP network with variable bit rate.

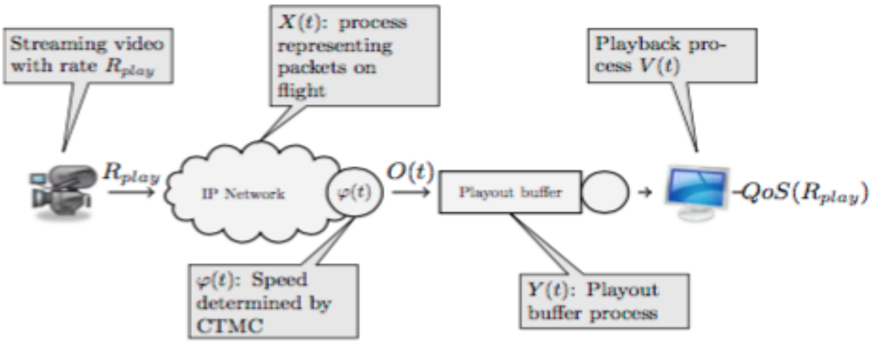
Streaming video with rate R_{play}
$X(t)$: process representing packets on flight
Playback process $V(t)$
R_{play}
IP Network
$\varphi(t)$
$O(t)$
Playout buffer
$QoS(R_{play})$
$\varphi(t)$: Speed determined by CTMC
$Y(t)$: Playout buffer process

Figure 6: Fluid flow model for media streaming (Bosman J.W., 2012)

The fluid queue is used to model the congestion. The input rate is in this case constant whereas the output rate is determined by a stochastic process modelled as a Continuous Time Markov Chain (CTMC) (Bosman J.W., 2012). In other words, CTMC is used to represent the dynamics of IP network that causes congestion and fluctuations in available bandwidth. CTMC determines the actual transmission rate the network. By playing with parameters like the level of play-out buffer, the system can be dimensioned such as to obtain certain probabilities for e.g., video stalling during play-out.

In reference (Fiedler M., 2014)) it is shown that disturbances on a mobile link can be described by a CTMC, and the (average) off-time plays a major role for the risk of freezes, which cannot be fully compensated by buffering. Indeed, off times have shown to have a strong impact on QoE (Shaikh J. F. M., 2012). Based on above described two-state CTMC model (Fiedler M. S. J., 2014) connections have been provided between the parameters describing the on-off model (average on and off times) and user ratings, which has been further developed in (Ickin, 2015). With help of these models, an immediate closed-form relationship between the dynamics of the channel and the resulting QoE can be established. In particular, the (average) off time plays a major role for the viability of the QoE provided to end-user. (Shaikh J. F. M., 2012; Shaikh J., 2015)

Encouraged by the usefulness of the CTMC model and availability of closed-form expressions (Fiedler M, 2016) advanced the concept of sustainable throughput as the maximal throughput that can be handled by a system (like a mobile or wireless link) with negligible impact on QoE (Fiedler M, 2016) also provides closed-form expressions for sustainable throughput based on the dynamics of the wireless channel, given a maximal QoE disturbance. This approach is also able to incorporate MMRPs that are able to mimic self-similar behaviours, as described in Section 4.3.

4.7 Other Models

Besides the above-mentioned models, there are several other models that are not captured in these categories. Several popular models are the $M/G/\infty$ model and the Transform-Expand-Sample (TES) model (Tanwir S. and Perros H., A Survey of VBR Video Traffic Models, 2013).

The $M/G/\infty$ model is advantageous in modeling the video signals because of the existent facilities to capture the correlation. In essence, the $M/G/\infty$ process is the stationary probability function of the number of busy servers in a $M/G/\infty$ queue. By varying the service time distribution G many forms of time dependence can be obtained, with the consequence that the $M/G/\infty$ models become good candidates for modeling many types of correlated traffic.

Results reported in (Krunz M. and Makovski A., 1998) indicate that only the $M/G/\infty$ model is capable to provide acceptable predictions of the queueing performance. Another advantage is in terms of lower number of computations to generate a $M/G/\infty$ trace, compared to, e.g., FARIMA.

TES processes are designed to simultaneously fit the marginal distribution and the autocorrelation function of the empirical data (Tanwir S. and Perros H., A Survey of VBR Video Traffic Models, 2013). The advantage is that the TES model better matches the autocorrelation function for long lags, because the Markov chain captures the longer-term scene change behaviour. The price is in form of high computational complexity.

4.8 Conclusions

A large number of models have been suggested so far for video traffic, which are focused on capturing different types of video and encoding formats by using different modeling techniques. These models are focused on different attributes of the video and encoding standard as well as other important characteristics.

Model	Video coding	Level	Scene changes	Sources	Strong points	Limitations
Autoregressive models	DPCM; MPEG, H.261, H.264	I/B/P frames	Mostly No, Yes with some models	Single/multiple	Simple to understand and to implement	No single AR model can capture different statistical characteristics

Markov-process-based models	DPCM, MPEG, MPEG-4, H.263, H.264	I/B/P frames	Mostly Yes	Single	Accurate compared to AR models and can be used to model different types of video traffic	Difficult to accurately define and segment video sources into different states in the time domain due to the dynamic nature of video traffic
Self-similar models	DPCM, MPEG	I/B/P frames	Yes	Single	Accurately capture the LRD in video traffic	High computational complexity, fail to capture SRD in video traffic
Wavelet-based models	MPEG, JPEG, H.264	I/B/P frames	Yes	Single	Accurately model both SRD and LRD in video traffic	Difficult to implement and to determine how many levels of decomposition are needed
M/G/∞ - process-based model	JPEG, MPEG-2	frame	Yes	Single	More accurate than AR and self-similar models	Good for SRD traffic only
TES-based models	JPEG, MPEG, H.261, MPEG-4	GOB, I/B/P frames	Yes	Single	Non-parametric method; can generate any marginal distribution or arbitrary close approximation	Requires TEStool that is not publically available, high computational complexity

Table 4: Summary of video traffic models (Tanwir S. and Perros H., VBR Video Traffic Models, 2014)

5 NETWORKING AND PERFORMANCE ASPECTS

When discussing performance modeling, evaluation and monitoring of CONVINcE networking solutions, three concepts emerge: Quality of Service (QoS), Quality of Experience (QoE) and energy consumption. QoS is normally defining the performance at lower layers in the TCP/IP protocol stack, i.e., from the physical layer up to network layer, and even transport layer. QoS refers to individual networking domains as well. Examples of relevant QoS parameters are throughput (bit rate), frame or packet loss, delay, delay variations (PDV) and out-of-order packets, delay variations (PDV) and out-of-order packets. For instance, the throughput $B(T_i)$ is defined to be the number of data units (bit, bytes, packets, etc) $DU(T_i)$ observed during a given time interval $T_i = t_i - t_{i-1}$

$$B(T_i) = \frac{DU(T_i)}{T_i}$$

On the other hand, QoE reflects the end-user subjectively perceived experience of the quality of the particular delivered service on an e2e basis and as such it is a subjective measure. QoE is an e2e parameter. It includes the complete e2e system effects, which means that the networks between source (Head End) and destination (Terminal) can be treated as a black box. QoE is however dependent on the QoS parameters characterizing the parts making up the black box. The impact of each QoS parameter on the application and thus on the QoE differ depending upon the application type. For instance, in real time video streaming the irregularities of QoS parameters can effectively be considered as packet loss. The multi-dimensional relationships existing between QoE and QoS parameters and the definition of adequate Key Performance Indicators (KPIs) are very difficult questions to be answered (COMBO, 2014). Definitions of QoE include keywords like “quality of perception”, “user behaviour” and “psychological measure”, whereas QoS refers to parameters relevant for network like throughput, e2e delay, jitter and error rates. User centric measurements are normally referred to in conjunction with parameterization methods like Mean Opinion Score (MOS), Degradation Opinion Score (DOS) and Media Delivery Index (MDI) (COMBO, 2014).

The standardization activity for QoE and QoS has been very intense. Most active standardization organizations are ITU, ETSI, IETF and IEEE. Table 5 shows, e.g., the activity of ITU (Soldani, 2010).

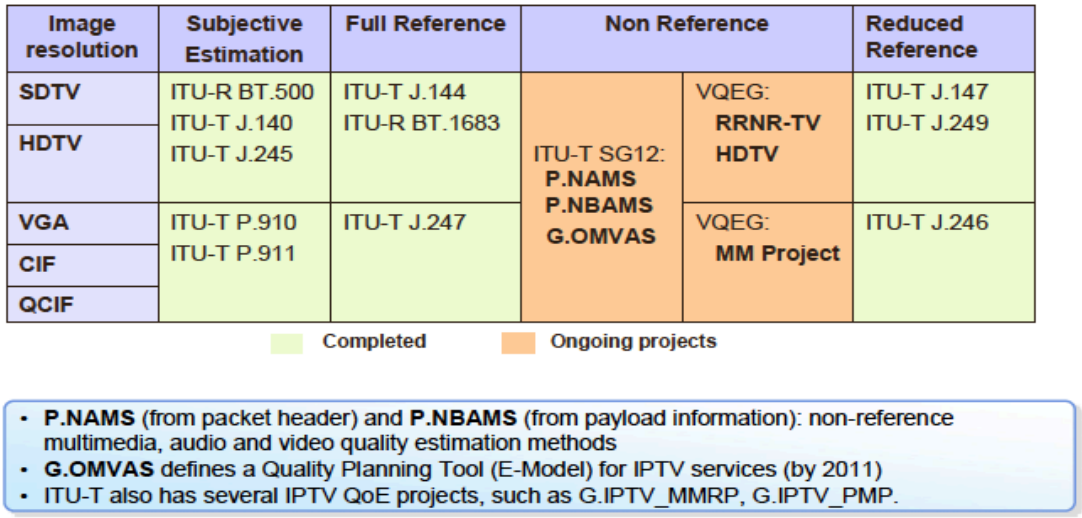

Image resolution	Subjective Estimation	Full Reference	Non Reference		Reduced Reference
SDTV HDTV	ITU-R BT.500 ITU-T J.140 ITU-T J.245	ITU-T J.144 ITU-R BT.1683	ITU-T SG12: **P.NAMS** **P.NBAMS** **G.OMVAS**	VQEG: **RRNR-TV** **HDTV**	ITU-T J.147 ITU-T J.249
VGA CIF QCIF	ITU-T P.910 ITU-T P.911	ITU-T J.247		VQEG: **MM Project**	ITU-T J.246

🟩 Completed 🟧 Ongoing projects

- **P.NAMS** (from packet header) and **P.NBAMS** (from payload information): non-reference multimedia, audio and video quality estimation methods
- **G.OMVAS** defines a Quality Planning Tool (E-Model) for IPTV services (by 2011)
- ITU-T also has several IPTV QoE projects, such as G.IPTV_MMRP, G.IPTV_PMP.

Table 5: ITU-T standardization activity on video QoE (Soldani, 2010)

Furthermore, it is also important to mention that today little work is available on end-user perception of offered services and their dependency on network quality, especially for Internet-based applications and networks. We simply have gaps between network measured QoS and user-perceived QoE (COMBO, 2014; Soldani, 2010). This indicates the strong need existing today for developing of mechanisms, procedures and tools to continuously monitor, operate and report QoE indicators function of QoS parameters. In our case, of particular importance is the development of mechanisms for QoE monitoring for network optimization, supervision and operation purposes for video distribution networks, under the condition of minimum e2e energy consumption.

