Context-aware Video Encoding as a Network-based Media Processing (NBMP) Workflow

Christoph Mueller
christoph.mueller@fokus.fraunhofer.de
Fraunhofer FOKUS
Berlin, Germany

Louay Bassbouss
louay.bassbouss@fokus.fraunhofer.de
Fraunhofer FOKUS
Berlin, Germany

Stefan Pham
stefan.pham@fokus.fraunhofer.de
Fraunhofer FOKUS
Berlin, Germany

Stephan Steglich
stephan.steglich@fokus.fraunhofer.de
Fraunhofer FOKUS
Berlin, Germany

Sven Wischnowsky
sven.wischnowsky@telekom.de
Deutsche Telekom AG
Berlin, Germany

Peter Pogrzeba
peter.pogrzeba@telekom.de
Deutsche Telekom AG
Berlin, Germany

Thomas Buchholz
thomas.buchholz@telekom.de
Deutsche Telekom AG
Berlin, Germany

ABSTRACT

Leveraging processing capabilities and resources in a network is a trending approach in accomplishing complex media processing tasks. At the same time, efficiently utilizing available resources while ensuring the potential for scalability and distribution is key. However, deploying, operating and maintaining such complex media service workflows on different cloud services, at the edge or on-premise can be a very complex and time-consuming task. In this paper, we will present an approach that addresses these challenges by utilizing state-of-the-art technologies and standards for advanced multimedia services such as the MPEG Network-based Media Processing (NBMP) standard. We will apply the presented approach for implementing bandwidth reduction and optimization strategies by using context aware video encoding. Implemented as an automated NBMP workflow, the context aware encoding method with the support of machine learning models avoids computationally heavy test encodes. The models are trained on complex datasets composed of 40+ video attributes and generate an optimal encoding ladder as an output (bitrate/resolution pairs). In comparison to the conventional per-title encoding method, we observed significant savings in terms of storage and delivery costs, while maintaining the same visual quality.

CCS CONCEPTS

• **Networks** → *Network architectures*; • **Computing methodologies** → *Artificial intelligence*; **Machine learning**.

MMSys'22, June 14–17, 2022, Athlone, Ireland

ACM ISBN 978-1-4503-9283-9/22/06.
https://doi.org/10.1145/3524273.3533250

KEYWORDS

Context-aware Video Encoding, Network-assisted Media Distribution, Network-based Media Processing, Machine Learning

ACM Reference Format:
Christoph Mueller, Louay Bassbouss, Stefan Pham, Stephan Steglich, Sven Wischnowsky, Peter Pogrzeba, and Thomas Buchholz. 2022. Context-aware Video Encoding as a Network-based Media Processing (NBMP) Workflow. In *13th ACM Multimedia Systems Conference (MMSys '22), June 14–17, 2022, Athlone, Ireland.* ACM, Athlone, Ireland, 6 pages. https://doi.org/10.1145/3524273.3533250

1 INTRODUCTION

In 2021, 82 % of consumer Internet traffic is made up of video streaming [8]. In 2019, 63% of global mobile bandwidth was used for video delivery, predicted to reach 76% in 2025 [2]. There is a continual yearly grow of video traffic on the internet mainly caused by social media, gaming, VoD and linear TV. Upcoming technologies such as 8K and XR with a demand for higher bandwidth due to their high resolution will strengthen this trend. Increasing peak time traffic during live events such as UEFA EURO 2020 or Olympic Games leads to shortages of bandwidth mainly in mobile networks. All that shows that there is a need to optimize the bandwidth resources in fixed and mobile networks for video streaming services.

To overcome the challenges faced in terms of processing complex media processing tasks, as well as address the aforementioned motivational factors, we propose a context aware encoding pipeline, a new workflow using Network-based Media Processing (NBMP) technology with an integrated AI-enhanced video analysis component. This architecture can not only utilize available resources efficiently, but with the help of context aware encoding, maintain a high QoE for the end customer as well, while reducing bandwidth costs.

In this paper, the proposed methodology will be applied to two primary use cases:

- Network-assisted media distribution: Using NBMP technology for efficient, dynamic and flexible media distribution.

- Distributed encoding and context aware encoding.

