A Survey on Virtual Network Functions for Media Streaming: Solutions and Future Challenges

Roberto Viola, Angel Martin, Mikel Zorrilla, Jon Montalbán, *Senior Member, IEEE*,
Pablo Angueira, *Senior Member, IEEE* and Gabriel-Miro Muntean, *Senior Member, IEEE*

***Abstract*—Media streaming services rely heavily on good and predictable network performance when delivered to large numbers of people. Media services need to ensure enhanced user perceived quality levels during content playback to attract and retain audiences, especially while the streams are distributed remotely via networks. Furthermore, as the quality of media content gets higher, the network performance demands are also increasing and it is challenging to meet them. To this end, Content Delivery Networks (CDN) employ diverse solutions to empower media streaming, including state-of-the-art streaming technologies, such as HTTP adaptive streaming (HAS). Network functions should help to further enhance media streaming services and cope with the high dynamics of network performance and user mobility. Furthermore, new networking paradigms and architectures under the 5G networks umbrella are bringing new possibilities to deploy smart network functions, which monitor the media streaming services through live and objective metrics and boost them in real time. This survey overviews the state-of-the-art technologies and solutions proposed to apply new network functions for enhancing quality of service (QoS), quality of experience (QoE) and other business and energy metrics in the context of media streaming.**

***Index Terms*—Media streaming, network functions, quality of service, network virtualization, network traffic, network forecast.**

I. Introduction

In the recent years, media streaming traffic is constantly growing. Wireless and mobile devices are becoming main sources for both rich media content generation and consumption. 5G networks must cope with this new traffic demand supporting higher bandwidth and reduced latency. It is estimated that 5G connections will handle nearly three times more traffic than a current LTE connection by 2023 [1]. New applications involving video streams are gaining relevance and are attracting an increased audience, including in areas in which there was little or no rich media presence. Examples of professional applications and application areas which can benefit from advanced media streaming include Industrial Internet of Things (IIoT), medical equipment and connected and autonomous vehicles. Moreover, 3D video formats enable support for new services, such as eXtended Reality (XR), Virtual Reality (VR) and Augmented Reality (AR). Finally, online gaming and video conferencing are also highly popular, especially in the last period. These services have increasing demands in terms of network support. However, although the networks have growing capabilities, there is a large increase in rich media streaming traffic, mostly fueled by the global COVID-19 pandemic. This pandemic is transforming users' habits to access the Internet [2], [3] and media content consumption [4], [5]. The Broadband Commission for Sustainable Development, a joint initiative of the International Telecommunication Union (ITU) and the United Nations Educational, Scientific and Cultural Organization (UNESCO), is also concerned about these user habit changes and it is implementing an Agenda for Action to push an emergency response to the pandemic, aiming at Internet access extension and boosting its capacity [6], [7].

All the above-mentioned factors are inevitably influencing the evolution of all services, and especially affect the rich media ones. It is therefore evident that there is a need for new network-related solutions to support high quality of service (QoS) for these applications. The current network traffic crosses networks working on a best-effort basis where no details regarding packet delivery (e.g. time) is guaranteed. Therefore, best-effort networked-transmitted media traffic may result in lower user quality of experience (QoE). A pragmatic example of this QoE degradation are stalls or artifacts during media playback on player devices. Employing content delivery networks (CDNs) is the most common solution to prevent negative quality effects and make video delivery more efficient. CDNs are geographically distributed hierarchical systems that cache and store video streams to foster efficiency and increase the service coverage. CDN price is decreasing, but the overall cost for the content provider is increasing, as the traffic from/to CDN is increasing [8].

Beyond CDNs, more advanced solutions based on Network Function Virtualization (NFV) technologies [9] are being investigated to support media streaming services. NFV allows the deployment of Virtual Network Functions (VNF) devoted to empower network abilities when delivering media streaming traffic in an optimized and cost-effective manner [10], [11]. VNFs enable flexible operations whose benefits are threefold. First, VNF-based networks monitor objective operational parameters, such as throughput or latency, representative for QoS of the media streaming dataflows, which have a direct

R. Viola is with the Vicomtech Foundation, Basque Research and Technology Alliance, 20009 San Sebastián, Spain, and also with the Department of Communications Engineering, University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain (e-mail: rviola@vicomtech.org).

A. Martin and M. Zorrilla are with the Vicomtech Foundation, Basque Research and Technology Alliance, 20009 San Sebastián, Spain.

J. Montalbán is with the Department of Electronic Technology, University of the Basque Country (UPV/EHU), 20018 San Sebastián, Spain.

P. Angueira is with the Department of Communications Engineering, University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain.

G.-M. Muntean is with the Performance Engineering Laboratory, Dublin City University (DCU), Dublin, Ireland (e-mail: gabriel.muntean@dcu.ie).

Manuscript received May 1, 2021.

influence on user satisfaction. However, QoS metrics do not perfectly map on user experience, as user perceived quality is highly subjective. Additionally, QoE which compiles subjective evaluation elements, including rewards for playback quality and smoothness, and penalties for image freezes and unstable or low quality [12], [13], needs to be considered, too. Secondly, the Content Provider (CP) has more control to shape the network traffic and allocate resources since business rules for VNF deployment and life-cycle management could be established. These rules allow balancing network resources and business costs trade-offs [14], so they are highly relevant. Last, as the volume, complexity and real-time nature of the media streaming traffic has an evident impact on energy consumption of the network and devices managing the content, an optimized streaming delivery through VNFs should also consider the energy efficiency.

In this context, the main contributions of this survey are:

- The survey discusses widely employed performance assessment solutions and metrics related to media streaming. Metrics are classified into four subgroups: QoS, QoE and fairness, business metrics and energy efficiency;
- The survey provides an extensive overview of the literature involving media traffic monitoring and analysis, including traffic characterization and analysis to enable forecasts. Most common tools for network performance monitoring and simulation are also presented;
- The survey analyzes and classifies the state-of-the-art performance-driven network functions for media streaming;
- The survey presents technologies considered by the telecommunications industry as key enablers for the next generation networks, and discusses remaining challenges.

A list of acronyms used throughout the paper is presented in Table I. The rest of the paper is structured as follows. First, section II presents the objective of this work in the context of related surveys. Section III contains an overview of media streaming technologies and protocols, while section IV describes the taxonomy of VNFs for media streaming. Section V covers methods for the assessment of performance metrics related to media streaming. In section VI, we provide an overview of the state-of-art on media streaming network traffic monitoring and analysis. Section VII describes the network functions employed to date to enhance the performance of media streaming services, while section VIII presents the current challenges in the virtualization process of network functions, inside 5G networks and beyond, to assess the open issues and scientific research directions. Finally, we highlight some very valuable international initiatives in section IX and assert our conclusions in section X.

II. Paper Objectives in the Context of Related Surveys

The objective of this survey is to perform an extensive literature review on the proposed solutions in the realm of VNFs applied to the field of media streaming. The paper also addresses future challenges in this research area. To better understand how VNF solutions fit with media streaming, performance metrics and network traffic monitoring and analysis are necessary aspects to consider. Network analysis allows to design an effective network function by enabling the management and orchestration operations according to network status at any moment. Performance metrics are instead useful to test the effectiveness of the deployed network function to empower the media streaming service. Then, before presenting and comparing state-of-the-art VNF solutions, the paper also covers two more related topics: performance assessment and network traffic monitoring and analysis.

In a media streaming context, performance assessment focuses on the evaluation of the system employed to stream the content. The performance assessment is done by employing metrics and collecting measurements and/or estimations of such metrics. These will involve quantifiable values to track and monitor a streaming session, i.e., video resolution and bitrate, or a related factor which can influence it, i.e., network bandwidth and latency. Metrics are usually collected alongside the streaming session for two main reasons:

- They can be exploited as source of information for the network function, i.e., bandwidth measurement;
- They can be used as a measure of the goodness of a proposed network function. For example, it is possible to collect the video representation level in order to evaluate the effects of the network function on a streaming session.

Network traffic monitoring and analysis is the process of recording and monitoring traffic to gain the required knowledge to back decisions for increasing the network and service delivery performance as well as for executing network operation and management. This is a relevant field of research since the raise of digital telecommunications networks. Network traffic monitoring and analysis has different subareas: traffic characterization or modelling and traffic analysis to allow forecasts.

Traffic characterization or modelling consists in statistical analysis of the traffic in order to create a model which approximately describes the behavior of the network. There is not a universal model which perfectly describes the network, but each one brings its own limitations and is more accurate under specific conditions (network traffic profiles, employed protocols, transmission medium, etc.). Such models are the basis for generating realistic network traffic. Traffic generation attempts to exploit traffic models and provide tools to simulate specific network conditions. The purpose of synthetically replicating real conditions is to test network applications in a known and controlled environment before they come into play in a real deployment.

Finally, traffic analysis allows to predict network events. Here, it is important to study the temporal variability of the network and construct time series models which estimate future traffic patterns based on past observations of the network. The advantage to forecast events is clear, as it facilitates the implementation of proactive actions that prevent from network malfunctions.

Over the years, several surveys focused on performance assessment, network traffic monitoring and analysis or VNF-

TABLE I
List of Acronyms used in the paper.

3GPP	3rd Generation Partnership Project	Multi-RAT	Multiple Radio Access Technology
5G	Fifth Generation	NAT	Network Address Translation
6G	Sixth Generation	NFV	Network function virtualization
AES	Advanced Encryption Standard	NFV-RA	NFV resource allocation
AES-CBC	AES block cipher mode	NFVI	NFV Infrastructure
AES-CTR	AES counter mode	NFVO	NFV Orchestrator
ANN	Artificial Neural Network	NS	Network Service
AR	Augmented Reality	O-RAN	Open RAN
C-RAN	Cloud-RAN	ONAP	Open Network Automation Platform
CAPEX	Capital Expenditure	OPEX	Operational Expenditure
CDN	Content Delivery Network	OSI	Open Systems Interconnection
CSI	Channel State Information	OSM	Open Source MANO
CMAF	Common Media Application Format	OTT	Over-the-top
CN	Core Network	P2P	Peer-to-peer
COTS	Commercial off-the-shelf	PoP	Point of presence
CP	Content Provider	QoE	Quality of Experience
CRM	Customer Relationship Management	QoS	Quality of Service
DASH	Dynamic Adaptive Streaming over HTTP	RAN	Radio Access Network
DNS	Domain Name System	RNI	Radio Network Information
DTN	Delay-tolerant Networking	RNIS	RNI service
ESN	Echo State Network	RNN	Recurrent Neural Network
ETSI	European Telecommunications Standards Institute	RTCP	Real-time Transport Control Protocol
FeMBMS	Further enhanced MBMS	RTMP	Real-time Messaging Protocol
GUI	Graphical User Interface	RTP	Real-time Transport Protocol
HAS	HTTP Adaptive Streaming	RTSP	Real Time Streaming Protocol
HLS	HTTP Live Streaming	SCTP	Stream Control Transmission Protocol
HTTP	HyperText Transfer Protocol	SDN	Software-defined network
IaaS	Infrastructure as a Service	SDR	Software-defined radio
IBN	Intent-Based Network	SLA	Service Level Agreement
IoT	Internet of Things	SON	Self-Organizing Network
IIoT	Industrial Internet of Things	SRT	Secure Reliable Transport
IM	Instant Messaging	STUN	Session Traversal Utilities for NAT
IP	Internet Protocol	SVA	Streaming Video Alliance
ISP	Internet Service Provider	SVM	Support Vector Machine
ITU	International Telecommunication Union	SVR	Support Vector Regression
KPI	Key Performance Indicator	TCP	Transmission Control Protocol
L1	Physical layer	TURN	Traversal Using Relays around NAT
L2	Data link layer	UAV	Unmanned Aerial Vehicle
L3	Network layer	UDP	User Datagram Protocol
L4	Transport layer	UE	User Equipment
L7	Application layer	UNESCO	United Nations Educational, Scientific and Cultural Organization
LL CMAF	Low Latency CMAF		
LL-DASH	Low Latency DASH	VIM	Virtual Infrastructure Manager
LL-HLS	Low Latency HLS	VNF	Virtual Network Function
LSTM	Long short-term memory	VNF-CC	VNF Chain Composition
LTE	Long-Term Evolution	VNF-FG	VNF Forwarding Graph
M3U8	HLS playlist	VNF-FGE	VNF Forwarding Graph Embedding
MANO	Management and Orchestration	VNF-SCH	VNF Scheduling
MBMS	Multimedia Broadcast/Multicast Service	VNFI	VNF Instance
MEC	Multi-access Edge Computing	VNFM	VNF Manager
MLP	Multi-layer Perceptron	VOD	Video-on-Demand
MOS	Mean Opinion Score	VR	Virtual Reality
MPD	Media Presentation Description	vRAN	Virtual RAN
MPTCP	Multipath TCP	WebRTC	Web Real-Time Communication
MMS	Multimedia Messaging Service	WSN	Wireless Sensor Network

based solutions, but their scope was limited and did not discuss on the relation between these topics. Table II shows a summary of related survey papers.