In the following, the edge-cloud based network architecture is considered as network model. This is a general architecture considered for study in CONVINcE (Monnier R., 2016). As observed in Figure 7, the **functional** block scheme contains the following elements: Video Data Center, transport network, delivery network and video consumers. It is also important to mention that, according to the document D1.1.3, for a particular networking scenario, the most e2e performance limiting networking functional segment is the wireless RAN – terminal segment (Popescu A. a. a., 2016).

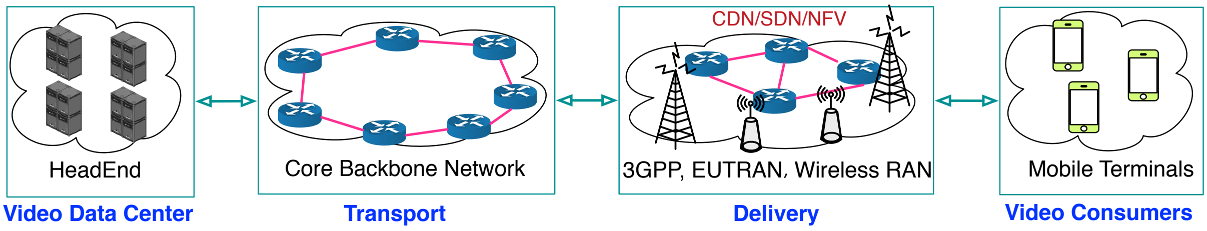

Figure 7: Typical networking scenario (Monnier R., 2016)

The situation however is more complicated in reality. A real network is of type inter-provider model, where two or more network providers share the end-to-end (e2e) path and, associated with this, the existent resources between the two edges (Figure 8, (Group, 2006)). By resources it is meant both physical resources and other categories of resources (e.g., bandwidth).

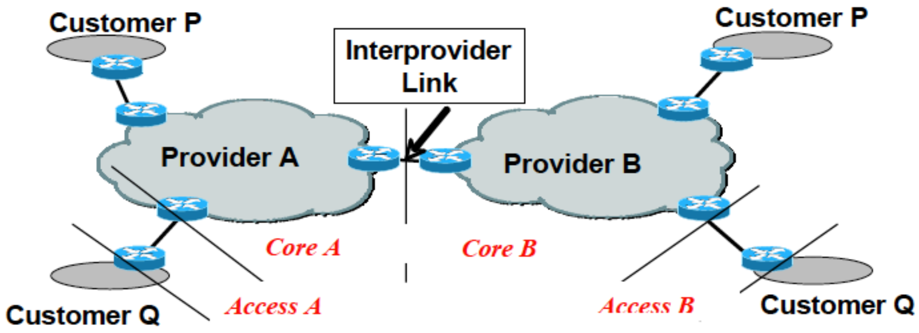

Figure 8: Real networking scenario (Group, 2006)

With reference to the example shown in the above figure, an important aspect is that the two service providers A and B are expected to offer services, which are concatenated with the services from the other provider, and provide so the end-to-end (e2e) service to end-user. For instance, in the case of fixed maximum e2e delay, the individual allowable delay may need to be negotiated among multiple carriers. In a similar way, the minimum e2e channel bandwidth can be negotiated and agreed among the participating teleoperators. Therefore, there is a need for monitoring the e2e performance experienced by a customer, which means that both service providers A and B are involved. This also means that a wide range of interconnection methods needs to be available to provide the e2e service. This does not refer to IP interconnection only, but also to other layers like Multi-Protocol-Label-Switching (MPLS) and layer 2 interconnection solutions (e.g., Ethernet) (Group, 2006).

In terms of interfaces, we have the so-called Customer Edge (CE) and the Provider Edge (PE). It is the responsibility of the customer to ensure that the traffic traversing the CE-PE link is correctly marked before it reaches the PE. This may enforce different aspects of the Service Level Agreement (SLA) and Customer to Provider Interface (CPI), where only one customer sends traffic (RFC, 2006).

An important element is regarding the Service Class (SC). These are defined by teleoperators and are primarily targeted for the transport of latency-sensitive applications like video and voice over IP. A general characteristic of these applications is their dependency on the e2e delay and, accordingly, the bandwidth available for service at different networking entities.

Another important question in this context is regarding the definition of Quality of Service (QoS) constraints for an e2e packet flow. These constraints are generally specified for the entire e2e path, which may traverse several networking domains owned by different Internet Service Providers (ISP). The problem is that each ISP allocates bandwidth and provides particular QoS guarantees independently from other providers (Anjum B. and Perros H., 2015). This means there is a need for co-operation to combine and to coordinate the QoS guarantees for the multi-domain routing specified by the entire e2e path shown, e.g., in Figure 8.

Using of average value as a performance indicator for performance metrics like the response time of a Web service or controlling the power budget in green Data Centers (DCs) may not provide the best e2e solution. This is because average values can be misleading, as they do not represent the range of values the metric under study may take. Towards this goal, a better performance indicator is the percentile of the particular metric (Anjum B. and Perros H., 2015). The percentile statistically bounds the behaviour of the system. As a definition, the q-th percentile (like, e.g., the 95th percentile) of a variable X is defined to be the value below which a given percentage of observations in a group of observations fall, i.e., X lies q% of the time. Furthermore, a difficult problem that must be solved is regarding how to add percentiles and how to partition them to individual components. Several interesting solutions are reported in (Anjum B. and Perros H., 2015) on the calculation of the weight function for the cases of Gaussian individual distributions, exponential individual components with identical rate parameters, exponential components with different rate parameters and two-stage Coxian distribution. Furthermore, other interesting solutions are reported in (Group, 2006), where the authors addressed the issue of how to allocate the e2e response time, packet loss and jitter across multiple networks owned by different operators. The authors of (Group, 2006) suggest using mean values for response time and percentile for inter-arrival time at destination for jitter in combination with a particular method suggested for adding the individual operators' jitter. It is reminded that the jitter is the one-way IP packet delay variation (IPDV).

Another important question related to the evaluation of performance on an e2e basis is regarding the particular traffic model for the video flow. As described above, there are several categories of models that can be used for this purpose, each of them with own advantages and drawbacks. Given that such models must be also related to the e2e performance in terms of QoS parameters and energy consumption as well as the end user performance in terms of QoE, we will consider in the following the concept of decomposition of general tandem queueing networks with MMPP input (Heindl A., 2000) as well as fluid flow models. The main advantage in this case is simplicity in combination with precise analysis.

The network in this case is partitioned into individual elements/nodes that are analysed in isolation. Further, when using decomposition algorithms for continuous-time analysis of networks, the output of a queueing system (particular network) is usually approximated by a renewal process, which serves as the arrival process to the next queueing system. By doing this, we easily obtain models based on Markov-modulated Poisson Processes (MMPP), which are simple and easy to use for doing sophisticated analysis studies (Heindl A., 2000). The drawback is that this model (tandem queueing networks) does not lend themselves to an exact analysis, especially because of concurrent non-exponential activities. Short details about MMPP models are presented above, in Section 4.2.

6 FUNCTIONAL QoE ARCHITECTURE

The definition of Quality of Experience (QoE) is the overall acceptability of an application or service, as subjectively perceived by the end-user (ITU-T R. , 2003).

There is not much work existing today on end-user perception of telecom services, especially based on Internet and their relationship to Quality of Service (QoS) measured in the network. That means there are gaps existing today between the network measured quality and end-user perceived performance. The fundamental problem in bridging the QoE and the QoE is because of the existing difficulties to measure and to interpret QoE indicators and to relate them to QoS parameters. The consequence of this is in form of difficulties to develop mechanisms, procedures and tools to continuously monitor, operate and report QoE indicators with the goal of network supervision, optimization and operation (Soldani, 2010), (Fiedler M. H. T.-G., 2010).

Figure 9 shows an example of relationship between QoE value and QoS disturbance, which may look differently, function of the particular application like, e.g., VoIP, Web browsing, IPTV services (Soldani, 2010), (Fiedler M. H. T.-G., 2010). The service satisfaction criterion refers to different parameters, which are depending upon the particular application like, e.g., "**Web paging loading time AND active session throughput**" **(Web service)**, "**Streaming mean active throughput AND streaming media active transport jitter AND streaming media loss**" **(video streaming)**, and "**Active session throughput**" **(video over HTTP)** (Soldani, 2010).

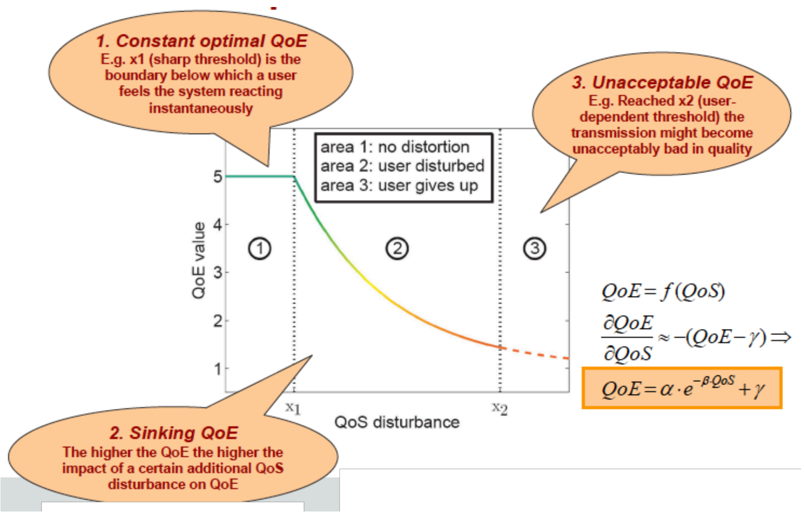

Figure 9: Relationship between QoE and QoS (Soldani, 2010; Fiedler M. H. T.-G., 2010).

To make things even more complicated, there are today two main quality assessment methodologies, which are subjective and objective (Soldani, 2010). A serious problem is that measuring and providing good QoE for video applications is very subjective in nature. The most commonly used subjective method for quality measurement is the ITU-T standardized Mean Opinion Score (MOS) (Soldani, 2010) (ITU-T R. , 2003). This is defined as a numeric value from 1 to 5, which corresponds to qualities varying from poor to excellent, respectively. MOS is basically a method used to provide the mapping function required to map the objective Video Quality (VQ) into the predicted subjective score (MOS_p) $MOS_p = x + yVQ$. The main drawback in this case is regarding complexity, high cost, time consuming.

Figure 10 below shows an example of mapping model for QoS and QoE parameters (Soldani, 2010).