Leveraging processing capabilities and resources in the network is a trending approach to achieve complex media processing with more resource efficiency and ensuring scalability. Deploying such complex services on top of micro-services, while each one of them could be running on a different Cloud, on an Edge, or on-premise, can be a very complex and time-consuming task. Figure 1 shows the overall NBMP architecture of the Network-assisted Media Distribution use case.

Realized as a NBMP workflow, for distributed encoding and context aware encoding Live or VoD sources are encoded using AI-based image processing and metadata extraction. These results in dynamic changing of bitrate depending on content characteristics. This technology creates optimal encoding bitrate ladders for individual streams and leads to a significant bitrate reduction, including lower network traffic and cost reduction.

The remainder of the paper is structured as follows: in Section 2, an overview of related work is presented. In Section 3, context-aware encoding is introduced, including key features and challenges. Further we discuss how the context-aware encoding use case is realized as a NBMP workflow. The paper then concludes with an outlook and summary.

2 RELATED WORK

The NBMP MPEG standard [6] provides a flexibility that is necessary for deploying complex media workflows on top of micro-services. NBMP defines interfaces, media, and metadata formats to address fragmentation and offer a unified way to perform media processing on top of any cloud platform and on any IP network. NBMP, illustrated in Figure 2, is composed of the following entities:

- Task: Implements one media processing function. Once a Function is loaded, it becomes a Task, which is then configured by the Workflow Manager through the Task API and can begin processing incoming media.
- Function Repository: Offers APIs for function discovery and the load of function description documents.
- Workflow Document: A document that includes a description of a Directed Acyclic Graph (DAG), where each node of the DAG represents a processing Task in the Workflow.
- Workflow Manager: Parses the Workflow Document and composes it together by initiating and configuring the involved functions.

Cock et. al. [1] describe a method to perform per-title encoding for video on demand (VoD) content. Due to the fact that diverse video content requires different bitrates and encoding settings in order to achieve a certain quality, the team introduced a context-aware encoding method where a recommended encoding ladder was tailored for each video, rather than applying the same encoding settings across all types of videos of a certain resolution size. Based on a complexity analysis, each source video is encoded in different parameters and resulting bitrate/resolution pairs are plotted on a graph, forming a curve called the 'convex hull'. The optimal encoding ladder is then derived from pairs positioned closest to the curve. They further improve this approach by applying a chunk-based multi-pass encoding process. As a result, title- and chunk-based encoding approaches outperformed conventional approaches in terms of storage savings and video quality.

As a follow-up to Netflix's per-title encoding method, Amazon Web Services (AWS) released a "Automated ABR Configuration" feature in AWS Elemental MediaConvert, a video transcoding service [4]. Like Netflix's per-title encoding method, an optimal encoding ladder is generated for each video input. However, this particular workflow uses a different starting point, Quality-Defined Variable Bitrate encoding mode, in which the output ladder is further optimized by removing repetitive renditions, and a sufficient resolution is found for each rendition.

3 CONTEXT-AWARE ENCODING

To enhance the media distribution workflow with AI methodologies, we propose a Context-aware Encoding Pipeline, in which the AI-driven video encoding toolset *Deep Encode* is operated by NBMP technology for VoD and Live scenarios. The idea is to shift the encoding/transcoding to the network edges and to generate bitstreams matching the available bandwidth of the last mile in a mobile network. This would lead to an efficient usage of the available bandwidth. Another application field is the dynamic adaption of the bitrate depending on the Quality of Service (QoS) requirements of a user.

3.1 Challenges of existing video encoding approaches

Video streaming content differs in terms of complexity and requires title-specific encoding settings to achieve a certain visual quality. Conventional "one-size-fits-all" encoding ladders ignore video-specific characteristics and apply the same encoding settings to all video files. In the worst-case scenario, this approach can lead to quality impairments, encoding artifacts or unnecessary large media files.

To combat these challenges, the concept of *per-title encoding* was introduced. Conventional per-title encoding solutions, such as the one originally by Netflix in 2015 [7], identify the optimal encoding settings for a single asset by creating multiple so-called *test-encodes* and determining their visual quality. Each test-encode is an encode of the original video asset, with varying encoding settings such as bitrate and resolution, ranging from very low-bitrate and low-resolution encodes, to high definition variants. After the test-encodes are created, the visual quality of each encode is determined by using common quality metrics such as VMAF (Video Multi-Method Assessment Fusion), a perceptual video quality assessment algorithm developed by Netflix [3]. Based on the test-encodes, the optimal encoding settings for the original video asset can be determined by calculating the convex hull for each resulting bitrate- and quality-value-pair and selecting test-encodes closest to the convex hull, which results in an optimized encoding ladder (multiple bitrate-resolution variants) for a single source asset.