Surveys related to performance assessment usually focus on a specific point of view, e.g., user's QoE or network QoS, meaning that they do not cover all the possible performance metrics. Chalmers et al. [16] provide a literature review of QoS assessment for mobile environment. QoS metrics are grouped into two categories: Technology-Based QoS and User-Based QoS. Jin et al. [17] focus their review on media applications and prefers to group QoS metrics depending on the layers of the end-to-end architecture they belong: Resource layer, Application layer and User layer QoS metrics. Akhtar et al. [27] provide the same classification, but also include QoE performance assessment. Even though QoE assessment should be based on subjective evaluation, a review of objective methods to assess QoE is presented. Objective methods infer QoS performance metrics to estimate subjective scores and are widely employed since they allow to reduce time and resources costs for the execution of the subjective evaluation. QoS and QoE are highly correlated each other, as it is presented by Alreshoodi et al. [13].

Surveys which review QoE assessment methods are wider presented in the literature compared to surveys dealing with

TABLE II
SUMMARY OF PREVIOUS SURVEYS ON VIRTUAL NETWORK FUNCTIONS AND MEDIA STREAMING.

Survey	Scope and topics	Performance assessment	Network traffic	Network virtualization	Year
Adas et al. [15]	Traffic models in broadband networks	-	Modelling	-	1997
Chalmers et al. [16]	QoS in mobile environment	QoS	-	-	1999
Jin et al. [17]	QoS specification for media applications	QoS	-	-	2004
Feng et al. [18]	Network traffic predictors	-	Analysis	-	2005
Chandrasekaran et al. [19]	Network traffic models	-	Modelling	-	2009
Mohammed et al. [20]	Network traffic models	-	Modelling	-	2011
Hoque et al. [21]	Energy efficient media streaming, wireless networks	Energy	Modelling, analysis	-	2012
Alreshoodi et al. [13]	QoS and QoE correlation models	QoS, QoE	-	-	2013
Baraković et al. [22]	QoE assessment over wireless networks, QoE optimization	QoE	-	-	2013
Seufert et al. [23]	QoE assessment for HAS, HAS adaptation strategies	QoE	-	-	2014
Juluri et al. [24]	QoE assessment in VOD services	QoE	-	-	2015
Su et al. [25]	Wireless and mobile networks, video coding, QoE assessment for mobile media streaming	QoE	-	-	2016
Zhao et al. [26]	QoE assessment and management in video streaming	QoE	-	-	2016
Akhtar et al. [27]	QoS and QoE assessment for audio-visual content	QoS, QoE	-	-	2017
Petrangeli et al. [28]	QoE assessment for HAS, QoE-centric management of HAS	QoE	-	-	2018
Skorin-Kapov et al. [29]	QoE assessment for HAS, QoE-centric management of HAS	QoE	-	SDN/NFV, MEC	2018
Barakabitze et al. [30]	QoE assessment, QoE management in SDN/NFV	QoE	-	SDN/NFV, MEC, Cloud/Fog	2019
Barman et al. [31]	QoE assessment and modelling for HAS	QoE	-	-	2019
Zhang et al. [32]	VNF design considerations	-	-	VNF, Cloud/Edge	2019
Navarro-Ortiz et al. [33]	5G Use cases and Traffic Models	-	Modelling	-	2020

QoS. Baraković et al. [22] present the state-of-the-art QoE assessment while using wireless networks, but they do not limit to media streaming services. Su et al. [25] also focus on wireless networks, but they limit to media streaming services. The authors also include reviews of wireless network technologies and video encoding as related topics. Seufert et al. [23], Petrangeli et al. [28] and Barman et al. [31] propose comprehensive surveys on QoE assessment while employing HTTP Adaptive Streaming (HAS) technologies to stream the media content. Performance metrics specific for HAS, referred as influence factors by the former, can directly influence the QoE evaluation. They also provide a review of server and client-side solutions to improve the QoE scores by optimizing the adaptation strategy. QoE assessment by Juluri et al. [24] includes instead a review of both real-time streaming and HAS metrics, but then focuses only on describing methods to assess QoE for Video on Demand (VOD) applications. Zhao et al. [26] include a similar review on state-of-the-art QoE assessment. Both Juluri et al. [24] and Zhao et al. [26] presents and classify objective QoE influence factors. Skorin-Kapov et al. [29] and Barakabitze et al. [30] are the most recent surveys on QoE assessment and describe also SDN/NFV-based approaches to enhance the streaming services.

Concerning energy efficiency performance, Hoque et al. [21] compare energy-related metrics and approaches to improve the efficiency of the streaming service. Solutions for energy efficiency are classified depending on the layer of the Open Systems Interconnection (OSI) model they belong. The authors also provide some considerations on traffic modelling and analysis, as they are enablers for reducing energy consumption.

Differently from the above-mentioned works, our survey covers performance metrics applied to media streaming from all the three different domains separately discussed in previous works: QoS, QoE and energy efficiency. Moreover, we add a review of performance metrics related to business aspect of media streaming services.

Surveys on network traffic do not cover media streaming use cases, but they remain generic, as they do not consider an application-specific traffic. Adas et al. [15] was the first survey on network traffic monitoring, but it was based on research studies on traffic generated before the 2000s. Some of the models are still considered valid, but some new models have been proposed, as described by more recent surveys, such as Chandrasekaran et al. [19] and Mohammed et al. [20]. Navarro-Ortiz et al. [33] is the most recent one and also considers the effects of specific 5G use cases/applications on traffic modelling. Feng et al. [18] lists different approaches for network traffic analysis. Our survey addresses only media streaming use case. Then, we include a review on media streaming-related traffic modelling and analysis.

Several surveys on network virtualization have been published in the last few years, in line with the increased interest in virtualization. Zhang et al. [32] provides generic considerations when designing VNFs. Limited to media streaming scope, as already mentioned before, Skorin-Kapov et al. [29] and Barakabitze et al. [30] discuss the use of virtualized solutions to improve the QoE assessment. Our survey wants to provide a similar review on VNFs, but we want to add other performance metrics and emphasize on media streaming traffic.

Therefore, our survey addresses a more specific scope. From one side, we want to widely present performance assessment, including less discussed metrics in literature, such as energy and business-related metrics. On the other side, we provide a review of network characterization due to media streaming traffic and present network solutions for its optimization. This survey discusses the relation between the VNFs and media streaming, also considering performance assessment and network traffic monitoring and analysis.

III. Media Streaming Overview

Media streaming refers to the delivery of media content (e.g., live television, video clip, etc.) from a streaming server to a streaming client over a certain network infrastructure. The media source can be either live or pre-recorded. In some cases, the Content Provider (CP) is also the owner of the infrastructure employed to stream the content, but recently diverse providers and operators have entered the market successfully with different roles in the media streaming process, e.g., Akamai, Netflix, etc. Some of them have their own proprietary media streaming solutions. However, first solutions were based on the Real-time Transport Protocol (RTP) [34] on top of the User Datagram Protocol (UDP) [35], where the Real-time Transport Control Protocol (RTCP) [34] was employed to monitor network metrics and update the rate control. The choice of UDP was based on its lower latency when compared to the Transmission Control Protocol (TCP) [36], even if it does not guarantee reliability when delivering packets, i.e., lost packets are not re-transmitted when employing UDP. The later explosion of Over-the-top (OTT) services, e.g., Netflix and Hulu, pushed the search for new solutions to deliver Video-on-Demand (VOD) contents, where latency was not a concern, but scalability to cover the increasing user demand for content. In OTT services, the CP streams its content over a public network and an Internet service provider (ISP) is in charge of the actual content delivery. HTTP adaptive streaming (HAS) [23] technologies were introduced to deliver OTT content and the use of TCP and HTTP made them attractive since these protocols are ubiquitous. Additionally, almost every device or User Equipment (UE) can establish HTTP-based communications. The HAS-based design has the following advantages over RTP/UDP-based solutions:

- Traverse networks: HAS communications are performed on top of HTTP/TCP stack and uses port 80 and pull-based streaming protocols. These cross current network infrastructure components, such as Network Address Translation (NAT) and firewall devices [37];
- Reuse and scalability: HAS-based media services can reuse existing CDN systems and caching infrastructures without modifications reaching wide audiences;
- User mobility and device heterogeneity: The dynamic content adaptation-enabled player mechanism is accommodated by all latest heterogeneous UEs, i.e., smartphones, tablets, which support user mobility.

Figure 1 illustrates the HAS-based adaptive streaming principle. HAS works in pull mode which means that the client pulls the data from a standard HTTP server, which simply hosts the media content. To reduce the effect of network fluctuations on the playback, HAS employs a dynamic content adaptation to provide a seamless streaming experience. The original media content is encoded at multiple representations, which differ from each other in terms of bitrate and/or resolution and are split into segments of fixed time duration (i.e., a segment is usually between 2 and 10 seconds). A manifest file is also generated and stored at the server, which contains information of the available representations including HTTP URLs indicating where to download the segments of each representation. During a typical HAS session, the client constantly measures certain parameters, such as available network bandwidth and playback buffer level. When it requests content, the client first receives the manifest file which is examined. Then, following an internal adaptation algorithm that processes the monitored performance parameters' values and takes decisions according to the desired adaptation policy, the client requests to download from the server the segment of an appropriate representation.

Apart from transport-layer protocols such as RTP, there are also application-layer protocols that are employed in media streaming. Table III summarizes different aspects of interest when identifying the best protocol candidates to be used when streaming videos. RTP and Real Time Streaming Protocol (RTSP) [38] perform low latency communications compatible with multicast media streaming. TCP-based Real-time Messaging Protocol (RTMP) [39] enables higher reliability compared to RTSP, but at the cost of having higher latency. Secure Reliable Transport (SRT) [40] simplifies the delivery by enabling both push and pull modes of operation. Web Real-Time Communication (WebRTC) [41] enables media streaming through a web browser by exploiting Session Traversal Utilities for NAT (STUN) [42] and Traversal Using Relays around NAT (TURN) [43] protocols provided by third party servers. Both SRT and WebRTC increase the security by including mandatory encryption support, while this is not always required for RTMP. HTTP Live Streaming (HLS) [44] and Dynamic Adaptive Streaming over HTTP (DASH) [45] increase latency due to an internal buffering to overcome network dynamics. In any case, violations on delivery timing could cause stalls and image freezes during the playback if the internal buffer gets empty. To minimise such issues, HAS allows dynamic adaptation mechanisms to track the variability of the network and select appropriate bitrate. Thus, sudden networking problems are prevented by an alternative bitrate selection from the manifest. Common Media Application Format (CMAF) [46] was a proposal to merge major streaming formats around HLS and DASH. Moreover, its Low Latency mode (LL CMAF) aims to reduce the latency by enabling HTTP chunked/push mode. Thus, the latency can be reduced and get closer to UDP-based streaming technologies. In practice, CMAF did not achieve to integration of HLS and DASH streaming formats since the implementations of Low Latency HLS (LL-HLS) [47] and Low Latency DASH (LL-DASH) [48] still present some differences. Thus, LL-HLS

TABLE III
FEATURES OF STREAMING TECHNOLOGIES.