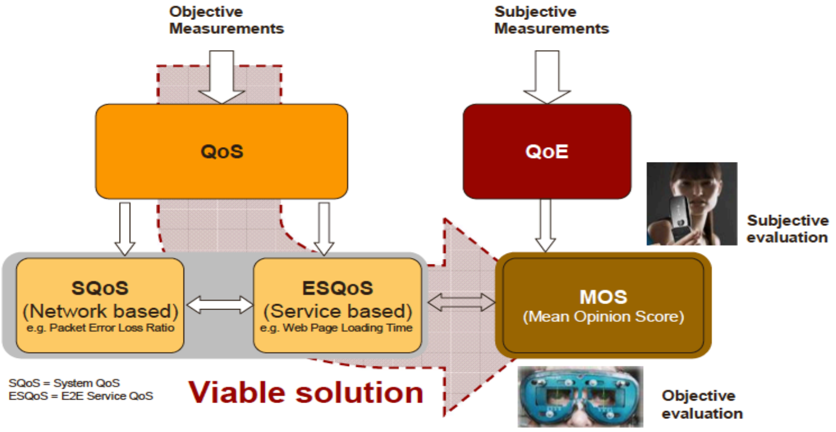

Figure 10: Mapping model QoS - QoE (Soldani, 2010)

Figure 11 shows an example of functional architecture to monitor and supervise QoE in a network (Soldani, 2010). Such architecture has five layers dedicated to specific tasks:

- Data acquisition layer, to collect information relevant to QoE like, e.g., packet loss error rate, throughput, delay jitter
- Data collection layer, to provide storage for collected information for continuous monitoring
- Data analysis layer, to provide data query, filtering, refinement, aggregation, correlation and processing as well as data conversion required for different data sources
- Presentation layer, to provide tools for visualization and report generation; the customer experience is the result of combining different network indicators
- Application layer, where the functional objectives use cases are developed

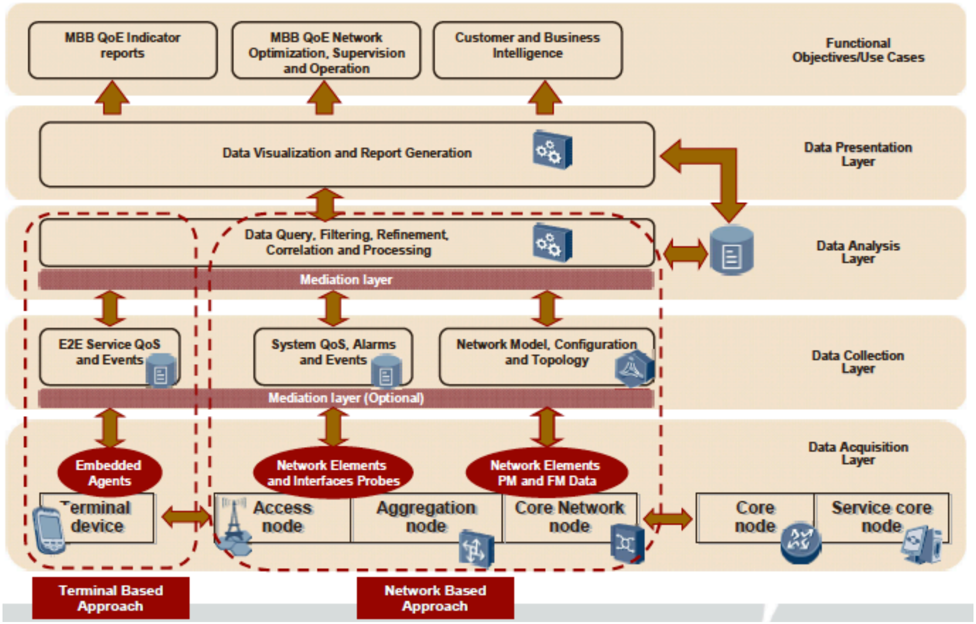
MBB QoE Indicator reports
MBB QoE Network Optimization, Supervision and Operation
Customer and Business Intelligence
Functional Objectives/Use Cases
Data Visualization and Report Generation
Data Presentation Layer
Data Query, Filtering, Refinement, Correlation and Processing
Data Analysis Layer
Mediation layer
E2E Service QoS and Events
System QoS, Alarms and Events
Network Model, Configuration and Topology
Data Collection Layer
Mediation layer (Optional)
Data Acquisition Layer
Embedded Agents
Network Elements and Interfaces Probes
Network Elements PM and FM Data
Terminal device
Access node
Aggregation node
Core Network node
Core node
Service core node
Terminal Based Approach
Network Based Approach

Figure 11: Functional QoE architecture (Soldani, 2010)

Figure 11 above gives a hint about the complexity of provision of end-user QoE performance. This involves information collection, data processing and different activities of control in the e2e chain to provide the expected performance. Furthermore, things become even more complex in the case of considering the energy consumption in the communication chain.

Based on this architecture, the so-called service satisfaction criteria in terms of QoE can be obtained, which depends upon the particular service and the particular parameters collected, as indicated in the table 6 below.

With regard to this table, it is observed that the main QoS parameters that influence the experienced service satisfaction (i.e., the QoE) at video services are the throughput and the jitter. Given that the e2e throughput performance at CONVINcE is limited by the networking element with the minimum throughput performance, this means important efforts must be laid on the provision of good e2e performance in terms of maximum e2e throughput and minimum e2e energy consumption and not only individually at the different networking elements.

Web service	Web paging loading time AND Active session throughput
Video streaming	Streaming mean active throughput AND Streaming media active transport jitter AND Streaming media loss
Video over HTTP	Active session throughput
VoIP	Call setup success ratio AND Call setup time AND Call cut-off ratio AND Speech quality

Table 6: Examples of service satisfaction criteria (Soldani, 2010)

The suggested concept for e2e performance optimization at CONVINcE is presented below.

7 PERFORMANCE OPTIMIZATION

Based on the information provided above, the next step is to further develop the concept of single-objective multi-constraint performance optimization advanced in the WP1 document D1.1.3 (Popescu A. e. a., CONVINcE D1.1.3 High-Level Architecture Design, 2016) and to extend it into a multi-objective multi-constraint performance optimization algorithm for CONVINcE.

The edge-cloud network architecture is still considered for this purpose, which is modelled as being composed by a number of networking elements, as shown in Figure 12. At the moment, we do not

specify the details of networking elements but only define the fundamental networking functions, as follows: **Head End – Edge Cloud – Wide Area Network – Radio Access Network – Terminal**.
As observed in the figure 12, **every networking element is characterized by a set of three performance indicators**: Throughput T, Quality of Service QoS (expressed in terms of other parameters than throughput, which are specific to the particular networking element, e.g., error rate for RAN, execution time/delay for Edge Cloud) and Energy consumption E. Besides this, we also have the Terminal where the Quality of Experience QoE is evaluated.

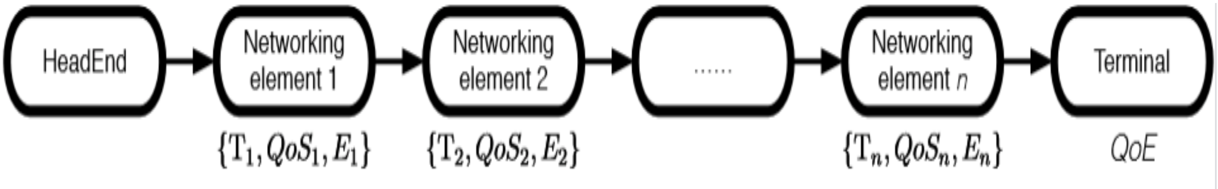

Figure 12: e2e networking model

The goal of CONVINcE performance optimization algorithm is to provide the best performance with regard to best possible QoE at end user and minimum e2e energy consumption. There are two optimization algorithms involved in this process.

- ***Optimization algorithm #1: best end-user QoE performance* AND *minimum e2e acceptable energy consumption***

Accordingly, this involves the following

- ***Optimization algorithm #2: best QoS performance for individual networking entities wrt the above-mentioned optimization algorithm; examples of QoS parameters are throughput, error rate, jitter***

In other words, we have a three-dimensional process of performance optimization. The QoS parameters for individual networking entities can be different from networking element to networking element, and they depend upon the specific of the individual networking entity. For instance, 3GPP suggests the following metrics (so-called Key Performance Indicators KPIs) for variable traffic in mobile networks (ETSI-2011, 2011):

- Mean, 5th, 50th, 95th percentile user throughput
- Served (cell) throughput
- Harmonic mean normalized cell throughput
- Normalized cell throughput
- Resource utilization

In particular, the metrics suggested by 3GPP to trade-off the quality versus energy consumption are relevant for us (ETSI-2011, 2011):

- Mean user throughput
- Throughput Complementary Distribution Function (CDF)
- Median and 5th percentile (worst) user throughput

The throughput is a measure of the rate at which data is successfully transmitted through the network or through the network entity. This corresponds to the amount of useful data transfers, and it is generally less than the amount of data injected into the network. This is due to the lost, corrupted, retransmitted or misdelivered packets. Sharing the network also lowers the throughput for every competing user. Finally, diverse protocol overheads contribute to reducing the throughput as well. Maximum obtainable throughput is therefore equivalent to the system's capacity.

Typical units of measure for throughput are bit/s, byte/s and packets/s. Recommended values for the time interval are one to five minutes interval.

Besides throughput, other relevant QoS parameters are error rate, jitter and e2e delay. The point is that these parameters can be translated into the throughput performance for the particular networking entity, as it is done for instance at Transport Control Protocol (TCP) (Popescu Adrian, 2015; Padhye J., 1998).

Furthermore, the fundamental law governing the behaviour of networking nodes must be respected in every node as well. This is about the so-called ***flow-conservation law*** (Ravindra K. Ahuja, 1993), which states that:

[Flow-out-of-a-node] – [Flow-into-a-node] = [Network-supply-at-the-node]

where by “network supply” we mean auxiliary flow(s) that enter the particular node, with the purpose of maintenance and control.

Given these elements, the CONVINcE performance optimization has therefore two parts:

- Optimization of low-layers networking elements, which refers to layers 1-3 (PHY/DLL/NL) in the TCP/IP protocol stack
- Optimization of the protocols in the layers above (MPEG-DASH/HTTP/TCP)

It is mentioned that this concept is valid for all categories of architectures and scenarios considered in CONVINcE.

7.1 Fundamental procedure

As mentioned above, the optimization algorithm has two main parts: the first one refers to low layers NL/DLL/PHY in the TCP/IP protocol stack along with the e2e chain as indicated in Figure 13. The goal in this case is to obtain the best trade-off between QoE at end-user and minimum e2e energy consumption on the whole e2e networking chain. The second part refers to the high layers MPEG-DASH/HTTP/TCP above. The goal in this case is to obtain best end-user QoE, which is controllable via the above-mentioned e2e protocols, on top of low-layers protocols.

As observed in figure 13, the e2e throughput is the fundamental parameter used to bridge the relationships among QoE, QoS and energy consumption.

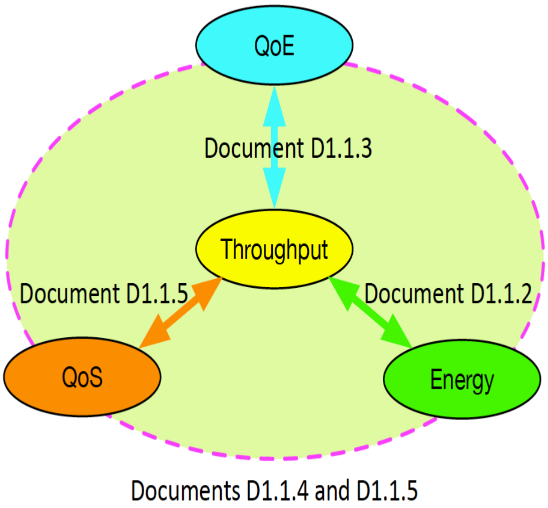

Figure 13: QoE- and energy-aware performance optimization at CONVINcE

The general procedure for performance optimization is as follows:

- Define the block scheme and the component elements like, e.g., access (eNodeB, HGW, …), backhaul and transport (switching,...), core network (element1, ..) and servers (Data Center, ..)
- Measure and model the power consumption of each element in steady state and also when serving video flows
- Measure and model the throughput provided by each networking element
- Do performance optimization for the individual networking elements in terms of best trade-off between the minimum energy consumption and highest possible throughput.
- On an e2e perspective, the throughput of individual networking elements is also related to the throughput of the other networking elements in the sense that one should avoid, as much as possible, low values for any of them. In other words, the goal is to provide as high as possible e2e throughput, which means high throughput in all networking elements, under the condition of minimum energy consumption in all networking elements. This concept is also depending upon the particular testbed. It is well known that the minimum individual throughput in the e2e chain of networking elements, also known as the bottleneck bandwidth (Constantine B., 2011) determines the throughput experienced by the end user, which may deteriorate the QoE performance in the case of low values.
- Do performance optimization at low layers NL/DLL/PHY for the whole e2e networking chain with regard to highest possible e2e throughput, best (terminal) QoE performance and minimum possible e2e energy consumption
- Do performance optimization at high layers MPEG-DASH/HTTP/TCP for the e2e networking chain with regard to maximum e2e throughput and minimum error rates
- Do system performance optimization with regard to the best trade-off: minimum e2e energy consumption AND highest e2e system throughput AND best QoE performance; this refers to the whole protocol stack MPEG-DASH/HTTP/TCP/NL/DLL/PHY
- Do system performance optimization for the testbeds considered in CONVINcE

The modeling and evaluation activities take into consideration the architectural blocks and the temporal behaviour of component elements as well. We also assume that the total traffic coming into the network is equal to the total traffic leaving out of network. A unit of work is defined to be the work associated with the transport of traffic accepted from one external interface to another external interface.