While this approach allows the calculation of an exact encoding ladder for any given source video, this method quickly becomes computationally expensive as well as time consuming and takes up a large amount of disk space. As a result, the brute-force approach

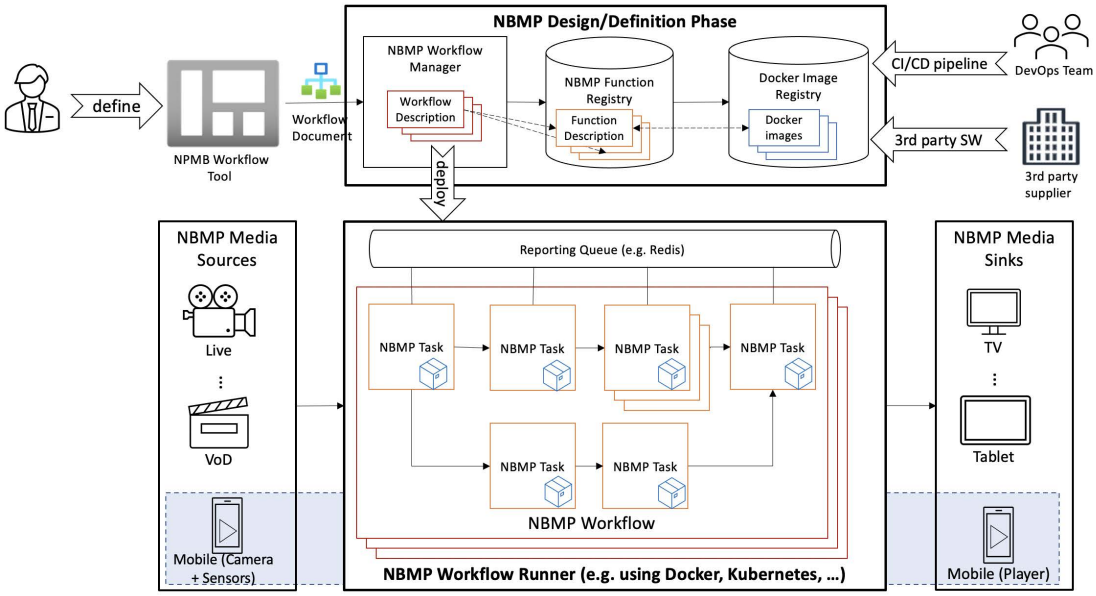

Figure 1: NBMP architecture and workflow overview

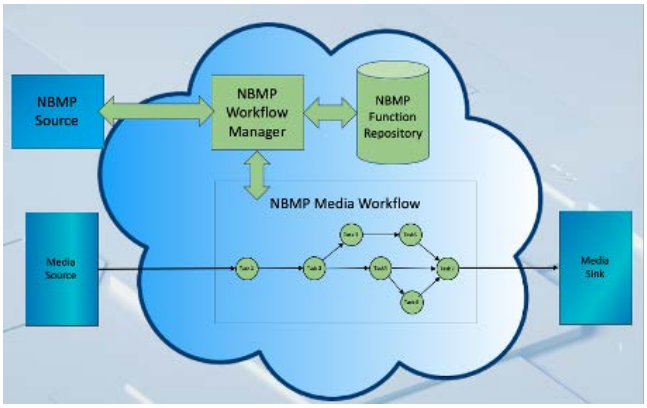

Figure 2: NBMP Network Entities Overview [9]

is inefficient for live-streaming scenarios and can prove costly, especially for smaller streaming companies.