Tech.	Transport	Manifest file	Common issues	Latency	Available bitrate	Bitrate adaptation	CDN compatible	Encryption
RTP	UDP	no	packets lost & artifacts	very low ($\leq$1sec)	RTCP	encoder	no	no
RTSP	UDP	SDP	packets lost & artifacts	very low ($\leq$1sec)	RTCP	encoder	no	no
RTMP	TCP	no	packets lost & artifacts	low (1-3secs)	RTMP control messages	encoder	no	AES-128 CBC
SRT	UDP	no	packets lost & artifacts	very low ($\leq$1sec)	SRT control messages	encoder	no	AES-128 / 265 CTR
WebRTC	UDP, QUIC-ready	SDP	packets lost & artifacts	very low ($\leq$1sec)	RTCP	encoder	no	AES-128 CTR
HLS	HTTP 1.X / 2.0 over TCP	M3U8	segment buffering & quality switch	high (5-30secs)	representation	player	yes	AES-128 CBC
DASH	HTTP 1.X / 2.0 over TCP, QUIC-ready	MPD	segment buffering & quality switch	high (5-30secs)	representation	player	yes	AES-128 CBC / CTR
LL-HLS	HTTP 2.0 over TCP	M3U8	chunks buffering & quality switch	low (1-3secs)	representation	player	yes	AES-128 CBC
LL-DASH	HTTP 1.1 Chunked over TCP	MPD	chunks buffering & quality switch	low (1-3secs)	representation	player	yes	AES-128 CBC / CTR

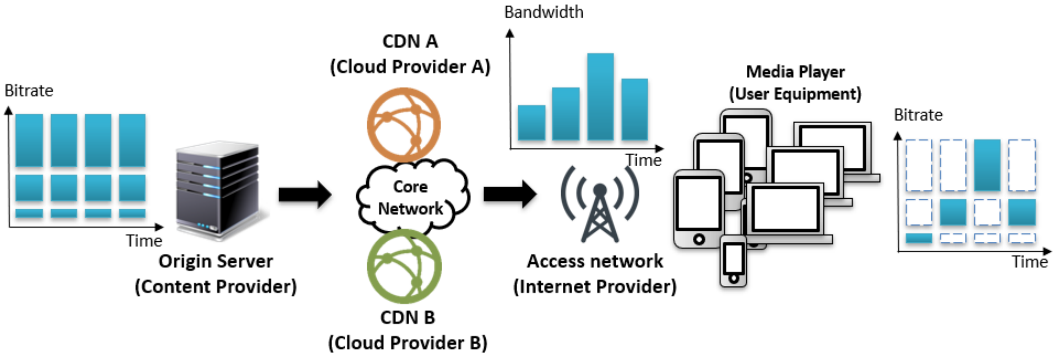


Fig. 1. HAS-based media streaming principle

and LL-DASH employ different approaches for HTTP transport and encryption schemes. For instance, a common feature to most HTTP-based solutions is the security by design where different encryption standards protect communications, such as Advanced Encryption Standard (AES) [49] with Cipher Block Chaining (AES-128 CBC) or Counter mode (AES-128 CTR).

Finally, even if most existing media streaming solutions employ UDP and/or TCP, some of them, such as DASH [50] and WebRTC [51], are already evolving and/or being tested with QUIC, a new transport protocol which is expected to substitute TCP when HTTP/3 will replace the current HTTP/2. QUIC lays on top of UDP to provide reduced latency, but with a connection control mechanism to guarantee the same reliability as TCP [52]. There are also proposals to use HAS-based media streaming with protocols such as Stream Control Transmission Protocol (SCTP) [53] and Multipath TCP (MPTCP) [54], which support multihoming, very important in recent heterogeneous network environments. Noteworthy is that MPTCP is backward compatible with the vanilla TCP, which is very useful for service deployment. Finally, efforts are already being made to develop a multipath QUIC [55] protocol to combine the benefits of these approaches, but so far no HAS-based media delivery solution has used it.

IV. TAXONOMY OF VIRTUAL NETWORK FUNCTIONS FOR MEDIA STREAMING

Following the discussion of media streaming solutions, this section reviews the motivations for the applicability of VNFs to improve the media streaming process. The taxonomy of VNFs applied to media streaming is shown in Figure 2.

Media streaming can leverage VNFs to enable higher network capacity and stability, media traffic optimization and other performance-related advantages. The final aim is to increase the performances of media streaming, including efficient use of network resources and end device capabilities [56] during their involvement in the streaming service. Media streaming performance indicators include Quality of Service (QoS), Quality of Experience (QoE) and Fairness, Business metrics and Energy Efficiency and illustrated in Figure 3. We will discuss media streaming performance assessment from these different perspectives in section V.

The employment of VNFs in media streaming is growing in the last few years, as the attention increases on media distribution over the newly deployed 5G networks [57]. VNFs are intrinsically designed to follow the principles of modularity, interoperability, scalability and flexibility. However, to use VNFs more effectively in media streaming, knowledge of the network is essential. Characterizing and

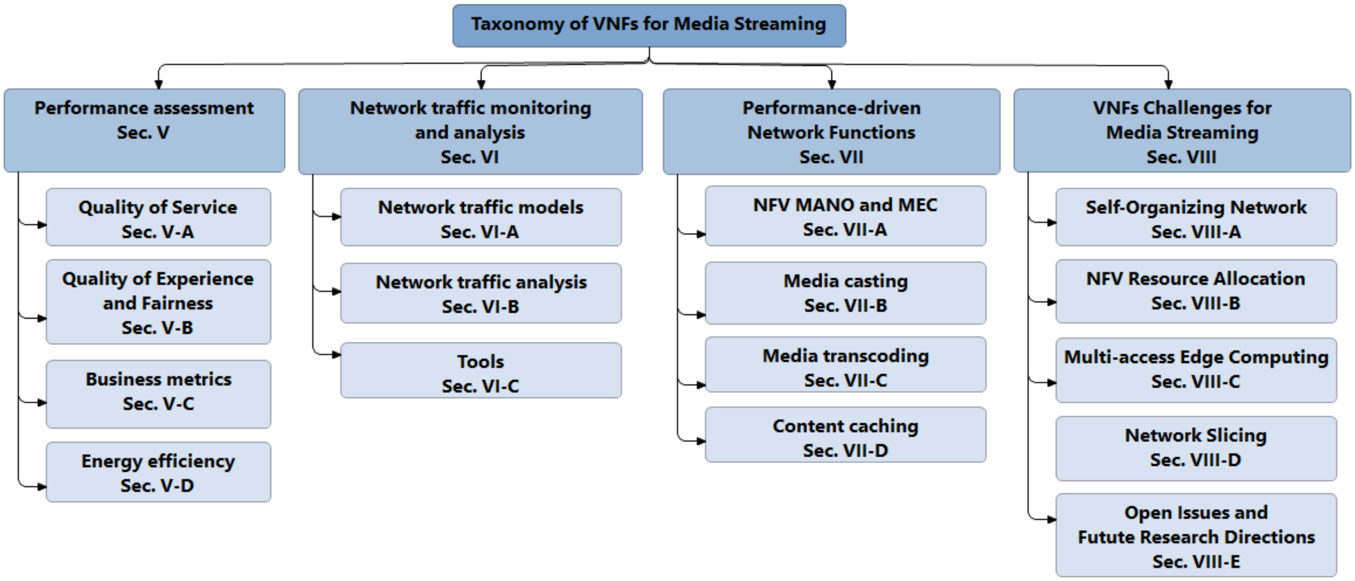

Taxonomy of VNFs for Media Streaming
Performance assessment Sec. V
Quality of Service Sec. V-A
Quality of Experience and Fairness Sec. V-B
Business metrics Sec. V-C
Energy efficiency Sec. V-D
Network traffic monitoring and analysis Sec. VI
Network traffic models Sec. VI-A
Network traffic analysis Sec. VI-B
Tools Sec. VI-C
Performance-driven Network Functions Sec. VII
NFV MANO and MEC Sec. VII-A
Media casting Sec. VII-B
Media transcoding Sec. VII-C
Content caching Sec. VII-D
VNFs Challenges for Media Streaming Sec. VIII
Self-Organizing Network Sec. VIII-A
NFV Resource Allocation Sec. VIII-B
Multi-access Edge Computing Sec. VIII-C
Network Slicing Sec. VIII-D
Open Issues and Futute Research Directions Sec. VIII-E

Fig. 2. Taxonomy of VNFs for media streaming.

modelling network behavior, as well as monitoring and analyzing its traffic provide useful information to be exploited while designing, deploying a VNF [10], [11] and managing its life-cycle [58]. Study of networks and traffic can be tackled from different points of view, as shown in Figure 4. There is a wide agreement that real world knowledge allows to design a more mature VNF [59]. This knowledge is collected from network monitoring and data analysis. Considerations on network traffic monitoring and analysis are included in section VI.

Based on the achievements in performance assessment and knowledge acquired in network and traffic characterization, several network solutions to enable a performance-driven management of the resources are already being employed and/or investigated, as shown in Figure 5. Section VII deals with performance-driven network functions, including a review of the solutions provided in literature.

Finally, in the current deployment of 5G networks the VNFs have a significant role, as 5G aims to having a fully virtualized network deployment. However, there are still several open issues and challenges that need to be address in the future, as shown in Figure 10. Section VIII discusses the future of VNFs in order to enable an improved media streaming process and enhanced user experience.

V. Performance Assessment

This section presents an overview of performance assessment avenues in the context of VNF-based media streaming. It involves performance aspects from multiple viewpoints, including QoS, QoE and fairness, business metrics and energy efficiency, as illustrated in Figure 3.

A. Quality of Service (QoS)

QoS is related to features which describe the status of network communications and/or the service supported by the network.

QoS properties should be physical and measurable. They are objective performance factors not affected by user's perception of the application, but they will definitely influence this perception. Due to the heterogeneity of networks and/or applications, there is not a unique set of widely accepted QoS properties. Different textbooks and publications introduce differently QoS, but in most cases they focus on a specific network and/or application context. When dealing with a network, QoS assessment consists of measuring network performance, e.g., *network bandwidth*, *packet loss* and *latency*. When considering a particular network application, QoS is linked to a wide range of properties including performance, responsiveness, availability, reliability, and application-related aspects. Each QoS property will be associated to a performance metric. Each metric will facilitate monitoring and characterization of the application property in order to understand the application behavior from that perspective. Furthermore, having a characterization of the application based on metrics associated with different QoS properties allows to put in place actions to optimize the operations of the application at run-time.

Focusing on media streaming and based on the proposals made in [17] and [27], QoS properties are classified depending on the level of abstraction from the underlying network and hardware/software capabilities. Three main groups of QoS properties (i.e. introducing a QoS layer structure) are defined in relation to different concerns: resources, applications and users. Typical QoS properties and parameters/metrics to measure each property for each these three groups are summarized in Table IV and are discussed next.