The procedure of performance optimization at CONVINcE is partitioned into two categories: low-layers network optimization and high-layers network optimization. The details are as follows.

7.2 Low-layers network optimization: multi-objective multi-constraint optimization

Low-layers network optimization is about performance optimization done at the three low layers: network layer (NL), data-link layer (DLL) and physical layer (PHY).

Basically, the concept of low-layers network optimization at CONVINcE is about the extension of the single-objective multi-constraint optimization algorithm advanced in the document "CONVINcE D1.1.3 High-Level Architecture Design" (Popescu A. a. a., 2016), which is applicable to the last networking segment Radio Access Network (RAN) – Terminal only. This optimization concept is now extended as a multi-objective multi-constraint optimization algorithm, which is applicable to the whole e2e networking chain Head End (HE) – Edge Cloud (EC) – Core Backbone Network (CBN) – Wireless RAN – Terminal. As shown in Figure 7, the complete networking chain provides the following functional chain: Video Data Center – Transport – Delivery – Consumers.

The functional networking chain covers the four categories of architectures considered for CONVINcE (as described in the document D1.1.1). Every such architecture has associated an objective function in form of minimum e2e energy consumption under the condition of best possible QoE at the end-user. In the following, focus will be given to the edge-cloud architecture as a particular study case. The procedure is however similar for the other three architectures as well.

Given	CONVINcE edge-cloud architecture with the network elements: Video Data Center (Head End), Edge Cloud (CDN), Core Backbone Network, Radio Access Network (RAN) and Terminal

Live Video Streaming Scenario

Do Power minimization for individual entities in terms of relevant parameters: number of nodes weighted by the number of active CPUs (Edge Cloud); best route with regard to optimum packet forwarding (Internet); maximum ratio RF output power / total AC input power (RAN); minimum number of replicating proxies (CDN); minimum power amplifier consumption (HFC); minimum power consumption and maximum QoE (terminal)

Also do Measure the throughput/channel performance for every individual entity in the e2e network chain

Compare the individual throughput/channel values obtained above

Optimize the process of power minimization for individual networking entities with regard to a given maximum difference accepted between the maximum and the minimum throughput of e2e component networking entities ($\Delta T = T_{max} - T_{min}$), which must not exceed a maximum given value ($\Delta T_{max} \geq T_{max} - T_{min}$). The value of ΔT_{max} is selected with regard to the acceptable QoE value at end user

Also do Select the individual entity that shows the minimal e2e throughput/channel and denominate this throughput as being the sustainable throughput for the whole system (Fiedler M, 2016)

Single-objective multi-constraint performance optimization in terms of minimum e2e power consumption and best QoE at terminal, under the condition of given acceptable throughput

It is observed that the multi-objective multi-constraint optimization algorithm described above becomes at the end a single-objective multi-constraint optimization algorithm, as presented in the document D1.1.3 (Popescu A. e. a., CONVINcE D1.1.3 High-Level Architecture Design, 2016). Furthermore, the same document D1.1.3 presents an example of such algorithm applied in the case of Live Video Streaming (LVS), where the e2e system bottleneck is given by the last networking segment RAN – terminal. The same concept is applicable in this case, after the procedure of multi-objective multi-constraint optimization algorithm is done.
Figure 14 shows the concept of single-objective multi-constraint system optimization for CONVINcE (Popescu A. e. a., CONVINcE D1.1.3 High-Level Architecture Design, 2016).

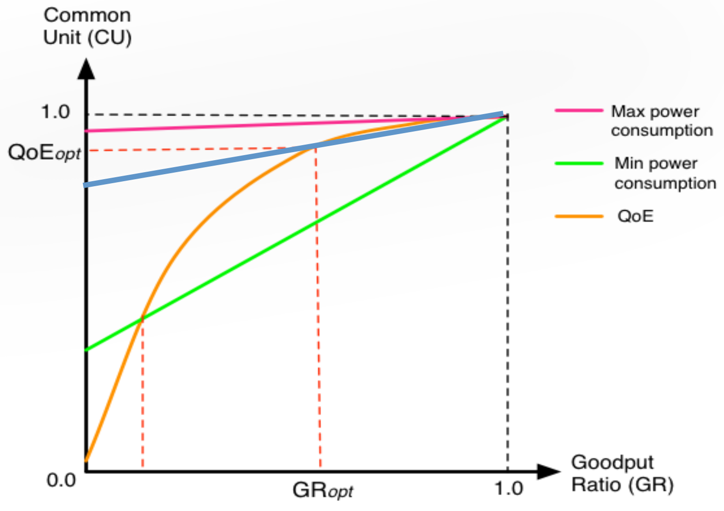

Figure 14: System optimization at CONVINcE

The y-axis represents the so-called parameter Common Unit (CU). This is an abstract parameter, which is used to represent both parameters QoE and Power, but with reference to their respective maximum values. In other words, this means that both parameters have a maximum value of 1.0 on

CONVINcE confidential

the CU axis. Furthermore, the x-axis represents the so-called goodput ratio (GR), which is used to represent the channel throughput of the particular system, with reference to the maximum value of throughput. This further means that this parameter has a maximum value of 1.0 on the GR axis.
With regard to figure 14, the performance optimization problem is as follows:

Given CONVINcE edge-cloud architecture with the network elements: Video Data Center (Head End), Edge Cloud (CDN), Core Backbone Network, Radio Access Network and Terminal

Live Video Streaming Scenario

Do Power minimization for individual entities

Also do Measure the e2e throughput/channel for every individual entity in the e2e network chain; also measure the QoE at the terminal

Select the individual entity that shows the minimal e2e throughput/channel and denominate this throughput as being the sustainable throughput for the whole system

Plot the function CU = f (GR)

Select the optimal value for this function; select accordingly the values for QoE_{opt} and GR_{opt}

Subject to Specific individual requirements for power minimization of entities in terms of, e.g., minimum power consumption for Head End $\cap$ minimum power consumption in Edge Cloud $\cap$ best route (Internet) $\cap$ minimum power consumption in RAN

The problem of selecting the optimal value for the function CU = f(GR) is as follows. The goal for QoE is towards GR = 1, whereas the goal for power consumption is towards GR = min. Figure 14 shows for instance the particular case of using a simple arithmetic average for the value of GR_{opt} ($GR_{opt} = 0.5$) in the case of optimal power consumption (as indicated by blue line). Based on this, a particular value is obtained for QoE_{opt}, which is acceptable, as observed in the figure. QoE is almost maximum in this case, which corresponds to CU = 1. Furthermore, it is also observed that in the case of maximum power consumption (red line) the optimal solution is in form of $GR_{opt} = 1$.
Similar constrained optimization problems can be defined for the other scenarios (On-Demand Video Streaming; Camera-Based Sensor Networks; Cloud Gaming) as well.

7.3 High-layers adaptive video streaming

High-layers adaptive video streaming is about video streaming and performance optimization done at the application protocol (AP), application layer (AL) and transport layer (TL). In the following, the protocol stack MPEG-DASH/HTTP/TCP is considered.
It is reminded that Ultra-High-Definition Television, also known as Ultra HD television, or UHDTV include 4k UHDTV (3840 by 2160p) and 8k UHD (7680 by 4320p) standards, which are approved by the International Telecommunication Union (ITU) in Recommendation BT.2100. This standard is built on the 2015 approved UHDTV Recommendation BT.2020.
The most common pitfalls of 4k and 8k video quality impairments are as follows (Elemental, 2016):

- Rules and thresholds for HD no longer apply: any impairment is much more prevalent in 4k and 8k video because of much higher bit rate, larger number of pixels and larger vieweing surface compared to HD video. The consequence is mainly in form of higher frequency with which errors may occur in a 4k or 8k vieweing experience compared to the one of HD. In other words, quality levels must be higher for 4k or 8k video to provide so vieweing experience similar to the one ar HD.
- Over-compression: this is mainly related to the limited existing processing capacity and bandwidth, which is necessary to create a satisfactory vieweing experience. More resources are needed to improve the video experience.

- Excessively variable packet and bitrate output: these limitations adversely impact the communication in the network with consequences on video quality. One needs to monitor the maximum peak variable bitrate (VBR) of the traffic coming from the encoder, and notify significant and fluctuating bursts. Use of statistical multiplexing is recommended as well.
- Varying quality of adaptive bitrate output streams: improper encoder settings may influence the video quality of output streams. Video quality needs to be monitored to detect these categories of problems and to compensate them.
- Group Of Pictures (GOP) misalignment across Adaptive Bit Rate (ABR) output streams. This is a problem that may appear when a client device switches from one ABR profile to another one, even within the same resolution. This problem demands that operators ensure their encoders with GOP-locking outputs for seamless playback.

In other words, this means that, because of higher resolution at 4k and 8k video, with physically larger vieweing space (big-screen television) and higher frame rates, the consumers are more likely to perceive visual defects in video playback. The operators therefore need to provide adequate settings for encoding, packaging and delivery systems, which must be complemented by adequate video quality monitoring systems to detect impairments in real time as well as to compensate them. Regarding the components of video quality, the most important element is the video viewer's perspective. With regard to this, the components of a quality online video experience can be partitioned into two broad groups, which are about the reliability of the playback system (playback stalls and video start time) and the quality of the video image (resolution, delivered bitrate – key metric, smooth playback).
In the following, a particular focus is given to video delivery systems.