3.2 Deep Encode - AI-supported context aware video encoding

In order to optimize the shortcomings of traditional per-title encoding approaches and, above all, to address the problem of costly test encodes, we developed *Deep Encode*, a solution that uses machine learning methods. The goal of this approach is to completely eliminate the need for test encodes and preceding quality metric calculations and instead use statistical predictions to determine the optimal encoding parameters for any given output video. These automated predictions are based on algorithms that use training data sets to develop statistical models that detect patterns and regularities in videos. The system uses multiple machine learning techniques, such as convolutional neural networks (CNNs) for image analysis, as well as models combining multiple approaches like *extreme gradient boosting* (XGBoost) and fully-connected, feed-forward neural networks (FF-FNNs) for regression and prediction tasks. These models are used to efficiently evaluate new, previously unknown videos. The trained models are able to make statements about the perceptual quality of different quality levels based on extracted characteristics of a video and can predict the bitrate and quality of a video based on a given set of encoding settings. The models were continuously trained and refined on large data sets of video content, and resulted in additional bitrate savings while maintaining the same perceptual quality. As such, the Deep Encode also observed significant savings in terms of computational time and storage. With this new approach, the previously necessary test encodes are no longer required.

Deep Encode is built as an API-driven micro-service architecture to enable automatic and effortless scalability. It consists of several worker instances that are responsible for video analysis and data extraction and video encoding, as seen in fig. 3. This architecture and implementation approach allows Deep Encode to be easily integrated into an NBMP workflow, which will be described in detail in the next chapter.

3.3 Overview: NBMP Workflow for Context Aware Video Encoding

Fig. 4 illustrates an overview of a live-streaming scenario using Deep Encode as a Context-aware Encoding Pipeline. The "NBMP Media Sources" represents the streaming URL, of which, is split into 5-second segments prior to the analysis step. An ingest worker (not

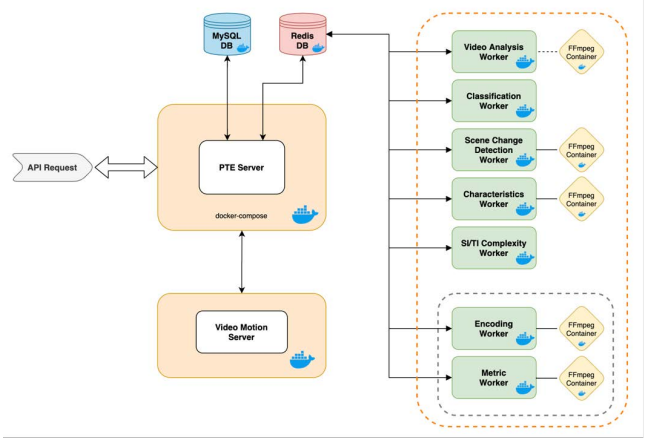

Figure 3: High-level Deep Encode architecture overview

Video Information	
Filename	11729_001__4-scenes-full-frame.mpg
Format	MPEG-PS
File size	3866 MiB
Duration	1 min 12 sec
Video Bitrate	53.7 Mb/s (uncompressed)
Format profile	4:2:2@High
Format settings, GOP	Variable
Bit rate mode	Constant
Resolution	1920x1080
Display aspect ratio	16:9
Framerate	25 fps
Chroma subsampling	4:2:2
Bit depth	8 bits
Scan type	Progressive

Table 1: Input Video Metadata (raw source)

pictured) processes the live source to trigger Deep Encode. Each segment is analyzed in terms of its complexity, which is operated by the worker instances.

As described in the previous chapter, this analysis eliminates the need for the test encodes that are normally created in the conventional per-title encoding process, and consists of extracting key features from the video, such as metadata and unique video features. The resulting values stemming from the complexity analysis are then sent to a machine learning model, which then predicts an optimal encoding ladder, made up of bitrate/VMAF pairs for each resolution. "Video Multi-method Assessment Fusion" (VMAF) is a video quality metric that was also developed by Netflix to capture the perceptual quality, which is commonly lacked in other video quality measures, such as the Peak Signal-to-Noise Ratio (PSNR) metric. VMAF predicts subjective video quality between the source and encoded video on a scale of 1 – 100, with 100 representing perfect quality (indistinguishable from the source video file).

Fig. 5 illustrates how the Context-aware Encoding Pipeline can be integrated in a telecommunications operator infrastructure. After the pipeline analysis is completed, each segment is encoded with its corresponding predicted bitrate/resolution pair. Based on the fact that the pipeline is a content-aware encoding process, only the needed encoding parameters are created. The encoded segment is then sent to the packager and made available to the content delivery network (CDN), of which, the live stream can then be played on any device.