The performance metrics at resource QoS layer quantify physical resource properties and are highly dependent on the hardware and platform employed. At this QoS layer, the most interesting ones for media streaming are *timeliness*, *capacity* and *reliability*. Properties at resource layer should fit the requirements needed by the streaming service exploiting

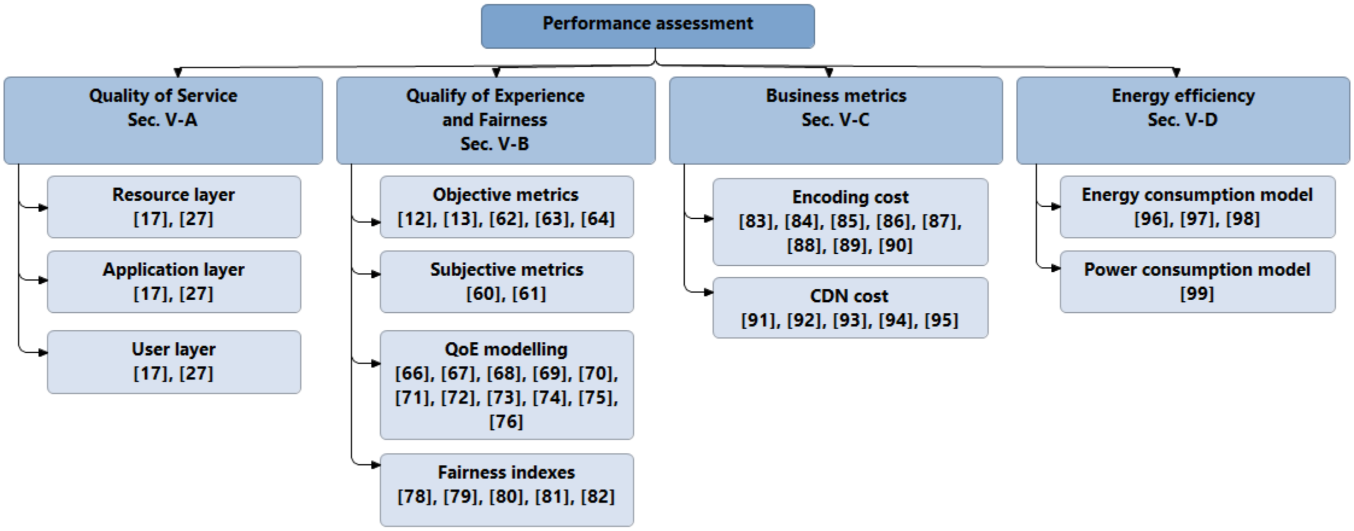


Fig. 3. Multi-dimensional performance assessment.

those resources. Metrics are not dependent on any particular application and/or user, rather than on the service requirements e.g., video conferencing has tight packet delay and jitter requirements, while video on demand (VOD) requires high throughput. Consequently, the performance metrics are generic for a range of applications and are measured on different OSI network model layers ranging from layer 1 (physical layer) to layer 4 (transport layer). Performance metrics at different OSI layers can characterize the same QoS property, but the abstraction from the physical resource becomes higher as the layer level increases.

At application QoS layer, similar properties to resource layer can be identified, but the metrics are now completely independent from the hardware and platform, as they are application-specific and can be mapped on the network application layer (OSI layer 7). For instance the performance metrics for media streaming at application layer are highly dependent on the video and audio encoding/decoding, streaming technology for the delivery and any other application-level media processing. These metrics are completely abstracted from the network protocols, meaning that they could remain valid even if the content is a file stored locally, as the application layer is agnostic about the origin of the content (local repository or remote server). The content production itself already provides QoS parameters, via audio and video codecs and their encoding bitrates. These parameters are fixed in the encoding and/or decoding process, and they do not depend on the underlying layers. These QoS attributes refer to the characteristics of the encoding at media server and multimedia capabilities at the player device. These content and device characteristics have fixed values during the streaming session and are easily known. In some situations, content characteristics are not enough to describe the QoS at application layer. Additional characteristics dependent on measurements at player side while presenting the content are used. Video impairments provide an objective measurement of the QoS level of the streaming.

Although application layer QoS properties provide high level objective metrics, they are not enough to describe user's point of view. User can be influenced by both objective and subjective factors. User layer QoS aims to identify objective metrics which describe the streaming service from user's point of view. The fact that they remain objective means that they can be measured. Here, these objective properties are completely different since they are neither based on physical resources, nor technological assets, while they deal with external features. We consider that there are two main categories to classify user layer performance metrics: technical metrics and economic ones. Technical metrics describe user device and streaming content. Economic metrics includes consideration on streaming service pricing.

Finally, user subjective metrics are instead not uniquely quantifiable as the user QoS layer ones. They are also referred to as QoE since they are focused on how the user perceives the media service. A more detailed explanation on QoE assessment is described in the next section.

B. Quality of Experience and Fairness

1) Quality of Experience (QoE): QoS performance metrics do not express well users' perceived quality and satisfaction with services. A major reason relies on the fact that human evaluation is influenced by subjective factors that cannot be easily defined by quantifiable parameters and then measured. Therefore, the term QoE is employed to define and describe how a user perceives the media streaming service. Having good values for QoS metrics is not enough to guarantee a certain level of QoE, as it does necessarily imply that the perceived quality is also good. QoS performance metrics can be considered as QoE objective metrics, but additional QoE subjective metrics are necessary. Table V shows widely used QoE subjective metrics.

The International Telecommunication Union (ITU) defines the Mean Opinion Score (MOS) [60] as the measure for the QoE evaluation. It is a widely consolidated way to evaluate the QoE and consists in five quality increasing levels: 1-Bad, 2-Poor, 3-Fair, 4-Good, 5-Excellent. MOS levels are shown in Table VI. MOS level achieved by a particular streaming service is assessed by arithmetic mean over all the individual

TABLE IV
QoS PERFORMANCE METRICS.

QoS Layer	Property / Category	Parameter / Metric	Description
Resource layer	Timeliness	Packet delay	Time taken to deliver a packet
		Packet jitter	Delay inconsistency between each packet
	Capacity	Channel bandwidth	Occupied frequency range
		SNR	Signal-to-Noise Ratio
		PSNR	Peak Signal-to-Noise Ratio
		MCS	Modulation and Coding Scheme
		Network bandwidth	Maximum (theoretical) data transfer rate
		Throughput	Effective data transfer rate
	Reliability	BER	Bit Error Rate
		PLR	Packet Loss Rate
		Outage probability	Probability that data transfer rate is less than the required threshold
Application layer	Timeliness	Startup delay	Time to receive and display the first video frame
		End-to-End Delay	Time elapsed from content production to its consumption
		Queuing Delay	Time the video frame waits in the playback queue before being displayed
		Audio&Video synchronization	Audio and video are synchronized (no lip sink error)
	Capacity	Audio bandwidth	Audio frequency range
		Audio sampling rate	Audio samples recorded every second
		Video resolution	Pixels in each dimension that can be displayed
		Video frame rate	Video frames recorded every second
		Audio&Video codecs	Codecs employed for audio and video encoding
		Audio&Video encoding bitrates	Bitrates employed for audio and video encoding with the given codecs
		Audio&Video representations	The representation levels presented in HAS
	Reliability	Video Frame Loss	Video frames lost while displaying
		Representation switches	Switches between audio and/or video representation levels
		Stalling ratio	Probability of stalling events
		Stalling duration	Duration of stalling events
User layer	Technical	Device type	Smartphone, tablet, TV, etc.
		Screen/window size	Size of output screen/window
		Content type	Video conferencing (real-time), Live Streaming, Video on Demand, etc.
	Economic	Pricing model	Flat-Rate or Pay-per-Use pricing
		Range of price	High, medium or low price

TABLE V
QoE SUBJECTIVE METRICS.

Category	Examples	Description
Contextual factors	Location	Home, office, car , etc.
	Environmental characteristics	Noisy or quite, crowded or uncrowded, etc.
	Motion	Sitting or moving, speed, etc.
	Time	Time of the day
Human factors	Age	User's age
	Mood	Emotional state at any time
	Attention level	Attention level at any time
	Goal	User's aim
	Motivation	Level of motivation

TABLE VI
MEAN OPINION SCORE LEVELS.

MOS	Quality	Impairments
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

ratings by subjects which take part in the evaluation test. Nevertheless, due to the unpredictability of the subjective factors, a considerable number of scenarios could be possible while assessing the QoE. Then, ITU addresses this issue by attempting to standardize the scenario and environmental variables where the QoE ratings are collected. ITU describes the procedures to assess MOS in the correct way [61]. The procedure intrinsically entails a long time since it requires to select a diverse group of people to represent a good approximation of a typical human audience for a given content. Then, the content should be shown to all the subjects of the chosen set and rated by them.

To simplify QoE assessment, the correlation between QoS and QoE is widely investigated in literature [12], [13] to profile the subjective human perception of the quality. Consequently, quality assessment based on Peak Signal-to-Noise Ratio (PSNR) when considering comparatively the viewed video frames and the original ones has been replaced by more accurate metrics, such as Structural similarity (SSIM) [62], SSIMplus [63], and Netflix' Video Multi-Method Assessment Fusion (VMAF) [64]. While PSNR and SSIM are limited to spatial analysis of video frames, SSIMplus and VMAF include both spatial and temporal analysis. VMAF also moves from employing statistical analysis methods to machine-learning algorithms. VMAF evaluates several elementary metrics which measure content characteristics, type of artifacts and degree of distortion and uses inference to deliver a more accurate final score. Furthermore, Netflix introduced the concept of Per-Title Encoding [65], the same metrics employed to evaluate the user's QoE can be exploited at the server-side while encoding the content. Per-Title encoding allows to select the encoding bitrate which maximizes the user's QoE depending on the type of the media content (i.e. news, sport, action movie, etc.).

TABLE VII
QoE MODELS FOR HAS.

Model	Description	MOS scale	Year
De Vriendt et al. [66]	Bitrate model, PSNR/SSIM model, chunk-MOS model and Quality level model	yes	2013
Yin et al. [67]	Normalized QoE	no	2014
Xue et al. [68]	Instantaneous and cumulative QoE with exponential decay	no	2014
DASH-UE (Liu et al.) [69]	DASH User Experience model	no	2015
Bentaleb et al. [70]	SSIMplus-based QoE	yes	2016
SQI (Duanmu et al.) [71]	Streaming QoE Index	yes	2016
U-vMOS (Huawei) [72]	User/Unified/Ubiquitous video Mean Opinion Score	yes	2016
ITU-T P.1203 [73]	Parametric bitstream-based quality assessment for HAS services	yes	2017
KSQI (Duanmu et al.) [74]	Knowledge-driven streaming quality index	yes	2019
De Fez et al. [75]	Modified Yin[67]-model, PSNR-based model, VMAF-based model	yes/no	2020
ITU-T P.1204 [76]	Bitstream-based/pixel-based/hybrid models for resolutions up to 4K	yes	2020

Focusing on HAS, diverse metrics have been considered to create QoE models based on its characteristics. QoE models span from pixel-level comparison between received frames and original ones (e.g. PSNR, SSIM, SSIMplus and VMAF) to content-agnostic models with sophisticated equations which consider a wide range of parameters, including available representation bitrates, frequency of bitrate changes and buffering duration. By focusing on objective metrics only, there is an inevitable loss of accuracy, but it has several practical advantages. The absence of human feedback on the QoE reduces the test time and result processing can be automated to be carried out online. Common QoE models are presented in Table VII. Even ITU defines MOS as the standard metric, MOS scale is not employed by all the QoE models as a measure of achieved QoE ratings.