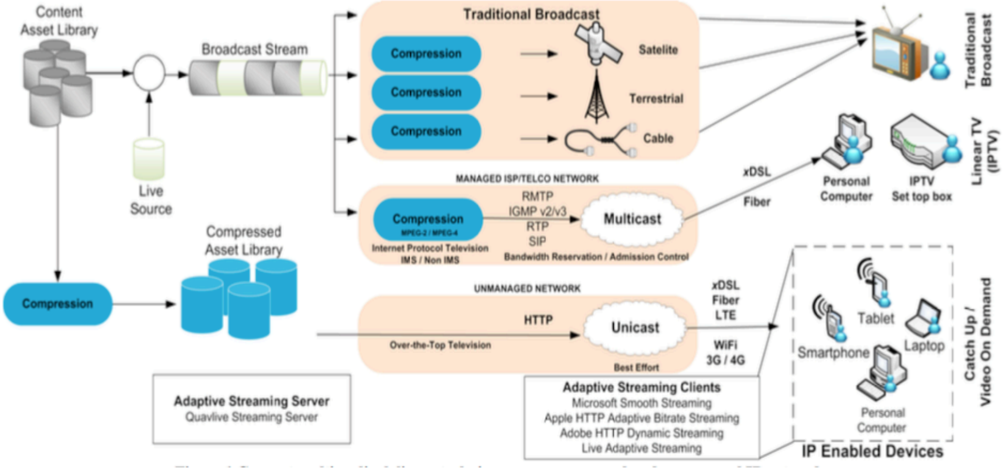

Figure 15: Video delivery over managed and unmanaged IP networks (Blair A., 2011)

Figure 15 above shows the main component elements used in video distribution across managed and unmanaged IP networks (Blair A., 2011). The focus in our case is not laid on the traditional broadcast technique. Furthermore, whereas IPTV demands for resource reservation as well as the need to create technical and commercial relationships between the content providers and underlying service providers, which can be difficult and costly to implement, the situation is different in the case of OTT. The OTT systems have no way to use layer 3 (IP) management facilities to provide QoS provisioning and to do resource reservation like, e.g., bandwidth reservation, packet prioritisation and overload protection. The consequence is that these limitations must be solved at higher layers in the TCP/IP protocol stack, which refers to transport protocol, application protocol and the application itself. The OTT architecture takes the existing content from the asset library, it compresses it and transmits it individually to IP devices. The HTTP protocol is used from which the distinct fragment files are served or alternatively the data may be multiplexed. The fragments are hosted on an HTTP server from which they are served to clients. Furthermore, the clients play the stream by requesting fragments from the HTTP server. Then the client plays the fragments continuously as they are downloaded. In the case the particular data is stored in separate fragment sequences, then all of them are downloaded to form the playback.

Adaptive streaming is used as well. This is basically a technique where a piece of content is encoded (as H.264, though HVEC or other video codecs are possible) into multiple copies of varying quality levels, each of them with different bit rates (Blair A., 2011). These copies are stored as ”chunks” of different durations, normally ranging between one and five seconds. They contain data encoded at varying levels of quality. Furthermore, based on the particular network conditions, mainly expressed in terms of e2e throughput, the quality is optimized by requesting a chunk suitable to the existent network conditions. The selected chunk is downloaded by using progressive file download.

Decision-making algorithms are used at client during run time and all together form the system of video streaming. The client monitors the current network conditions and reacts to any change that occurs and affects the transmission time and accordingly the content quality. The main parameter that is relevant for the video streaming system is the throughput. Other parameters like, e.g., packet loss, can be easily converted into throughput performance. Also, the transmission time varies depending upon the quality of the video segment and the network conditions.

The protocols relevant for the video streaming system are therefore the Transport Control Protocol (TCP) and the Dynamic Adaptive Streaming over HTTP (MPEG-DASH). Figure 16 shows the block scheme of a multimedia streaming system (Chowrikoppalu Y. P. G., 2013). The main components are the encoder, streaming client, media transfer protocol and the underlying physical network.

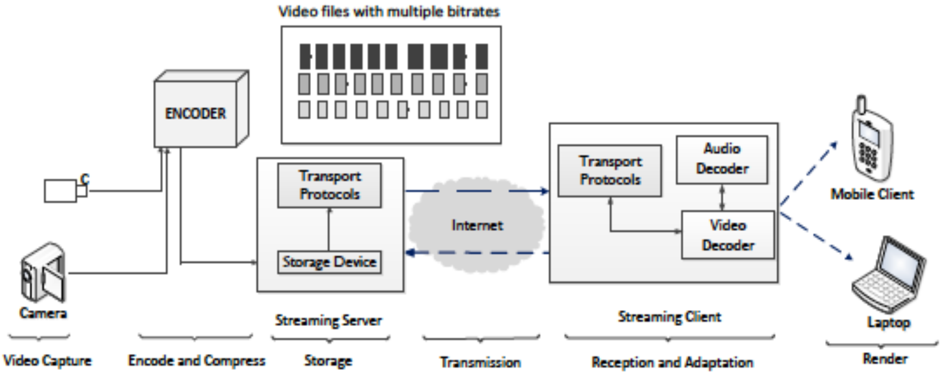

Figure 16: Multimedia streaming system (Chowrikoppalu Y. P. G., 2013)

A short presentation of the protocols, together with their benefits and limitations, is presented in the following. Related to this, solutions suggested to minimize the limitations are presented as well.

7.3.1 Dynamic adaptive streaming over HTTP (DASH)

MPEG DASH is a Web-compatible international standard used to describe multi-rate encoded multimedia, which was ratified in December 2011 and published by the International Organization for Standards (ISO) in April 2012 (ISO/IEC, 2012). The standard describes the multi-rate encoded multimedia, allowing so intelligent client playback applications to dynamically choose which portions of a media presentation to select for rendering to the user, which is based on the network dynamics or available device resources. DASH also provides a framework for hosting a common encryption mechanism, usable by multiple digital rights management (DRM) systems. (ISO/IEC, 2012).

MPEG DASH is actually an efficient and flexible Internet-compatible distribution platform that scales to the rising demands for video communication and distribution. Basically, MPEG DASH is a standard for optimizing the multimedia delivery over the open Internet. It is not a specification for system, or protocol, or presentation, or codec, or middleware, or client. Rather, MPEG DASH is more like a neutral enabler aimed at providing several formats that foster efficient and high-quality delivery of streaming media services over the Internet. Furthermore, MPEG DASH can be viewed as an amalgamation of the three prominent industrial standards for adaptive streaming protocols, namely Adobe HDS, Apple HLS and Microsoft Smooth Streaming (Iraj, 2011; Andy, 2012; Fisher Y., 2014).

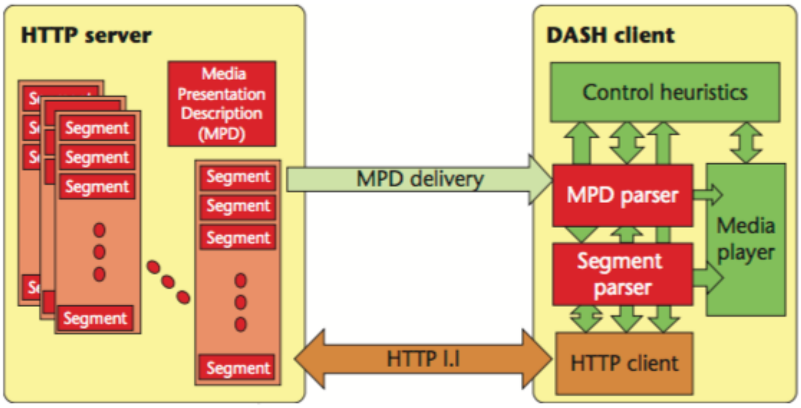

Figure 17: MPEG DASH model (Sodagar I., 2011)

Figure 17 shows for instance the scheme of a DASH client (Sodagar I., 2011). It is observed, e.g., that the content stored at server has two parts, namely Media Presentation Description (MPD) and Segments. More technical details are presented in (Andy, 2012; Fisher Y., 2014).
Though MPEG DASH appears to be strong and popular, in reality the scarcity of widely adopted and feature-rich clients makes it look better in theory than in practice. In spite of being the most popular protocol used by video service providers, difficulties of different nature complicates the situation like, e.g., no dominating format, dependence on the client devices served, DRM requirements inherited from content owners (Fisher Y., 2014). Furthermore, there are several directions expected to be considered in the future to improve the performance as well as the protocol acceptance. These are as follows:

- Multi-path delivery, which is about using of more paths for data transfer. Examples of protocols used for this purpose is the Multipath-TCP (MP-TCP), which is about using of several TCP flows for data transfer. The goal is to improve the e2e throughput performance and therefore the end-user QoE, in spite of technical difficulties associated with this like, e.g., need for flow synchronization and for adaptation logic. Furthermore, this can be also considered as a possible solution to the problem of protocol aggressivity against competing flows at the level of application as well as against the transport protocol itself.
- DASH over protocols others than http like, e.g., HTTP2.0, Content Centric Networking (CCN). This is however a brand new direction not yet investigated and evaluated. Large research efforts are needed to understand and to evaluate the advantages and drawbacks of this method.

In the following, the focus is laid on the streaming protocols.

7.3.2 Streaming protocols

Many protocols have evolved to support multimedia transmission over the Internet. Connected with this, it is important to distinguish between the video delivery over an open and a closed network. Video delivery over an open network refers to the open, uncontrollable Internet, with no involvement from Internet Service Providers like, e.g., in the case of OTT service. On the other hand, video delivery over a closed network refers to networks managed by Internet Service Providers (ISPs), like in the case of IPTV service.
Two of the most widely protocol sets used to provide streaming services are the Real Time Protocol (RTP)/User Datagram Protocol (UDP) and the HyperText Transfer Protocol (HTTP)/Transport Control Protocol (TCP). Each of these protocol sets provides different streaming services and performance. For instance, RTP/UDP based streaming was specifically developed for media transfer and it is characterized by own packet structure and session management facilities. Both multicast and broadcast transmissions are supported.
Although RTP/UDP works well in managed networks, it has several drawbacks. The most important one is that RTP packets are blocked by firewalls and they cannot pass through the Internet.
On the other hand, the HTTP/TCP streaming has seen an explosive development over the last years, in spite of some drawbacks like, e.g., it supports only unicast transmission and media is delivered in

large segments. HTTP/TCP based streaming offers important advantages compared to RTP/UDP and, in combination with the easiness of deployment in the current Internet, has had the consequence that this is today the preferred solution for streaming. Other important advantages are that HTTP/TCP are foundation protocols for World Wide Web (WWW) and as such they are part of the protocol stack in every Internet-connected device. They also provide a simplified connectivity as HTTP/TCP flows can traverse network firewalls and Network Address Translators (NATs). Finally, another important advantage is that HTTP/TCP based streaming can be managed without the need to maintain a session state on the server, which means that the system scalability is improved.

At the same time, the HTTP/TCP based streaming has several important limitations, which are regarding performance limitations and the lack of support for multi-path streaming. Performance limitations are mainly due to fluctuations in the e2e throughput, and previous studies suggest that TCP requires twice the bitrate of the video for uninterrupted video streaming and good QoE (Wang B., 2004). Other limitations are because adaptation is done based on the measured bandwidth only and it does not consider the fullness of the playout buffer, switching between different video versions does not consider the impact of varying the video quality of user experience as well as erroneous and inefficient bandwidth estimations (Lederer S., 2012; Chowrikoppalu Y. P. G., 2013). The consequence therefore is that the TCP throughput performance and the associated algorithms for performance increase and optimization are of paramount importance for video streaming and good QoE performance.

It is also important to indicate that, in the case of limitations in the e2e TCP throughput, i.e., the TCP throughput is reduced to a lower level than the current playback rate, then the buffer may eventually become exhausted, something which may create stops in video playback, with the consequence of disruptions that may impact the perceived QoE. An interesting solution to this problem is suggested in (Mok R., 2011), where three application performance metrics are suggested to complement the e2e throughput metric. These metrics are as follows:

- Initial buffering time
- Mean duration of a rebuffering event
- Rebuffering frequency

Network QoS parameters have a direct impact on these metrics. These metrics are also applicable to adaptive video streaming. Subjective experiments reported in (Mok R., 2011) indicate that it is the rebuffering frequency that is the main factor responsible for variations in QoE. Another important factor influencing the QoE is the temporal structure of the video traffic.