3.4 Demo

As part of the demo track, we intend to present the end-to-end workflow combining DeepEncode with NBMP. The demo features a hands-on UI, showcasing the features of Deep Encode itself as well as the benefits that are gained from utilizing DeepEncode as an NBMP workflow. The hands on UI contains, among others: two video players with live video streams running simultaneously, while relevant metadata about the core technologies (such as: adapted video encoding ladders, NBMP workflow configuration, power consumption, etc.) is displayed and updated automatically as the live stream progresses (see Fig. 6).

3.5 Assessment

3.5.1 Test Setup. To validate the Context-Aware Encoding Pipeline within the NBMP workflow and to make the assessment reproducible, we set up a clean EC2 instance (*c5a.4xlarge*). The only processes running on this machine were the components of DeepEncode and the other necessary modules for the NBMP tasks, as seen in fig. 4.

To assess the potential bitrate savings of the the Context-aware Encoding Pipeline, we used a high resolution source file with a total length of 1min and 12s, consisting of four scenes in total (as detailed in table 1). The source file was made available to the NBMP workflow as a continuous live source, which simulates a live stream (with four recurring scenes). This source file is then analyzed with a standardized per-title encoding approach (as described in [7]) and with the Context Aware Encoding Pipeline.

Upon starting the live stream, the Context Aware Encoding Pipeline in the NBMP workflow automatically picks up the live source, detects scenes in the video and finally predicts the optimal encoding settings for each of the previously detected scenes. Table 2 presents the first four predicted encoding ladders. One scene with high redundancy and low movement ("Scene 1") only required 1141 kb/s, while another scene with low redundancy and a high amount of movement required a higher bitrate of 3316 kb/s at 1080p ("Scene3"), both achieving the same predicted VMAF quality of 90.

3.5.2 Test Results. Table 3 compares the results in terms of overall file size, bitrate, and VMAF quality. In order to compare the videos encoded by the Context-aware Encoding Pipeline to the reference files, the target quality (VMAF) within DeepEncode was adjusted to match the input files, and the VMAF was again calculated for each video, to ensure the videos had the same visual quality.

In comparison to the conventional per-title encoding method, at 1080p, the Context-aware Encoding Pipeline was able to save 20% in both bitrates and file sizes, while achieving the same quality. Based on these results, it can be assumed that at the same perceptual quality of 90, predicting an encoding ladder for each segment can