ITU proposes several models, including ITU-T P.1203 [73] and ITU-T P.1204 [76]. ITU-T P.1203 [73] is a parametric bitstream-based model for HAS services which expresses the result in terms of MOS. The model considers both audio and video features, the impact of buffering on perceived quality and also takes into account information on the employed display device. Due to the performed bitstream analysis, a real implementation [77] of this model is computationally intensive as content is analyzed on a per media segment and per video frame basis. The model introduces four modes of operation, from 0 to 3, to tune the trade-off between accuracy and complexity. Lower modes are less accurate to reduce complexity, while higher mode increase the complexity to gain accuracy. Finally, modes 3 and 4 also raise security issues, as the bitstream must be unencrypted/decrypted to access the required input information. ITU-T P.1204 [76] the newest standard from ITU and it is not actually a unique model, but it groups models of different type: bitstream-based, pixel-based and hybrid models. ITU-T P.1204 is meant to be an extension to ITU-T P.1203, as it is focused on higher resolutions (up to 4K). Unfortunately, both ITU models show intrinsic computational complexity.

As an alternative to the complex ITU models, other QoE models are proposed in the literature. In equation (1) De Vriendt et al. [66] formulate a general expression for QoE models to predict the results of HAS services on MOS scale.

$$M_{pred} = \alpha * \mu - \beta * \sigma - \gamma * \phi + \delta \tag{1}$$

where α, β, γ δ are tunable coefficients. μ and σ are average of the quality of the displayed HAS representations and its standard deviation, respectively. Finally, ϕ takes into account both average duration and frequency of freeze events. From the equation, it is clear that the quality estimation is influenced by some major factors of HAS services: quality associated to each representation, switches between representations and stalling/buffering events. The coefficients are tuned by minimizing the Mean Square Error (MSE) between the predicted MOS values (M$_{pred}$) and the real ones assessed by rating a set of different video clips on different devices, as shown is equation (2).

$$\frac{\sum_{n=1}^{N} (M_{pred,n} - MOS_n)^2}{N} \tag{2}$$

The selection of values for μ and σ leads to different types of models, as several ways to define the quality of a representation are possible. De Vriendt et al. [66] state that there are at least 4 ways to select the quality:

- Bitrate model: the quality is defined by the bitrate of the representation.
- PSNR or SSIM model: the quality is defined by the average PSNR or SSIM over all the frames of a segment.
- Chunk-MOS model: the quality is not calculated from the representation, but it is part of the same MSE minimization process.
- Quality level model: the quality levels are equally spaced between a minimum and a maximum value.

The authors conclude that the chunk-MOS model has the best performance, with more flexibility for optimization since two more parameters (μ and σ) are varied to improve the model.

Yin et al. [67] suggest a similar approach that considers the same variables to assess a normalized QoE. Later models start from a similar optimization problem expressed by the equation (1) and aim to expand by including further variables or defining differently the quality associated to each representation. Liu et al.'s DASH-UE [69], SQI (Duanmu et al.) [71] and Huawei's User/Unified/Ubiquitous video MOS (UvMOS) [72] also include the startup (initial) delay in the equation. They assert that startup delay has negative effects on the user's QoE. Xue et al. [68] perform instantaneous QoE score estimations and introduce an exponential decay to emulate the forgetting curve of human perception when evaluating the cumulative QoE score.

Bentaleb et al. [70] employ SSIMplus [63] to assess the quality related to each representation instead of the four approaches proposed by De Vriendt et al. [66]. Duanmu et al.'s Knowledge-driven Streaming Quality Index (KSQI) [74] considers the same variables, aims to include a human visual system (HVS) analysis result to improve QoE modelling. The authors derive a system of linear inequalities from QoE subjective studies which allows to improve the optimization problem of QoE modelling.

Finally, De Fez et al. [75] propose three different models. The first one is a modified Yin et al. [67] model which improves accuracy by using the actual video segment bitrate instead of the average value of the segment representation. The second and the third ones are a PSNR-based model and a VMAF[64]-based model, respectively. These models employ PSNR and VMAF metrics to evaluate the quality of each segment instead of the actual video bitrate.

2) Fairness: While the number and diversity of approaches to assess QoE is large, there are very few metrics to measure fairness. Jain's fairness index [78] is one of the most widely used such metric and was originally introduced to express the fairness of throughput distribution across multiple flows that share a common distribution infrastructure. However, its applicability can be extended to any set of values x_i, which are measured on a scale, where $i = \overline{1, N}$. Note that the minimum Jain's fairness value is $\frac{1}{N}$ and the maximum is 1.

$$J(x_1, x_2, ...x_N) = \frac{(\sum_{i=1}^{N} x_i)^2}{\sum_{i=1}^{N} {x_i}^2} \tag{3}$$

Unfortunately, even though many networking parameters can be measured on ratio scales, there are some (i.e., QoE is among them), which are expressed on interval scales, such as the 5-point MOS scale, for instance. For one of these situations, Hoßfeld et al. [79] have proposed a QoE Fairness index based on the lowest L and highest H bounds of the rating scale. In (4) σ is the standard deviation and measures the degree of dispersion of the values. Hoßfeld's fairness index has values in the interval $[0, 1]$, where 0 is associated with total unfairness and 1 with perfect fairness.

$$F = 1 - \frac{2\sigma}{H - L} \tag{4}$$

There are some situations when classic fairness metrics do not reflect well the actual distribution of values. A more generic product-based fairness metric, presented in equation (5), was discussed in [80] along with other fairness metrics. In equation (5) f is a transformation function which can be defined according to the desired effect, allowing for very high flexibility in the fairness assessment.

$$\mathcal{P}(x) = \prod_{i=0}^{N} f\left(\frac{x_i}{\max(x)}\right) \tag{5}$$

The simplest product-based fairness index which uses a linear function $f(x) = x$ is represented in equation (6).

$$LP(x) = \frac{\prod_{i=0}^{N} x_i}{\max(x)^N} \tag{6}$$

TABLE VIII
BUSINESS COSTS FOR MEDIA STREAMING.

Category	Examples	Description
Capital expenditure	Buildings and furniture	Buy buildings, server racks, etc.
	Equipment	Servers, laptops, monitors, etc.
	Intangible assets	Purchased licenses or patents
	Software	Commercial proprietary software
Operational expenditure	Utilities	Electricity, water, etc.
	Employees	Salaries and benefits
	Research and development	Develop/improve media service
	Encoding	Costs due to content preparation
	CDN usage	Costs due to content delivery

Two other product-based fairness indexes, G's and Bossaer's, are defined using $f(x) = \sin(x\pi/2)^{\frac{1}{k}}$ and $f(x) = x^{\frac{1}{k}}$, respectively. While the first emphasizes the values closer to $\max(x)$, the later inflates the values closer to 0.

Other approaches include the general fairness model proposed by Lan et al. [81] and min-max and max-min-based fairness indexes introduced by Radunovic et al. [82].

C. Business Metrics

Achieving higher QoS and QoE values comes at a cost for the Content Provider (CP), because there are generally increased expenses in terms of network/services resources. Nevertheless, CP's strategies should focus on minimizing the business costs, while guaranteeing the same or even higher QoS/QoE. Table VIII provides a list of common business costs for a typical CP-based media streaming.

Capital Expenditure (CAPEX) refers to expenses incurred by the CP for the acquisition or improvement of fixed assets that are necessary for the business. CAPEX includes intangible assets, specific software as well as expenses related to licenses and patents payments. In this sense, the use of video codecs is the most evident example. Moving Picture Experts Group (MPEG) codecs require payments of licenses (royalties) for commercial use. H264 remains widely used, and it is still supported by most end devices. The royalties for using HEVC have increased [83] and this fact prevents some CPs from using HEVC (and maybe its successor VVC) and drives their interest towards royalty-free alternatives [84]. VP8 and VP9 were developed and released by Google, which later joined the Alliance for Open Media (AOM) with other mayor tech companies to work on the AV1 video codec. Nevertheless, as the MPEG-Google/AOM codecs struggle is still on-going, other factors may influence CP decisions on the employed codec, including limitations from device manufacturers (hardware encoding and decoding capabilities) and/or browser capabilities [85].

Operational Expenditure (OPEX) refers to on-going costs for running the business and inherent to the operation of the assets. Except from expenses common with every business, encoding and CDN usage are the most relevant and specific to media streaming services. Once the codec has been chosen, the encoding operations may have other operational costs that vary depending on the encoder choice (i.e., open-source or commercial) and where the encoder runs (i.e., cloud or on-premise encoding). Cloud encoding prices are established by cloud providers [86], while on-premise coding depends on the hardware selection and maintenance. On the other side, on-premise encoding allows to have more control on the processed data and content compared to cloud encoding [87], [88]. The total encoding cost is expressed in equation (7).

$$\begin{aligned} Enc_{cost} = & codec_{royalties} + encoder_{price} + \\ & + server_{cost} + processing_{cost} \end{aligned} \tag{7}$$

In equation (7), codec$_{royalties}$ + encoder$_{price}$ expenses belong to CAPEX, while processing$_{cost}$ is an OPEX. server$_{cost}$ depends on the strategy, a cloud encoder generates an OPEX, while an on-premise encoder needs an equipment investment which is a CAPEX. Focusing on processing$_{cost}$, employing Per-Title encoding and CMAF can reduce the OPEX. Per-Title encoding enables optimization of the encoder adjustment [89], and provides a more effective bitrate and resolution choice to optimize the trade-off between QoE and processing resources. CMAF guarantees compatibility between different HAS technologies, meaning that the encoded content can be shared between them. Thus, a single encoding operation is necessary for both DASH and HLS [90].

The cost of CDN resources depends on their ongoing utilization [91]. Not all providers publish their own pricing plans since in most of the cases they offer personalized plans to each customer. Nevertheless, [92], [93] and [94] reveal common factors that influence the OPEX for CDN resources, such as the outbound network traffic, storage occupancy and usage time. Thus, CDN OPEX can be expressed as (8) [95]:

$$\begin{aligned} CDN_{cost} = & \sum_{i=1}^{N} (\alpha_{loc_i} * Tr_i + \beta_{loc_i} * K_{req_i} + \\ & + \gamma_{loc_i} * T_i + \delta_{loc_i} * St_i + \epsilon_{loc_i}) \end{aligned} \tag{8}$$

In equation (8), *Tr*$_i$ and K_{req_i} represent the traffic volume and the number of HTTP requests producing this traffic. T_i and *St*$_i$ are the utilization time for a CDN (active sessions from video players) and the employed storage at CDN, respectively. Finally, α_{loc_i}, β_{loc_i}, γ_{loc_i}, δ_{loc_i}, and ϵ_{loc_i} are multiplicative coefficients established by the CDN provider and are dependent on the location of CDN resources (the cost of a cloud server depends on the geographical location). CPs usually employ simultaneously more than one CDN to increase coverage and in consequence the addition means a sum over the *N* available CDNs. The values of the coefficients depend on the business model and the pricing plan of each CDN provider. On the contrary, the variables which depend on the CDN usage (*Tr*$_i$, K_{req_i}, T_i and *St*$_i$) can be exploited by the CP to optimize the CDN resource selection and achieve a trade-off between QoS/QoE and cost.

D. Energy Efficiency

Efficient usage of energy has become a worldwide critical challenge. There is a very strong motivation for researchers to propose and develop energy efficient techniques in order to manage the power consumption in both current and future network environments. The range of green networking solutions covers a wide area. There are centralized network-centric approaches, where operators would deploy and positively influence large scale systems. A different strategy is based on individual solutions, which can be deployed considering a user-centric paradigm. There are energy preservation solutions which target equipment functionality and others which influence data exchange protocols, mechanisms which involve single components and others which target communication and cooperation between units, solutions deployed at a single network layer or across multiple layers, schemes which are made public and approaches which are proprietary, etc.