Furthermore, another important quality metric that should be considered in the case of DASH systems is regarding the automatic switching between quality levels and the associated metric quality transitions (Mok R. L. X., 2012). For instance, results reported in (Mok R. L. X., 2012) indicate that the end user could prefer a stable video stream with fewer quality transitions at the expense of an overall higher bitrate.

7.3.3 TCP throughput

The Transport Control Protocol (TCP) is today the most widely used transport protocol. It provides logical e2e connection between applications running on end hosts. Basically, TCP is a connection-oriented protocol that provides reliable in order byte delivery of data over unreliable underlying network. The TCP protocol was first defined in IETF RFC 793 (DARPA, IETF RFC 793, 1981). The goal was to reliably operate over almost any transmission medium regardless of transmission rate, delay, data corruption, data duplication or data reordering of segments. Later on, advances in networking technology have resulted in higher transmission speeds, and a number of TCP extensions have been developed and defined in other IEEE RFC documents.

TCP focuses on reliable and accurate delivery rather than timely delivery, which means that if a data packet is lost or corrupted on the way to destination, it will be retransmitted, with the consequence of delay. Also, buffering is used to handle delays in the system. All together, TCP exhibits complex behaviour, especially because the traffic conditions in Internet are quite complicated.

Basically, TCP contains a set of algorithms used to provide the expected services to application layer, on an e2e basis. These are as follows (DARPA, IETF RFC 793, 1981):

- Process-to-process communication
- Encapsulation and decapsulation
- Multiplexing and demultiplexing

- Unicast transport
- Fully reliable delivery
- Flow control, which refers to eliminating the congestion in the end buffers
- Error control
- Congestion control, which refers to eliminating the congestion in the network as well as providing the e2e throughput performance

The most important algorithm, which influences the transport performance, is the congestion control. TCP was initially designed to alleviate the congestion collapse in Internet and to provide so the stability in Internet. These mechanisms are embedded in the TCP congestion control mechanism in the form of slow-start and congestion avoidance phases. Also, the so-called congestion window limits the amount of data transmitted by the TCP sender. Related to this, there are several versions of TCP, e.g., Tahoe, Reno, Vegas, Compound, CUBIC, which are dedicated to different purposes (DARPA, IETF RFC 7323, 2014). Relevant information is provided in:

- RFC 1323 (TCP Extensions for High Performance)
- RFC 2018 (TCP Selective Acknowledgment Options)
- RFC 2582 (The NewReno Modification to TCP's Fast Recovery Algorithm)
- RFC 2883 (An Extension to the Selective Acknowledgment SACK Option for TCP) and
- RFC 3517 (A Conservative Selective Acknowledgment-based Loss Recovery Algorithm for TCP)

Given that HTTP DASH systems use TCP for transport, it is necessary to be aware of that both TCP and DASH may influence the performance of the application due to limited throughput performance as well as that the throughput performance is adaptive and therefore it is necessary to increase this performance as much as possible in order to obtain good QoS and QoE performance.
Maximum achievable throughput for a single TCP connection is determined by several parameters, e.g., maximum bandwidth of the lowest link in the e2e path, bit errors/packet loss and the available space in the end buffers. In other words, variations in these parameters are always observable in the form of new values for e2e throughput.
Assume for instance a TCP congestion control mechanism with maximum congestion window size $W_{\max}$ (expressed in number of TCP segments), round-trip time RTT and (maximum) segment size MSS. For a given probability of packet loss p, the TCP throughput $B(p)$ is given by (Padhye J., 1998):

$$B(p) \approx \min\left(\frac{W_{\max}}{RTT} * \frac{1}{RTT\sqrt{\frac{2bp}{3}} + T_0 \min\left(1.3 * \sqrt{\frac{3bp}{8}} p(1+32p^2 \right)} \right)$$

There are several methods for the estimation of RTT, each of them with different advantages and drawbacks, as indicated in RFC 1122 and RFC 2988. Basically, all TCP implementations attempt to estimate the current RTT by observing the delay patterns for recent segments only, and set the timer to a value somewhat larger than the estimated RTT. As the RTT changes, TCP should accordingly modify its timeouts (RTO). Several popular methods for RTT estimation are simple average, exponential average, RTT variance estimation (van Jacobson's algorithm), Karn's algorithm and exponential Retransmission TimeOut (RTO) backoff (Popescu Adrian, 2015).
It is also important to mention that things become more sophisticated in the case of cellular communication and wireless Local Area Networks (LANs), which is because environmental conditions and terrestrial obstructions and reflections may lead to high unpredictable error rates and models (Xylomenos G., 2001).
Although the above formula provides an accurate analysis, it is certainly beyond the level of detail relevant for our purposes. A more popular formula that can be used in CONVINcE is (Popescu Adrian, 2015):

$$B(p) \approx 1.22 * \frac{MSS}{RTT * \sqrt{p}}$$

TCP performance problems typically arise when the bandwidth * delay product is large (DARPA, IETF RFC 7323, 2014). A network having such paths is known as being a “long, fat network (LFN)”. There are two fundamental performance problems in the case of TCP over LFN, namely window size limit (the bandwidth * delay product exceeds the TCP receive window limits for maximum TCP throughput) and recovery from losses (packet losses in an LFN may have a catastrophic effect on throughput). Good solutions to these limitations are in form of larger window sizes and provision of Selective Acknowledgment (SACK) facility to handle multiple packets dropped per window (DARPA, IETF RFC 7323, 2014).

As mentioned above, determination of e2e throughput performance is of paramount importance in CONVINcE. Towards this goal, one of the most important sources of information is (IETF RFC 6349, 2011), which is about “Framework for TCP Throughput Testing”. This document recommends a methodology for measuring the TCP throughput to provide meaningful results with respect to user experience. The definition of TCP throughput is the amount of data per unit of time that TCP transports when in the TCP equilibrium state.

This document also suggests a methodology that must be performed in addition to traditional tests on the layers 2/3 below. In other words, **it is necessary to perform both layer 2/3 measurements and layer 4 measurements. Layer 2/3 measurements are focused on verification of network integrity (e.g., packet-layer tests focused on packet throughput, loss and delay measurements) whereas layer 4 measurements are focused on e2e throughput.**

The suggested testing methodology for measuring of e2e TCP throughput is as follows (Constantine B., 2011):

- Identify the Path Maximum Transmission Unit (Path MTU): this is necessary in order that TCP Throughput Test Device (TTD) is properly configured to avoid segmentation
- Baseline Round-Trip Time Bandwidth: to establish the inherent, non-congested Round-Trip Time (RTT) and the Bottleneck Bandwidth (BB) of the e2e network path
- TCP Connection Throughput Tests: with baseline measurements of RTT and BB, single- and multiple-connection throughput tests are conducted to baseline the network performance

Further technical details are provided in (Constantine B., 2011). It is also important to mention that, besides TCP, there are several more transport protocols developed for specific purposes like, e.g., Multipath TCP (MPTCP), Stream Control Transmission Protocol (SCTP), User Datagram Protocol (UDP), Lightweight User Datagram Protocol (UDP-Lite), Datagram Congestion Control Protocol (DCCP) and others (Fairhurst, 2015). Out of these, the most important protocol for our purposes is Multipath TCP (MPTCP).

7.3.4 Multipath TCP

Multipath TCP (MPTCP) is basically an extension of TCP to support multihoming, which is defined in RFC 6824 “TCP Extensions for Multipath Operation with Multiple Addresses” (Ford A., 2013). MPTCP is designed to be as transparent as possible to middleboxes, and it does so by establishing regular TCP flows between a pair of source and destination endpoints as well as multiplexing the application’s streams over these flows. By doing so, the e2e throughput is increased.

Figure 18 shows the MPTCP protocol stack.

<table>
<tr><td colspan="2"><b>Application (MPEG-DASH/HTTP)</b></td></tr>
<tr><td colspan="2"><b>Multipath TCP (MPTCP)</b></td></tr>
<tr><td><b>TCP (subflow 1)</b></td><td><b>TCP (subflow 2)</b></td></tr>
<tr><td><b>IP</b></td><td><b>IP</b></td></tr>
<tr><td><b>DLL</b></td><td><b>DLL</b></td></tr>
<tr><td><b>PHY</b></td><td><b>PHY</b></td></tr>
</table>

Figure 18: MPTCP protocol stack (Chowrikoppalu Y. P. G., 2013)

MPTCP has been designed to increase the throughput, do not harm and to balance the congestion. For doing this, MPTCP uses TCP options (TCP header) for the control plane. These options are used to signal multipath facilities as well as to negotiate data sequence numbers, advertise other available IP addresses and establish new sessions between pairs of endpoints. Basically, MPTCP provides similar features as TCP as well as the facility to multiplex a byte stream over separate e2e paths belonging the same MPTCP session. By this, higher throughput than provided by TCP can be obtained. Moreover, MPTCP provides other transport features as well. These are regarding congestion control with load balancing over multiple connections, endpoint multiplexing of a single byte stream, sub-flows can be run over IPv4 or IPv6 for the same session, resilience to network failures and/or handover (Ford A., 2013). Several deployment experiments have been reported so far (Bonaventure O., 2014) (Bonaventure O. Paasch C. and Detal G., 2015). Furthermore, popular devices today support MPTCP like, e.g., the iOS7 (Apple, April 2015).

The multi-path communication over TCP demands however for solving several problems in order to be accepted by multimedia content providers. These are related to some weaknesses observed when the underlying paths are not homogeneous. It has, e.g., been observed some performance degradations when the underlying paths are not homogeneous as well as the lack of support in middleboxes (Bonaventure, 2015). A second problem is regarding the presence of the so-called head-of-line blocking phenomenon, which is more frequent at MPTCP than at TCP. All together, these limitations may result in short abrupt throughput drops, with the consequence of limited performance in video streaming. It is however important to indicate that these are performance limitations that can be solved, e.g., as reported in (Corbillon X., 2016). In order to provide the expected quality for the e2e communication, a number of additional buffers must be considered when multi-path is used at the transport layer. These are as follows (Corbillon X., 2016; Popescu Adrian, 2015):

- Application sending buffer, of type First-Input-First-Output (FIFO), to manage the data input/output
- Multi-path sending buffer, to provide a unique socket to application as well as the individual congestion and flow controls; it is further mentioned that the individual congestion and flow controls are coupled among paths
- TCP sending buffer, to provide the global management of the TCP connection
- Receiving buffer, to queue all incoming packets from all subflows and to reorder the packet before the delivery to application
- Application receiving buffer

Furthermore, there is need to develop and investigate a cross-layer scheduler, the so-called video-content aware scheduler over MPTCP, to reorder the data transmission and to prioritize the most significant parts of video. The purpose is to improve the quality of the decoded video at the client side (Corbillon X., 2016). A good example of optimization criteria used in this case is to maximize the number of decodable video units and to improve the QoE performance.

Finally, there is need to investigate the problem of unfairness of MPTCP to other single TCP flows, which may create problems at higher level, i.e., network operation and management.

7.3.5 High-layers network optimization

High-layers adaptive video streaming is about e2e performance optimization done at the three highest layers in the TCP/IP protocol stack: application protocol (AP), application layer (AL) and transport layer (TL). Basically, the concept of high-layers performance optimization at CONVINcE is about optimization of protocols and algorithms used at the above-mentioned layers to compensate for limitations experienced in the low-layers e2e communication and to provide the expected Quality of Experience at the end user under the condition of minimum energy consumption associated with the particular video flow. Optimization algorithms are adaptive bit rate streaming, dynamic adaptive streaming over HTTP (DASH), adaptive TCP flow control also known as TCP tuning and video streaming over Multi-Path Transmission Control Protocol (MPTCP). The algorithm is as follows.