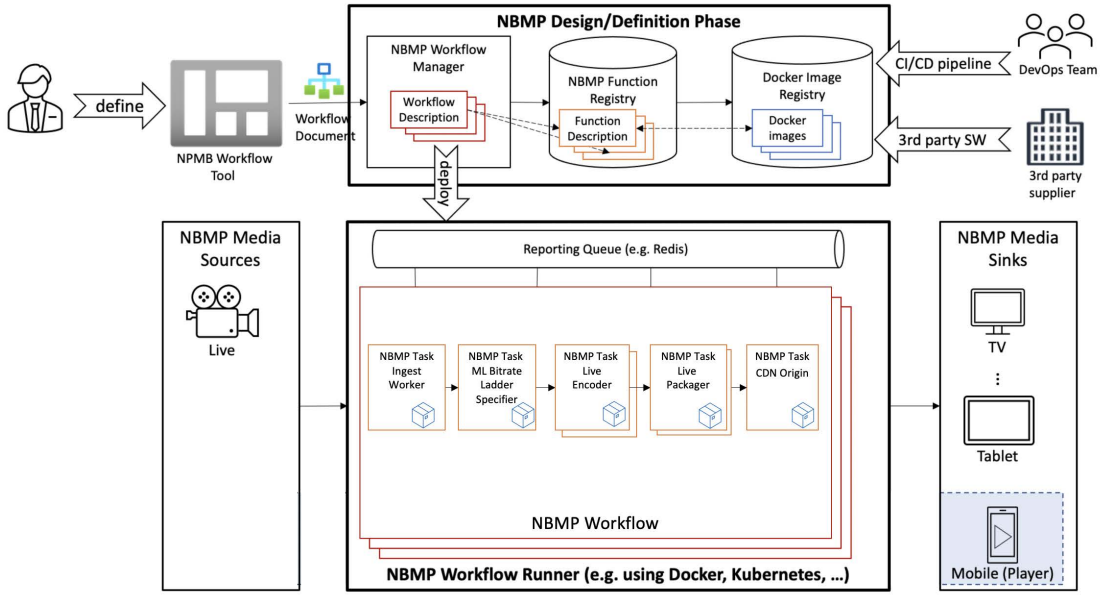

Figure 4: Context-aware Encoding Pipeline as an NBMP workflow

	Scene 1		Scene 2		Scene 3		Scene 4	
Resolution	**Bitrate (kb/s)**	**VMAF**	**Bitrate (kb/s)**	**VMAF**	**Bitrate (kb/s)**	**VMAF**	**Bitrate (kb/s)**	**VMAF**
480x270	200	40	594	59	555	40	200	63
640x360	200	60	1080	66	798	60	200	71
960x540	403	75	1400	75	1258	75	284	75
1280x720	600	80	2195	85	1668	80	600	85
1920x1080	1141	90	3915	90	3316	90	1517	91

Table 2: Video encoding ladder and predictions from the Context Aware Encoding Pipeline for scenes 1-4

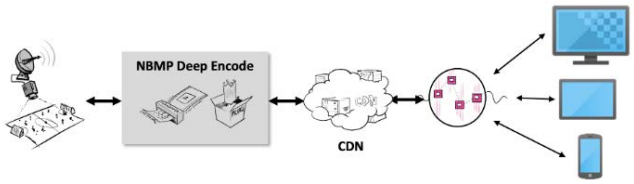

Figure 5: Context-aware Encoding Pipeline workflow in a telecommunications operator infrastructure

Method	**Bitrate (kb/s)**	**VMAF**	**File size (MByte)**
Context-aware Encoding (AI PSE)	2472	90.07	17.8
Conventional Per-Title Encoding	3111	89.79	22.4
Conventional Encoding (VBR)	5800	90.04	41.8

Table 3: Comparison: Bitrates, Quality (VMAF) and file sizes for different encoding approaches for the encoded input source clip (only 1080p)

save significant bitrates and storage in comparison to predicting an encoding ladder for an entire video source. Furthermore, distributed encoding as an NBMP workflow (in contrast to a static setup used in [7]) allows for optimization of bitrate usage in mobile networks, where the bitrate and resolution is dynamically adapted based on the Quality of Service (QoS) demands of the user.

3.6 Remaining encoding workflow challenges

After the encoding process, a packager is responsible for producing DASH and HTTP Live Streaming (HLS) streams. The two streaming formats enable player support across all relevant devices in today's over-the-top (OTT) ecosystem. The challenge for players that comes with Live encoding is that the bitrate of a quality level can vary significantly across different scenes. The ABR logic within players assumes a stable bitrate (with a certain margin) to make bitrate switching decisions. Changing the bitrate too much in an upcoming

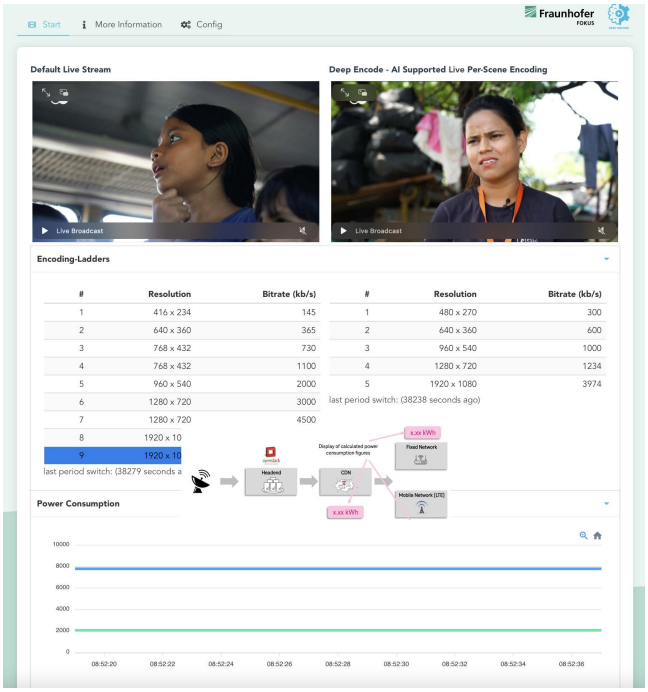

Figure 6: Demo UI showcasing the combination of DeepEncode and NBMP (livestream overview page)

scene, especially when the bitrate is increased, can lead to buffer underruns, because the ABR logic assumes that there is sufficient bandwidth for the current scene's bitrate.