In this complex energy-aware research, there is a natural interest on solutions for energy efficient delivery of multimedia content with focus on end-user terminal devices. The latest wireless smart mobile devices are deployed with limited battery-based power resources, while computational and content presentation-related complexity has increased exponentially. Kennedy et al. [96] studied mobile device components' energy consumption. These authors have noted that screen, CPU, audio and network units scored the highest in terms of energy consumption, with a large gap between minimum and maximum values for the presentation components (i.e., screen and speakers). Lately, by using hardware optimization solutions for content presentation, the consumption associated to screen and audio interfaces has been reduced at the cost of increased processing complexity as well as increased data transfer. It is therefore fundamental to achieve energy efficiency for rich media content exchange between smart device and other sources or in-between such devices in order to extend operational activity of the devices and support high user QoE.

For a device, energy consumption E is the sum of the energy consumed in data transmitting mode (Tx), data receiving mode (Rx), sleeping mode (Sl) and during state transition (Sw).

$$E = E_{Tx} + E_{Rx} + E_{Sl} + E_{Sw} \tag{9}$$

Energy efficiency is generally defined as information bits per unit of transmission energy. A typical function of energy efficiency calculation for an additive white Gaussian noise channel is shown in equation (10) [97]:

$$\eta = \frac{2R}{N_0(2^{2R} - 1)} \tag{10}$$

where the channel capacity R is defined as in equation (11):

$$R = \frac{1}{2} \log{(1 + \frac{P}{N_0 B})} \tag{11}$$

and P represents the transmit power, N_0 represents the noise power spectral density and B represents the system bandwidth.

However, most solutions require dynamic computation of energy consumption and the components with the largest contribution to the overall energy budget are E_{Tx} and E_{Rx}. Even though transmission energy consumption is expected to exceed the ammount required by reception functions, the literature associates the energy consumption with network interfacing activity in general, and not with any specific communication.

There are two models oftenly used in the literature. In the context of a mobile device, equation (12), proposed by Trestian et al. in [98], calculates the energy consumption as follows:

$$E = t(r_t + Th * r_d) \tag{12}$$

where E is the estimated energy consumption (Joule) for a RAN, t represents the transaction time (seconds), r_t is the mobile device's energy consumption per time unit (Watt), Th is the throughput (Kbps) and r_d is the energy consumption rate for the data stream (Joule/Kbyte). The parameters r_d and r_t are device specific and differ for various network interfaces present at the device side.

A second power consumption model for a sensor node was introduced by Zou et al. in [99] and is described by equation (13). According to this model, the theoretical power consumption p_r of a wireless interface r is proportional to the throughput Th_r, as indicated in equation (13).

$$p_r(Th_r) = \alpha_r * Th_r + \beta_r + \gamma_r \tag{13}$$

In equation (13), p_r is the power expressed in Watts, α_r is the energy consumption rate for data in mJ/Kb for the interface r, Th_r denotes the data rate in $Kbps$ on interface r, β_r is the energy consumption per unit time in $mWatt$ for the interface r and γ_r is a constant which is a tunable value associated with the background energy consumption for interface r. If the node is equipped with a total of R interfaces, the total power consumption is calculated as in equation (14):

$$P = \sum_{r \in R} p_r(Th_r) \tag{14}$$

VI. Network Traffic Monitoring and Analysis

This section overviews major research activities related to network and traffic monitoring. Studies are performed from different points of view, as shown in Figure 3. Research activities include investigations on statistical models which can approximate network traffic behavior and analysis of the network traffic to acquire valuable information to be used to improve the network performance. In this section we also include a review of tools employed for performance monitoring and network simulation.

A. Network Traffic Models

Over the years, network traffic models have been thoroughly studied within the communication networks domain to describe the behaviour of discrete entities, namely, packets, connections, etc. In statistics and probability theory, this kind of traffic is described as a Point Process [105]. Many models have been proposed, with advantages and disadvantages, appropriate or not to different types of networks (i.e. Ethernet, Wi-Fi, LTE/5G, etc.) and with support for diverse scenarios. The choice of traffic model to employ depends on the particular network under study and the demand characteristics. The models are useful to perform any optimization and to produce a robust and reliable network infrastructure design. Moreover, they are essential to design and experiment network services since they can be employed to recreate a realistic traffic scenario in a controlled environment (laboratory). Thus, network services are tested and validated before being deployed in production. Obviously, each model has some assumptions that limits its usage. In other words, the models are not perfect, but their approximation is good enough for experimentation purposes.

TABLE IX
Application-agnostic network traffic models.

Model	Description
Poisson [100]	Memoryless distribution of the arrivals from independent sources (Poisson sources)
Non-stationary Poisson [101]	Non-stationary Poisson behavior at multi-second time scales
Log-normal [102]	Inter-arrival times from aggregated sources modelling
Pareto [102], [103]	Inter-arrival times from aggregated sources modelling, End-to-end delay modelling
Weibull [104]	Inter-arrival processes (packets, flows and sessions) modelling
Markov [15]	Model of activities of a traffic source with exponentially distributed time between state transitions
Embedded Markov [15]	Model of activities of a traffic source with arbitrary probability distributed time between state transitions

In [19] and [20], the most common application-agnostic network traffic models are presented, i.e., these models focus on characterizing generic network packet arrivals. The Poisson distribution model is one of the oldest models, but it is still widely employed across the literature to model packet arrivals from independent sources [100]. The authors of [101] carried out a deep analysis of network traffic to study the limitations of the Poisson model. The authors propose a non-stationary Poisson model as the Poisson model accurately characterizes traffic only at sub-second time scales. At multi-second time scales the traffic seems to have a non-stationary behavior. The Log-normal and Pareto distributions are employed to model inter-arrival times from aggregated sources [102]. Moreover, the Pareto distribution also models end-to-end network delay [103]. The Weibull distribution describes inter-arrival processes at different levels, meaning that it fits with packets, flows and sessions arrivals by tuning its parameters [104]. Markov and Embedded Markov models are used for network sources with a finite number of states, e.g., voice telephony has idle, busy and transmit states [15], and they differ in describing the time between state transitions. Table IX presents the most employed application-agnostic models.

Other studies have focused on media specific applications instead on traffic-agnostic ones. In [33], the authors employ some models proposed in literature to describe the traffic generated by specific 5G use cases/applications. In [106], a traffic analysis of an IPTV CDN network is presented. The authors find that the bitrate of multicast flow is relatively stable and depends on the number of live broadcast channels, while the bitrate of a unicast flow varies along the day and presents differences between weekdays and weekend. In [107], the traffic characteristics of Netflix and YouTube were analyzed, and the findings reveal that data is transferred through ON-OFF cycles, whose duration is dependent on the user's device

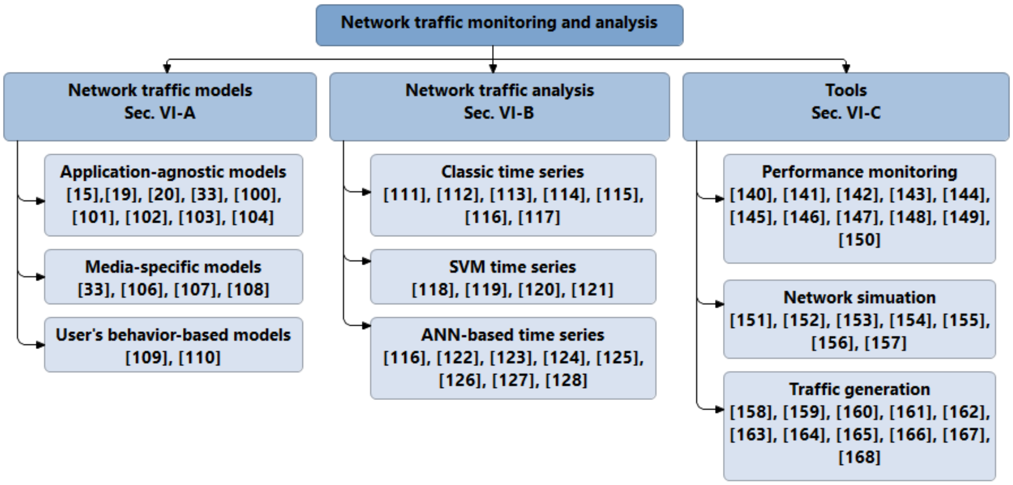


Fig. 4. Network traffic monitoring and analysis.

and browser. In [108], a study of YouTube traffic reveals that the traffic is highly dependent on the hour of the day. Moreover, the inter-arrival between two consecutive video requests depends on the popularity of the video.

Finally, [109] and [110] present two studies of user behavior while accessing streaming services. In [109], the authors focus on VOD streaming and they note that the user inter-arrival rate can be modelled by a modified Poisson distribution. Once streaming was accessed, the session length varied depending on the video duration. Furthermore, they find that more than half of the overall sessions end within ten minutes, while more than one third ended within 5 min. The user behavior has also some variations depending on the day of the week, as during the weekend, the video requests increase. In [110], the authors also consider live streaming. They find that a Poisson distribution is less accurate when modelling user inter-arrival for live streaming services than for VOD ones.

B. Network Traffic Analysis

The ability to model and generate realistic network scenarios offers the possibility to design and deploy network functions that adjust to the network traffic at any moment. Analyzing network traffic and applying time series analysis means a further step since it allows to forecast future network traffic. Network functions could move from reactive to proactive approach by exploiting predicted future conditions of the network. Actions are proactively taken when performances are going to not be satisfied. Thus, network under-performance and outages are prevented.

There are many methods proposed in literature for time series analysis. Among them, we distinguish classic time series, Support Vector Machine (SVM) time series and Artificial Neural Network (ANN)-based time series [129]. The choice of a predictor based on one of the different time series approaches depends on the characteristics of the network and different approaches are suitable for traces from different sources [130]. Moreover, in the same scenario different approaches could be combined to predict both long-term traffic demand and short-term network metrics [131]. Classic time series approaches are well known, as they were defined prior to the raise of telecommunication networks. On the contrary, SVM and ANN-based solutions can be seen as new contenders to classic ones, as Machine Learning (ML) application for time series prediction [132] is relatively new (SVM and ANN are two different supervised learning approaches). Time series methods employed for network forecasting are shown in Table X. The table also presents the main differences between them, such as the number of input and output variables and the selection of internal parameters. ML (SVM and ANN) models take the advantage from the knowledge of several variables as input (multivariate), while classic ones are limited to one (univariate). The same is valid for output variables, ML models can output more than one. The outcomes of [133] and [134] reveal that a higher number of input variables improves the traffic predictions of a ML model (the authors employ a Long short-term memory model) when compared to a classic one (the authors employ an autoregressive integrated moving average model).

The autoregressive integrated moving average (ARIMA) [135] is one of the oldest time series method and widely employed in literature as reference method for evaluating any other time series approach. In [111], ARIMA is employed to predict the workload of cloud services. Historical observed requests are exploited to predict the volume of requests during the next time interval. The authors find limitations to track traffic peaks accurately. In [115], ARIMA is instead employed to predict the request number and the amount of data traffic. Other limitations to ARIMA are found in [113] and [114] when modelling QoS attributes which have non-linear behaviors, i.e., time between QoS violations. Thus, they do not fit the linear assumption of ARIMA. Self-exciting threshold autoregressive moving average (SETARMA) [113] and generalized autoregressive conditional heteroskedastic (GARCH) [114] are integrated with ARIMA in hybrid linear and non-linear models to overcome ARIMA limitations.