Given Live Video Streaming Scenario

Video flow model VF(a_1, a_2, a_3,..) with the parameters a_n indicating relevant statistics characterizing the selected video traffic model, e.g., mean value, peak rate, variance, Hurst parameter

Particular e2e network configuration composed, e.g., by Head End – Core Backbone Network – Wireless Network - Terminal

MPEG-DASH/HTTP/MPTCP high-layers protocol stack

MPTCP based transport service

Do Determine the number of e2e MPTCP flows needed for the particular video flow

Determine the buffer space needed at the MPTCP sender (for traffic splitting) and receiver (for recomposition)

Also Do Run experimental tests to validate the correct functioning of the e2e system, which includes MPTCP traffic sharing and reunification

Determine by analysis and/or measurements the e2e flow performance in terms of metrics relevant for the video flow; main metric is the e2e throughput

Determine the QoE performance at the end user under the condition of minimum e2e energy consumption; improve the QoE performance, if necessary

Similar constrained optimization problems can be defined for the other scenarios (On-Demand Video Streaming; Camera-Based Sensor Networks; Cloud Gaming) as well.

8 CONCLUSIONS

The goal of this report is to first define the theoretical models relevant for the video distribution scenarios in CONVINcE. This is about models like Autoregressive (AR) models, Markov-based models, self-similar traffic models, Wavelet-based models and fluid-flow models. Out of these models, two categories are considered in the future work. These are the Markov-based and fluid-flow traffic models. The next part of the report is dedicated to aspects related to video streaming and the associated e2e performance optimization. There are two categories of performance optimization in CONVINcE, which are dedicated to low-layers network optimization and high-layers adaptive video streaming. Detailed presentation of these algorithms is done.

It is also important to highlight that these optimization algorithms are one step further with regard to the suggested optimization algorithm in the document D1.1.3, which was focused only on low layers in the protocol stack.

The future work is to continue this work and develop a theoretical model for a particular scenario in CONVINcE, to do performance evaluation and to validate these results with the performance expected to be obtained in the CONVINcE testbed(s).

9 BIBLIOGRAPHY

3GPP. (2015). *LTE-Advanced Pro.* 3GPP.

Abry P. and Veitch D. (1998). Wavelet Analysis of Long Range Dependent Traffic. *IEEE Transactions on Information Theory , 44* (1), ss. 2-15.

Akhshabi S., A. L. (2012). What Happens When HTTP Adaptive Streaming Players Compete for Bandwidth? *NOSSDAV'12* (ss. 9-14). Toronto: ACM.

Alheraish A.A. (2004). Autoregressive Video Conference Models. *International Journal of Network Management , 14* (5), ss. 329-337.

Alsharif M.H., N. R. (2013). Survey of Green Communications Networks: Techniques and Recent Advances. *Journal of Computer Networks and Communications* (ID 453893), ss. 1-13.

Andy, S. (2012). *MPEG DASH: A Technical Deep Dive and Look at What's Next.* Hämtat från 2016 Spring Technical Forum CableLabs - NCTA - SCTE.

Anjum B. and Perros H. (2015). *Bandwidth Allocation for Video under Quality of Service Constraints* (ISBN 978-1-84821-746-1 uppl.). ISTE Ltd and John Wiley & Sons.

Apple. (April 2015). *iOS: multipath TCP support in iOS7*. Hämtat från Blog post.

Arlos P. (2003). *Multi-Timescale Modelling of Ethernet Traffic.* Licentiate Thesis, Blekinge Institute of Technology, Karlskrona.

Bardon, J. (2016). *CONVINcE - Encoder-Transcoder recommended settings.*

Bianzino A.P., C. C.-L. (2012). A Survey of Green Networking Research. *IEEE Communications Surveys & Tutorials , 14* (2).

Bing, B. (2010). *3D and HD Broadband Video Networking.* USA, USA: Artech House.

Bircher W.L. and John L.K. (2012). Complete System Power Estimation Using Processor Performance Events. *IEEE Transactions on Computers. 61 No. 4*, ss. 563-577. IEEE.

Blair A., P. G. (2011). A Unified Architecture for Video Delivery Over the Internet. *ISBN: 978-1-902560-25-0 © 2011 PGNet* .

Boche H., S. M. (6 2011). A Generalization of Nash Bargaining and Proportional Fairness to Log-Convex Utility Sets with Power Constraints. *IEEE Transactions on Information Theory , 57*, ss. 3390-3404.

Bonaventure O. (2014). *Experience with Multipath TCP.* IETF presentation.

Bonaventure O. Paasch C. and Detal G. (October 2015). *IETF Internet Draft, draft-ietf-mptcp-experience-03*. Hämtat från Use Cases and Operational Experience with Multipath TCP: www.rfc-editor.org

Bonaventure, O. (January 2015). *Multipath TCP through a strange middlebox*. Hämtat från Blog post.

Bosman J.W., v. d.-Q. (2012). A Fluid Model Analysis of Streaming Media in the Presence of Time-Varying Bandwidth. *International Teletraffic Conference (ITC24).* Krakow: ITS.

Camel, A. (2015). Hämtat från http://www.camel.apache.org

CEET. (2015). *Measurement Based Network Element Power Modeling.* University of Melbourne, Victoria, Australia, Center for Energy Efficient Telecommunications. Bell Labs and University of Melbourne.

Chen L., W. B. (2011). Utility-Based Resource Allocation for Mixed Traffic in Wireless Networks. *IEEE INFOCOM* (ss. 91-96). Shangai: IEEE.

Chen T., K. H. (2010). Energy Efficiency Metrics for Green Wireless Communications. *International Conference on Wireless Communications and Signal Processing* (ss. 1-6). Suzhou, China: IEEE.

Chen Y., Z. S. (June 2011). Fundamental Trade-Offs on Green Wireless Networks. *IEEE Communications Magazine* .

Chowrikoppalu Y. P. G. (2013). *Multipath Adaptive Video Streaming over Multipath TCP.* University of Saarland, Germany.

CISCO. (u.d.). *Cisco Visual Networking Index: Forecast and Methodology 2014-2019.* Hämtat från http://www.cisco.com/en/us/solutions/collateral/service-provider/ip-ngn-next-generation-network/white_paper_c11-481360.html

CISCO. (2012). *IP/MPLS Networks: Optimize Video Transport for Service Providers.* CISCO.

Claise B., P. J. (September 2014). *Energy Management Framework.* IETF.

COMBO. (2014). *Monitoring Parameters Relation to QoS/QoE and KPIs.* COMBO.

Constantine B., F. G. (August 2011). Framework for TCP Throughput Testing. *Internet Engineering Task Forec (IETF) Request for Comments (RFC) 6349* . IETF.

Corbillon X., A.-P. R. (2016). Cross-Layer Scheduler for Video Streaming over MPTCP. *7th ACM International Conference on Multimedia Systems.* New York: ACM.

Cox D.R. (1984). *Long-Range Dependence: A Review.* Iowa State University Press, Iowa.

DARPA. (2014). *IETF RFC 7323.* Hämtat från RFC 7323: www.rfc-editor.org

DARPA. (1981). *IETF RFC 793.* Hämtat från RFC 793: www.rfc-editor.org

E.800, I.-T. (1994). *Terms and Definitions Related to Quality of Service and Network Performance Including Dependability.* ITU-T.

Elemental. (2016). *Perfecting 4K Video Delivery.* Elemental.

Energy, U. D. (March 2011). Hämtat från Best Practices Guide for Energy-Efficient Data Center Design: http://www1.eere.energy.gov/femp/pdfs/eedatacenterbestpractices.pdf

Esteban J., B. S. (2012). Interactions Between HTTP Adaptive Streaming and TCP. *NOSSDAV'12.* Toronto: ACM.

Estevez R.P., G. L. (July 2013). On the Management of Virtual Networks. *Communications Magazine, IEEE* .

ETSI. (2009). *Energy efficiency of wireless access network equipment.* DTS/EE-00007 V0.0.17.

ETSI. (2011). *Environment Engineering (EE) Energy Efficiency of Wireless Access Network Equipment.* ETSI.

ETSI TR 102 64, H. F. (2009). *Quality of Experience (QoE) requirements for real-time communication services.* ETSI.

ETSI-2009. (2009). *Energy efficiency of wireless access network equipment.* DTS/EE-00007 V0.0.17.

ETSI-2011. (2011). *Environment Engineering (EE) Energy Efficiency of Wireless Access Network Equipment.* ETSI.

Fairhurst, G. a. (2015). *Services Provided by IETF Transport Protocols and Congestion Control Mechanisms.* Hämtat från Draft-IETF-Taps-Transports-07: www.rfc-editor.org

Fiedler M, P. A. (2016). QoE-Aware Sustainable Throughput for Energy-Efficient Video Streaming. *2016 IEEE International Conference on Sustainable Computing and Communications (SustainCom 2016).* Atlanta: IEEE.

Fiedler M. (2014). On the Limited Potential of Buffers to Improve Quality of Experience. *PERCOM Workshop.* Budapest: IEEE.

Fiedler M., H. T. (2010). Quality of Experience Related Differential Equations and Provisioning-Delivery Hysteresis. *21th International Teletraffic Specialists Seminar (ITC) on Multimedia Applications - Traffic, Performance and QoE.* Miyazaki, Japan: ITC.

Fiedler M., H. T.-G. (March 2010). A Generic Quantitative Relationship Between Quality of Experience and Quality of Service. *IEEE Network Communications Magazine* .

Fiedler M., S. J. (4 2014). Exponential On-Off Traffic Models for Quality of Experience and Quality of Service Assessment. *Praxis der Kommunikationstechnik (PIK) , 37*, ss. 297-304.

Fisher Y. (2014). *An Overview of HTTP Adaptive Streaming Protocols for TV Everywhere Delivery.* Hämtat från 2016 Spring Technical Forum CableLabs NCTA SCTE.

Ford A., R. C. (2013). *IETF RFC 6824.* Hämtat från RFC 6824: www.rfc-editor.org

Forum, D. (2006). *Technical Report TR-126 Triple-play Services Quality of Experience (QoE) Requirements.* DSL Forum.

Gamma E., H. R. (1994). *Design Patterns: Elements of Reusable Object-Oriented Software.* Addison Wesley.

Gelman A.D., H. S. (1991). On Buffer Requirements for Store-and-Forward Video on Demand Service Circuits. *GLOBECOM* . Phoenix, AZ, USA: IEEE.

Georgopoulos P., E. Y. (2013). Towards Network-wide QoE Fairness Using OpenFlow-assisted Adaptive Video Streaming. *FhMN'13* (ss. 15-20). Hong-Kong: ACM.

Group, Q. o. (2006). *Inter-Provider Quality of Service.* Quality of Service Working Group.

Guvenc L., Q. T.-P. (June 2013). Heterogeneous and Small Cell Networks, Part 2. *IEEE Communications Magazine , 51* (6).

Haesik, K. (2012). Next Generation Wireless Systems. Oulu: VTT.

Hamdoun H., L. P. (April 2012). Survey and Applications of Standardized Energy Metrics to Mobile Networks. i I. T. Springer-Verlag, *Annals of Telecommunications* (Vol. 67, ss. 113-123). Springer.

Han T., A. N. (Third Quarter 2013). On Accelerating Content Delivery in Mobile Networks. *IEEE Communications Surveys & Tutorials , 15* (No. 3).