To deal with this challenge, we package scenes as multiple periods for DASH. This approach is often used for ad-insertion, where the advertisement can be encoded differently from the main content. This way, a DASH player and its ABR logic is prepared to consider different bitrate ladders for each scene/period. For native HLS playback on Apple devices the bitrate variation is limited by the HLS Authoring Specification. [7] As a result, for HLS, we can only vary the bitrate across scenes within certain thresholds.

Additionally, the use of the Context-Aware encoding pipeline introduced additional processing time to calculate the complexity of the input and generate the prediction for optimal encoding settings of around 1.5 seconds. However, the additional time can be further decreased by adjusting the granularity and accuracy of the extracted data from the source video. More work is needed to determine the lowest possible delay to real-time is achievable with these adjustments, especially for critical live-streaming events such as major sports-events.

4 CONCLUSION

The majority of Internet traffic and mobile bandwidth is made up of video streaming and is expected to increase significantly by 2025. With this trend, energy consumption and costs grow exponentially in parallel. Using the UEFA Euro 2024 game as an example, assuming that one million streaming customers will be watching all 51 games, the overall energy consumption would be 20,4 GWh for HD quality and 61,2 GWh for UHD [5]. Using the Context-aware Encoding Pipeline as an NBMP Workflow for these scenarios could not only save up to 20% more bitrate than conventional per-title encoding approaches (resulting in over 50% potential cost-savings for streaming providers), but also significantly reduce the carbon footprint of online video streaming, by eliminating unneeded test-encodes and minimizing the bitrate of streaming video, while maintaining the same visual quality.

Optimizing processing capabilities and resources while reducing bandwidth is a growing need for streaming service providers. The Context-aware Encoding Pipeline described in this paper uses NBMP technology to operate the complex media workflows, with an integrated component named Deep Encode. Deep Encode is the analysis component that enhances the conventional per-title encoding method. By using machine learning methods to predict optimal encoding settings based on video content complexity, it completely eliminates the previously needed computationally heavy test encodes. With the optimized predictions the Context-aware Encoding Pipeline was able to save 20% of bandwidth and storage space compared to conventional per-title encoding approaches while maintaining the same video quality. Combining the potential reductions in bitrate and storage with the effortless and automatic up and down-scaling of multiple machines in distributed networks within an NBMP workflow results in significant savings for streaming service providers in charge of running encoding pipelines.

REFERENCES

[1] Jan De Cock, Zhi Li, Megha Manohara, and Anne Aaron. 2016. Complexity-based consistent-quality encoding in the cloud. In *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1484–1488.
[2] Ericsson. 2020. *Ericsson Mobility Report June 2020.* Retrieved January 12, 2022 from https://www.ericsson.com/49da93/assets/local/mobility-report/documents/2020/june2020-ericsson-mobility-report.pdf
[3] Zhi Li et. al (Netflix). 2016. *Toward A Practical Perceptual Video Quality Metric.* Retrieved April 15, 2022 from https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652
[4] Amazon Gehred, D. 2021. *Encode video in a smarter way using Automated ABR.* Retrieved January 13, 2022 from https://aws.amazon.com/de/blogs/media/introducing-automated-abr-adaptive-bit-rate-configuration-a-better-way-to-encode-vod-content-using-aws-elemental-mediaconvert/
[5] Borderstep Institut. 2020. *Videostreaming: Energiebedarf und CO2-Emissionen.* Retrieved January 18, 2022 from https://www.borderstep.de/wp-content/uploads/2020/06/Videostreaming-2020.pdf
[6] The Moving Picture Experts Group (MPEG). 2021. *Network Based Media Processing.* Retrieved January 18, 2022 from https://mpeg.chiariglione.org/standards/mpeg-i/network-based-media-processing
[7] Netflix. 2015. *Per-Title Encode Optimization.* Retrieved January 22, 2022 from https://netflixtechblog.com/per-title-encode-optimization-7e99442b62a2?gi=5b4a98028f83
[8] Cisco Systems. 2021. *Cisco Global 2021 Forecast Highlights.* Retrieved January 12, 2022 from https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2021_Forecast_Highlights.pdf
[9] ISO/IEC 23090-8:2020 Information technology. 2020. *Coded representation of immersive media — Part 8: Network based media pro-cessing.* Retrieved January 10, 2022 from https://www.iso.org/standard/77839.html