Exponential smoothing [136] is a subset of classic time series method. Holt's linear trend method (secondary or double exponential smoothing) is employed in [112]. The authors find it complementary to ARIMA when predicting throughput in an

TABLE X
METHODS FOR TIME SERIES ANALYSIS APPLIED TO NETWORKS.

Method	References	Approach	Number of variables	Configuration / parameters	Description
ARIMA	[111], [112], [113], [114], [115]	classic	univariate	regression, integration and moving average parameters	Autoregressive integrated moving average
SETARMA	[113]	classic	univariate	regression, moving average and threshold delay parameters	Self-exciting threshold autoregressive moving average
GARCH	[113]	classic	univariate	regression and lag length parameters	Generalized autoregressive conditional heteroskedastic
Holt's linear trend	[112]	classic	univariate	smoothing factor	Secondary or double exponential smoothing time series
Holt-Winters' seasonal	[116], [117]	classic	univariate	smoothing factor	Cubic or triple exponential smoothing time series
SVR	[118], [119]	SVM	multivariate	weight vector and offset	Support Vector Regression
H-SVM	[120]	SVM	multivariate	weight vector and offset	Hierarchical Support Vector Machine
Multi-class SVM	[121]	SVM	multivariate	weight vector and offset	Multi-class Support Vector Machine
Feed-forward NN	[122]	ANN	multivariate	weight and bias	Feed-forward neural network
MLP	[116], [123]	ANN	multivariate	input vector, weight vector and bias	Multi-layer Perceptron
FNN	[123]	ANN	multivariate	input vector, weight vector and bias	Fuzzy Neural Network
RNN	[124]	ANN	multivariate	input, output and forget factors	Recurrent neural network
LSTM	[125], [126], [127]	ANN	multivariate	input, output and forget factors	Long short-term memory
ESN	[128]	ANN	multivariate	input, reservoir and output weights	Echo State Network

LTE network. ARIMA outperforms the exponential smoothing on weekdays, while the exponential smoothing prediction are more accurate on weekends. Holt-Winters' seasonal method (cubic or triple exponential smoothing) is instead employed in [116] and [117]. In [116], it is employed to implement an anomaly detection, while, in [117], its aim is to predict cloud resource provisioning.

Among SVM time series [137], in [118], a Support Vector Regression (SVR) model is employed to predict TCP throughput. A similar approach with SVR is presented in [119], but it aims to predict network links load and not limited to TCP traffic. In [121], the authors use Channel State Information (CSI) and handover history to determine a user's mobility pattern by means of a Multi-class SVM. The next cell can be predicted based on the previous crossed cells, user's trajectory, and CSI. The problem of estimating the location of mobile nodes is investigated also in [120], but limited to an indoor wireless network, and employing a hierarchical SVM model composed of four different levels. The same method is also employed to estimate channel noise.

Concerning ANN-based approaches, in [122] a Feed-forward Neural Network (Feed-forward NN) for predicting the execution time of services while varying the number of requesters is presented. In [124], a Recurrent Neural Network (RNN) is instead employed to forecast the end-to-end delay from RTT metrics. In [125], a Long short-term memory (LSTM) model, a particular type of RNN, is proposed to process downlink control information (DCI) messages, such as resource blocks, transport block size, and scheduling information. LSTM is also employed in [126] to solve a problem of traffic matrix prediction and in [127] to forecast stalling events during a video streaming session. An Echo State Network (ESN), also a kind of RNN, is employed in [128] to predict traffic volume in a city for various network applications, such as Multimedia Messaging Service (MMS), Web, media streaming, Instant Messaging (IM) and Peer-to-peer (P2P) communication. In [116], a Multi-layer Perceptron (MLP) model is employed to detect anomalies in network traffic. MLP is used jointly with a Fuzzy Neural Network (FNN) in [123] to forecast one-step ahead value of the MPEG and JPEG video, Ethernet, and Internet traffic data. The combined results of the two ANNs outperforms the results achieved by employing only one method.

Being able to forecast network traffic and performances is definitely interesting to provide proactive actions in response to future network issues. In any case, there is not an optimal method, as the better performing method depends on the considered metrics and scenarios. As a result, some hybrid solutions are also being investigated to exploit both the advantages of classic methods and ML (SVM or ANN) ones [138], [139].

C. Tools

1) Performance monitoring: Performance monitoring, including metrics collection and visualization, can be done though several visual analytics tools, as shown in Table XI. Tools are classified depending on the application domain.

Prometheus [140], InfluxDB [141], Grafana [142] and Elastic Stack [143] are open-source and general-purpose solutions. Prometheus [140] and InfluxDB [141] are time series database to collect, monitor and visualize real-time information. Anyway, their visualization capabilities are limited, and thus, they are usually employed jointly with external tools to create and visualize interactive data charts. Grafana [142] is the most common tool for these interactive data charts. It can connect to both Prometheus and InfluxDB or any other database to access data and manage them to create interactive web visualization. Several data charts can also be visualized at the same time by generating a unique dashboard to simplify decision-making operations. Elastic Stack [143] is an alternative to Grafana, but it comes with its own database, called Elasticsearch, and data visualization component, called Kibana, to generate data charts and dashboards. It has a modular architecture to allow adding optional add-ons to

TABLE XI
TOOLS FOR PERFORMANCE COLLECTION AND VISUALIZATION.

Domain	Processing model	Automation mode	Forecast Skills	Name	Description
General	Real-time inputs	API and manual GUI	Not Applicable	Prometheus [140]	General-purpose monitoring system and time series database
General	Real-time inputs	API and manual GUI	Not Applicable	InfluxDB [141]	General-purpose monitoring system and time series database
General	Real-time inputs	API and manual GUI	Not Applicable	Grafana [142]	General-purpose platform for decision-making with focus on customizable data charts
General	Real-time inputs	API and manual GUI	Predictions	Elastic Stack [143]	General-purpose platform for monitoring and decision-making with focus on customizable alerts and data charts
Business	Batch and real-time inputs	No API, manual GUI	Predictions and Simulations	Board [144]	Business-purpose platform for decision-making with focus on customizable CRM data charts
Business	Batch, scheduled and real-time inputs	No API, manual GUI	Not applicable	Tableau [145]	Business-purpose platform for decision-making with focus on customizable CRM data charts
Web sessions	Real-time inputs	No API, manual GUI	Not Applicable	Citrix Analytics [146]	General-purpose web activity including user session performance and application usage
Web sessions	Real-time inputs	API and manual GUI	Not Applicable	Google Analytics [147]	General-purpose web activity
Media	Batch, scheduled and real-time inputs	API and manual GUI	Not applicable	Akamai Media Analytics [148]	Media streaming service-specific Analytics solution
Media	Real-time inputs	API and manual GUI	Not Applicable	Conviva Streaming Analytics [149]	Media streaming service-specific Analytics solution
Media and Data	Real-time inputs	API and manual GUI	Not Applicable	Amazon Kinesis [150]	Media streaming service-specific Analytics solution

increase its capabilities. Among these add-ons, a Machine Leaning (ML) one can enable algorithms to analyze the data.

Board [144] and Tableau [145] are commercial software intended for business analytics. They focus on Customer Relationship Management (CRM) and data charts creation for enabling decision-making. Citrix Analytics [146] and Google Analytics [147] aim to track web activities (browser video players). While Citrix Analytics is a commercial solution, Google Analytics has both a commercial and a free version. This free version is usually enough for research activities. Akamai Media Analytics [148] and Conviva Streaming Analytics [149] are meant for media-specific application, as they manage metrics related to online streaming. Finally, Amazon Kinesis [150] is focused on both media and generic data streaming, as it allows to collect real-time data from heterogeneous sources, such as video and audio, application logs and IoT telemetry.

2) Network simulation and traffic generation: Achievements in network traffic modeling and analysis are widely exploited to develop utility software which simulates real networks and/or generates realistic traffic for experimentation. Table XII shows several tools enabling research activity and experimentation with network traffic.

Network simulators allow to simulate networks without having to deploy a real one. A single node running a simulator is employed to generate a network whose capabilities and performance are configurable. Network simulators replicate the physical layer (L1), wireless (Wi-Fi, LTE, 5G) or wired (Ethernet) [151], [152], [153], and configure network typologies to be employed during the experiments [154], [155]. Moreover, almost all simulators are designed to enable the exchange of packets belonging to different L2/3/4 protocols (Ethernet, IP, UDP/TCP). In some cases, they also allow to reproduce more specific network technologies or environments (IoT, WSN, DTN) [156], [157]. Definitely, they are useful when testing through a real network is not feasible due to several reasons, such as equipment costs or physical space for assets.

On the contrary, if a real network is available for experimentation, it is necessary to ensure that the traffic crossing the network has similarities with a real one. In this sense, the use of traffic generators become prominent to guarantee that the network exhibits a realistic behavior. There is a huge number of network simulators having a wide range of capabilities. Basic tools are already provided by Linux kernel-based OS distributions [158], [159] or provide a more user-friendly access to Linux kernel modules to generate traffic [160], but their capabilities are usually limited when needing to generate a specific packet distribution profile. More sophisticated solutions allow to select a specific traffic patterns generated at different OSI layers. The simplest ones are limited to model L3/4 packets [161], [162], [163], while others enable also L7 [164], [165], [166], [167], [168]. While L3 generation aims is to characterize IP flows and L4 generation is mostly limited to choose between employing UDP or TCP-based packets, at L7 there is a wide range of applications. Then, each traffic generator that works at such layer has to specify which applications can be simulated. Different solutions allow to simulate Web traffic, e.g., HTTP/HTTPS [164], [165] or VoIP [166], [167], and also 3rd Generation Partnership Project (3GPP) protocols [168].

VII. PERFORMANCE-DRIVEN NETWORK FUNCTIONS

This section presents an overview of VNF-based solutions designed to improve the performance of media streaming. These solutions employ knowledge that comes from network studies and data acquired from live monitoring of network traffic. Figure 5 illustrates the major avenues that performance-driven VNF solutions take. First we introduce NFV Management and Orchestration and Multi-access Edge

TABLE XII
TOOLS FOR NETWORK SIMULATION AND TRAFFIC GENERATION.