Hasan Z., B. H. (2011). Green Cellular Networks: A Survey, Some Research Issues and Challenges. *IEEE Communications Surveys & Tutorials , 13* (4).

Heindl A. (2000). Decomposition of General Tandem Queueing Networks with MMPP Input. i e. a. Haverkort B.R., & S.-V. B. Heidelberg (Red.), *TOOLS 2000* (ss. 86-100).

Hossain E., B. B. (2012). *Green Radio Communication Networks.* Cambridge University Press.

Hossfield T., F. M. (2011). The QoE Provisioning-Delivery-Hysteresis and Its Importance for Service Provisioning in the Future Internet. *7th Euro-NGI Conference.* Kaiserlautern: IEEE.

Husain Bohra A.E.H. and Chauhdary V. (2010). VMeter: Power Modeling for Virtualized Clouds. *IEEE International Workshop on Parallel and Distributed Processing (IPDPSW).* IEEE.

Ickin S. (2015). *Quality of Experience on Smartphones: Network, Application, and Energy Perspectives.* PhD thesis, Blekinge Institute of Technology.

IETF RFC 6349. (2011). Hämtat från RFC 6349: www.rfc-editor.org

IETF, R. 6. (August 2011). *Framework for TCP Throughput Testing.* Hämtat från Internet Engineering Task Forec (IETF): https://tools.ietf.org/html/rfc6349

Iraj, S. (oct-Nov 2011). MPEG-DASH: The Standard for Multimedia Streaming Over Internet. *IEEE Multimedia* .

ISO/IEC 23009-1. (2011). *Dynamic Adaptive Streaming Over Internet.* Hämtat från Short tutorial on MPEG-DASH.

ISO/IEC. (2012). *Information Technology - Dynamic Adaptive Streaming over HTTP (DASH) - Part 1: Media Presentation Description and Segment Formats .* 23009-1.

ITU-T, E. (1994). *Terms and Definitions Related to Quality of Service and Network Performance Including Dependability.* ITU-T.

ITU-T, R. (2003). *ITU-T Rec. P.800.1: Mean Opinion Score (MOS) Terminology.* ITU-T.

Kaup F., H. D. (2013). Optimizing Energy Consumption and QoE on Mobile Devices. *21st IEEE International Conference on Network Protocols (ICNP).* IEEE.

Kharitonov, D. (2012). Green Telecom Metricvs in Perspective. *18th Asia-Pacific Conference on Communications (APCC)* (ss. 548-553). IEEE.

Kim M., J. Y. (2014). A Simpe Model for Estimating Power Consumption of a Multicore Server System. *International Journal of Multimedia and Ubiquitous Engineering*, *9, No.2*, ss. 153-160.

Krunz M. and Makovski A. (June 1998). Modeling Video Traffic Using m/g/infinity Input Processes: A Compromise Between Markovian and LRD Models. *IEEE Journal on Selected Areas in Communications* , *16* (5), ss. 733-748.

Lange C., K. D. (2011). Energy Consumption of Telecommunication Networks and Related Improvement Options. *IEEE Journal of Selected Topics in Quantum Electronics* , *17* (2), 285-295.

Lara A., K. A. (First Quarter 2014). Network Innovation Using OpenFlow: A Survey. *IEEE Communications Surveys & Tutorials, Vol. 16, No. 1.* IEEE .

Le Callet P., M. S. (2012). Qualinet White Paper on Definitions of Quality of Experience. *European Network on Quality of Experience in Multimedia Systems and Services* . Lausanne: COST Action IC 1003.

Lederer S., M. C. (2012). Av Evaluation of Dynamic Adaptive Streaming over HTTP in Vehicular Environments. *ACM Multimedia Systems Conference* . ACM .

Lei L., Z. Z. (June 2013). Challenges in Wireless Heterogeneous Networks for Mobile Cloud Computing. *IEEE Wireless Communications* .

Lent R. (2011). Evaluating the Performance and Power Consumption of Systems with Virtual Machnies. *Third IEEE International Conference on Cloud Computing Technology and Science* (ss. 778-783). IEEE.

Liberal F., T. I.-O. (2013). Dealing with Energy-QoE Trade-Offs in Mobile Video. *Journal of Computer Networks and Communications* .

Möbius C., D. W. (June 2014). Power Consumption Estimation Models for Processors, Virtual Machines, and Servers. *IEEE Transactions on Parallel and Distributed Systems* , *25* (6), ss. 1600-1614.

Mannersalo P., N. I. (2002). *Multifractal Products of Stochsstic Processes: Construction and Some Basic Properties.* Applied Probability Trust.

Mediaentertainmentinfo. (u.d.). Hämtat från Mediaentertainmentinfo.com: http://www.mediaentertainmentinfo.com/2013/04/2-concept-series-what-is-the-difference-between-ott-and-iptv.html

Minoli D. (2011). *Designing Green Networks and Network Operations: Saving Run-the-Engine Costs* (ISBN 9781439816387 uppl.). CRC Press.

Mok R., C. E. (2011). Measuring the Quality of Experience of HTTP Video Streaming. *IFIP/IEEE International Symposium on Integrated Network Management.* IFIP/IEEE.

Mok R., L. X. (2012). QDASH: a QoE-Aware DASH System. *3rd Multimedia Systems Conference.* New York: ACM.

Monnier R., P. A. (2016). CONVINcE: Towards Power-Optimized Video Distribution Networks. *19th ICIN Conference.* Paris: ICIN.

Monnier, R. (2014). *CONVINcE Consumption Optimization in Video Networks.*

Monnier, R. e. (2014). CONVINcE: Consumption Optimization in Video Networks. *CELTIC-PLUS project CP2013/2-1, http://celticplus.eu/Projects/Project-info/PI-call_2013.asp.*

MPEG-DASH. *http://en.wikipedia.org/wiki/Dynamic_Adaptive_Streaming_over_HTTP.*

Nakao A., A. T. (2012). *Advanced Network Virtualization: Definition, Benefits, Applications, and Technical Challenges.* White Paper, Version 1.0.

Ohm J-R., S. G. (December 2012). Comparison of the Coding Efficiency of Video Coding Standards - Including High Efficiency Video Coding (HEVC). *IEEE Transactions on Circuits and Systems for Video Technology , 22* (12), ss. 1669-1684.

Open, N. F. (u.d.). *Software-Defined Networking: the New Norm for Networks*. (ONF, Producent) Hämtat från https://www.opennetworking.org

OpenStack. http://www.openstack.org.

P.10/G.100, I.-T. R. (2006). *Vocabulary for Performance and Quality of Service.* ITU-T.

Padhye J., F. V. (1998). Modeling TCP Throughput: A Simple Model and its Empirical Validation. *ACM Sigcomm 1998 conference on Applications, Technologies, Architectures and Protocols for Computer Communication (SIGCOMM 1998).* New York, USA.

Papazoglou, M. (2012). *Web Services & SOA: Principles and Technology.* Pearson Education Limited.

Popescu A. (2014). Greening of IP-Based Video Distribution Networks: Developments and Challenges. *International Conference on Communications.* Bucharest, Romania: IEEE.

Popescu A. (2008). *Traffic Analysis and Control in Computer Communications Networks.* BTH.

Popescu A., a. a. (2016). *High-Level Architecture Design.*

Popescu A., e. a. (2015). *CONVINcE D1.1.1 Application Scenarios.*

Popescu A., e. a. (2015). *CONVINcE D1.1.1 Application Scenarios.*

Popescu A., e. a. (2015). *CONVINcE D1.1.2 System Requirements.*

Popescu A., e. a. (2016). *CONVINcE D1.1.3 High-Level Architecture Design.*

Popescu Adrian. (2015). Lecture on "Transport Layer". *Course on "Internetworking with TCP/IP"* .

Popescu Adrian, E. D. (2011). ROMA: A Middleware Framework for Seamless Handover. i D. K. Kouvatsos (Red.), *Next Generation Internet: Performance Evaluation and Applications* (Vol. Lecture Notes in Computer Science 5233). Springer Verlag.

Rao A., L. Y. (2011). Network Characteristics of Video Streaming Traffic. *SIGCOMM.* ACM.

Ravindra K. Ahuja, T. L. (1993). *Network Flows: Theory, Algorithms and Applications.* USA: Prentice-Hall.

RFC, 4. (2006). *IETF - Internet Engineering Task Force.* Hämtat från BGP/MPLS IP Virtual Private Networks (VPNs): http://www.rfc-editor.org

Rostami A. (2014). Software Defined Networking. *ITC 26.*

Sezer S., S.-H. S. (July 2013). Are We Ready for SDN? Implementation Challenges for Software-Defined Networks. *IEEE Communications Magazine* (ss. 36-43). IEEE.

Shaikh J. (2015). *Network-Based Monitoring of Quality of Experience.* Blekinge Institute of Technology, Telecommunication Systems. Karlskrona: BTH.

Shaikh J., F. M. (2012). Modeling and Analysis of Web Usage and Experience Based on Link-Level Measurements. *International Teletraffic Conference (ITC) 24. 24.* Cracow, Poland: ITC.

Sodagar I. (2011). The MPEG-DASH Standard for Multimedia Streaming over the Internet. *IEEE MultiMedia* (18), ss. 62-67.

Soldani, D. (den 21-22 September 2010). *Bridging QoE and QoS for Mobile Broadband Networks.* Hämtat från ETSI: http://www.etsi.org/WebSite/NewsandEvents/QoSQoEUserExperience.aspx

Storage, B. a. (2016). *Bandwidth and Storage.* Hämtat från http://www.broadbandchoices.co.uk/guides/internet/watching-tv-online

Suarez L., N. L.-M. (2012). An Overview and Classification of Research Approaches in Green Wireless Networks. *EURASIP Journal on Wireless Communications and Networking , 2012:142*.

Tanwir S. and Perros H. (Fourth Quarter 2013). A Survey of VBR Video Traffic Models. *IEEE Communications Surveys & Tutorials , 15* (4).

Tanwir S. and Perros H. (2014). *VBR Video Traffic Models.* Wiley.

Tudor B.M. and Teo Y.M. (2013). On Understanding the Energy Consumption of ARM-based Multicore Servers. *SIGMETRICS'13* (ss. 267-278). NY, USA: ACM.

Wang B., K. J. (2004). Multimedia Streaming via TCP: an Analysis Performance Study. *12th Annual ACM International Conference on Multimedia* (ss. 908-915). New York, USA: ACM.

VCODEX. (den 14 04 2016). *Historical Timeline of Video Coding Standards and Formats.* Hämtat från https://www.vcodex.com/historical-timeline-of-video-coding-standards-and-formats

Vereecken W., H. W. (June 2011). Power Consumption in Telecommunication Networks: Overview and Reduction Strategies. *IEEE Communications Magazine* .

Wilde J. (2011). *Constrained Optimization.* http://www.econ.brown.edu.

Xylomenos G., M. P. (April 2001). TCP Performance Issues over Wireless Links. *IEEE Communications Magazine , 39* (4).

Yao Y. (2014). *A Software Framework for Prioritized Spectrum Access in Heterogeneous Cognitive Radio Networks.* Blekinge Institute of Technology, PhD thesis.

Zambelli A. (2009). *IIS Smooth Streaming Technical Overview.* Microsoft Corporation.

Zhang X., Z. J. (2013). On the Study of Fundamental Trade-offs Between QoE and Energy Efficiency in Wireless Networks. *Transactions on Emerging Telecommunications Technologies , 24*, 259-265.