Category	Name	OSI layers	Description
Network simulator	OMNeT++ [151]	L1/2/3/4	Simulation of communication networks, multiprocessors and distributed or parallel systems
Network simulator	NS-2 [152] / NS-3 [153]	L1/2/3/4	The Network Simulator (NS) -2 / -3, Simulation of TCP, routing, and multicast protocols over wired and wireless networks
Network simulator	OPNET [154]	L1/2/3/4	Optimized Network Engineering Tool (OPNET), Simulation of network typologies, nodes and flows
Network simulator	Mininet [155]	L1/2/3/4	Instant Virtual Network to develop and experiment with SDN
Network simulator	NetSim [156]	L1/2/3/4	Simulation of heterogeneous networks and protocols (5G NR, IoT, WSN, Cognitive Radio, TCP)
Network simulator	The ONE [157]	L1/2/3/4	The Opportunistic Networking Environment (ONE) simulator, Evaluation of DTN routing and application protocols (sparse mobile ad-hoc networks)
Traffic generator	iPerf [158]	L3/4	Tool for active network performance measurement
Traffic generator	packETH [159]	L3/4	Packet generator tool for Ethernet
Traffic generator	pktgen [160]	L3/4	Testing tool included in the Linux kernel
Traffic generator	Moongen [161]	L3/4	Flexible high-speed packet generator
Traffic generator	Brute [162]	L3/4	Brawny and RobUstT Traffic Engine (Brute), Generation of traffic workloads having common traffic profiles
Traffic generator	Harpoon [163]	L3/4	Application-independent tool for generating representative packet traffic at the IP flow level
Traffic generator	Ostinato [164]	L3/4/7	Generation of specific traffic flows with various protocols
Traffic generator	TRex [165]	L3/4/7	TRex - Realistic Traffic Generator, Emulation of L3-7 traffic
Traffic generator	D-ITG [166], [167]	L3/4/7	Distributed Internet Traffic Generator, Synthetic network workload generator to emulate various applications (DNS, Telnet, VoIP and network games)
Traffic generator	Seagull [168]	L3/4/7	Multi-protocol traffic generator test tool

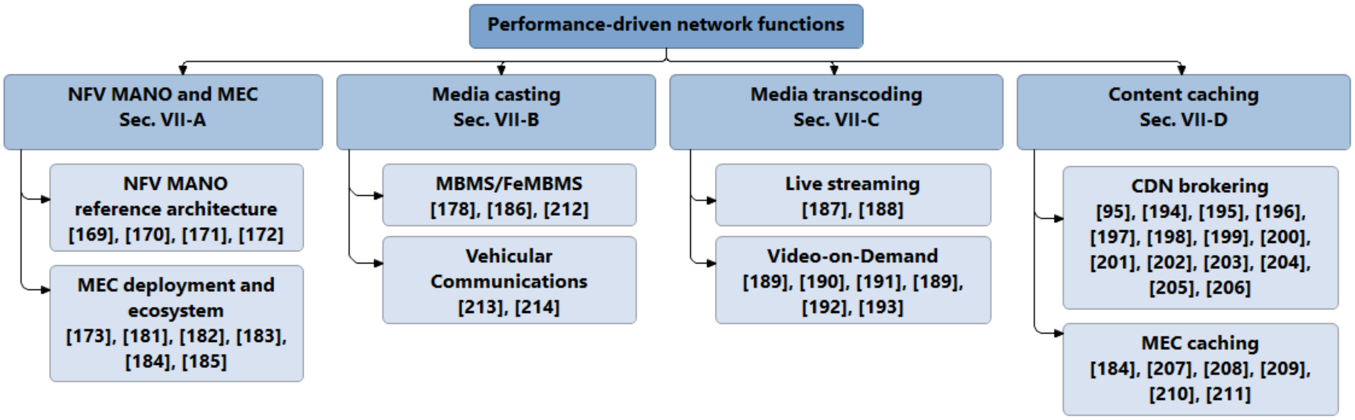


Fig. 5. Performance-driven Network Functions.

Computing, as VNFs rely on these paradigms introduced by European Telecommunications Standards Institute (ETSI) and embraced by 5G networks. Then, we discuss the state-of-the-art of most relevant media-related functions such as media casting, media transcoding and content caching.

A. NFV Management and Orchestration and Multi-access Edge Computing

Apart from the performance leaps on Key Performance Indicators (KPI) in terms of speed, capacity, mobility, and reliability, brought by 5G radio technologies, the network core is also fully engaged in a revolution, involving its own digital transformation. The concept that one network fits all is over. It is time to adapt the network according to applicable resources efficiency and delivery performance trade-offs. The goal is to allow network management systems to coordinate the systems comprising an agile, programmable and efficient network. This vision is being fueled by the transformation of network functions into dynamically controllable and configurable software components, which are virtualized exploiting cloud technologies and their scalable mechanisms, where orchestration of distributed network functions is done on top of the dynamic configuration of software systems. Going

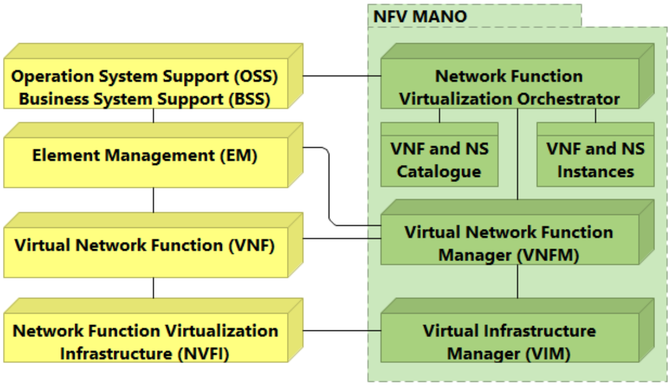


Fig. 6. ETSI NFV MANO architecture.

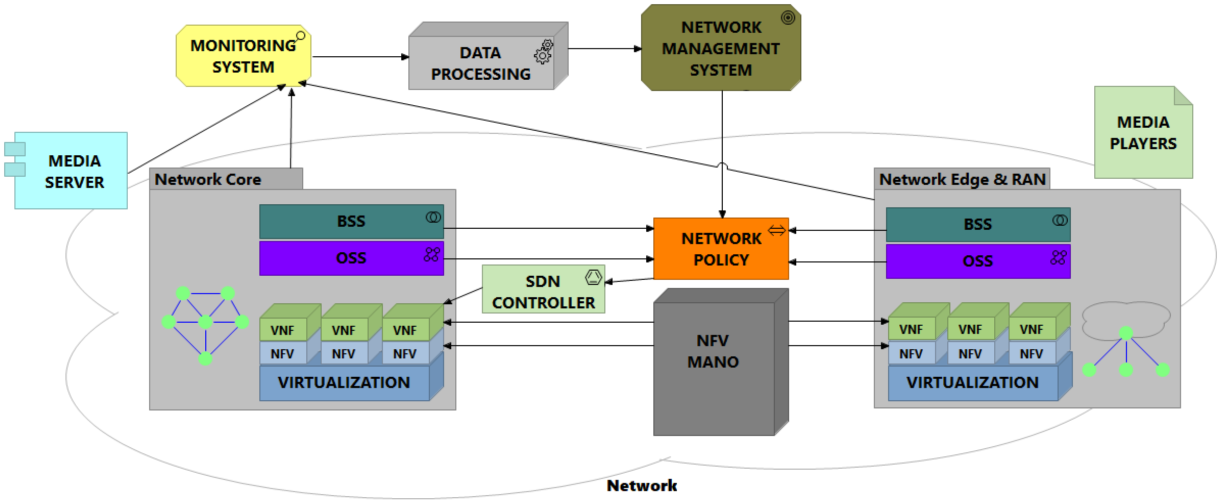


Fig. 7. ETSI NFV architecture applied to Media streaming services.

beyond, catalyzed by the network slices concept, the network would also connect groups of virtualized functions devoted to specific data flows or groups of users of specific services, handling independently Service Level Agreements (SLAs) of multiple points of presence (PoPs) over a common bare-metal infrastructure.

To achieve it, 5G network embraces NFV and VNF [169] concepts and comes with a NFV Management and Orchestration (MANO) architecture [170], standardized by ETSI.

NFV brings the primary virtualization step, providing computing, memory, storage and network resources from a bare-metal infrastructure (NFV Infrastructure or NFVI). The utilization of NFV contributes to the deployment of a network providing hardware and software decoupling. Thus, commercial off-the-shelf (COTS) hardware can be used to run every network function having a software implementation (VNF). This architecture is mainly employed by cloud vendors in order to provide Infrastructure as a Service (IaaS) solutions. Thus, hosting for systems on top of hardware and connectivity setup is performed on demand.

VNFs goes a step further in virtualization deploying specific network functions on top of NFVI. VNFs can be deployed, configured, started or stopped in a programmable manner. Thus, VNFs are intended to enable modularity, interoperability, scalability and flexibility when a media streaming service is managed, and the generated traffic is delivered.

NFVI and VNFs are managed and orchestrated by NFV MANO, whose reference architecture is shown in Figure 6. Its functional blocks are:

- Virtual Infrastructure Manager (VIM): It manages and controls physical and virtual resources (compute, storage and networking resources). Once a VNF is instantiated (VNF Instance or VNFI), it provides the VNFI with the resources it requires.
- VNF Manager (VNFM): It is responsible for the management of the life cycle of VNFI through the resources provided by the VIM.
- NFV Orchestrator (NFVO): It combines more than one VNF to create end-to-end services. Several VNFs could share VIM resources and be meant to be used for the deployment of a unique Network Service (NS), e.g., one VNF deploys the back-end and another one the front-end, the combination of the two VNFs constitute the NS.

Since 5G architecture allows for both public and private network deployment, existing NFV MANO-compliant solutions encompass both commercial and open-source alternatives for each of the three components. Some examples are Open Source MANO (OSM) [171], whose development is promoted by ETSI, and Open Network Automation Platform (ONAP) [172], supported by Linux Foundation.

All the described technologies that turn network functions into virtualized software systems facilitate a high level of automation and orchestration by network management systems. This trend is being deeply explored and investigated in the current generation of mobile networks (5G) and it will be key pillar for next ones (beyond 5G) and Multi-access Edge Computing (MEC) infrastructures [173]. MEC architectures enable context-aware applications. It opens computing infrastructures co-located with the base stations to host services closed to the mobile users exploiting the capillary distribution of cloud computing infrastructures at the edge of the cellular Radio Access Network (RAN).

The application of NFV and VNF technologies at the edge and the evolution of the RAN towards software components boosted by open-source software, such as OpenAirInterface [174] or srsLTE [175], eases the integration of MEC services with RAN systems. These solutions implement the Mobile Packet Core (Evolved Packet Core for LTE, 5G Core for 5G) and the RAN on top of open-source hardware enabling the deployment, management and orchestration through NFV MANO of both the mobile packet core [176], [177] and RAN [178]. A RAN deployment through NFV and VNF is usually referred as virtual RAN (vRAN). vRAN is also evolving towards the concept of Open RAN (O-RAN) [179], having open interfaces and network intelligence as key enablers to manage and tailor the network based on vendors and operators' requirements. O-RAN enables multi-vendor vRAN deployments, resulting in a more competitive and richer ecosystem [180]. In this context, MEC is a NFV MANO-compliant platform that comes also with a specific API to access Radio Network Information (RNI) [181].

Figure 7 shows how the virtualized (NFV) and softwarized (VNF) systems at the network core and edge are monitored and orchestrated according to business and technical policies which ask for changes in the NFV MANO system or SDN controller. Thus, any dynamic changes of the network can be applied over a widely-employed technology stack.

While focusing on the edge architecture, Figure 8 illustrates the MEC components and their interactions with the rest of the building blocks of RAN and Core Network (CN). The MEC host manages the User-plane, while the Data-plane communication is managed by the CN (LTE Evolved Packet Core or 5G Core). Depending on whether the deployment is within an LTE or 5G network, MEC host is equipped with User-plane Serving and Packet Gateways (SGW-U and PGW-U) or User Plane Function (UPF), respectively. These components are connected directly to the base station (eNB for LTE or gNB for 5G) and provide access to Internet. Inside the MEC Host, the RNI service (RNIS) oversees collecting RAN information which is later consumed by the application VNFs. Specifically VNFs can be designed to exploit such information to increase the overall system performance.

Beyond this design, VNF is also applicable for media-specific network functions beyond the 5G core and RAN, involving:

- media casting, in order to perform massive delivery of live data flows,
- media transcoding, such as streaming rate matches network available bandwidth, resulting in higher quality at destination and
- content caching, including storing popular data to help improved high traffic conditions, and managing alternative endpoints to balance the data requests.

All these network functions perform specialized functions of the media applications in order to improve network efficiency, saving bandwidth overheads and favoring the allocation of idle resources to other network flows, and to enhance quality of experience with enforced KPIs according to SLAs.

ETSI includes several use cases related to media streaming to be considered for MEC deployment [182] empowering traditional media streaming applications, which are based on interaction between remote server (origin server or CDN) and client, as shown in Figure 9. MEC platform can host diverse VNFs, which exploit RNI to get a wider view of the local conditions to enhance media streaming service. In this line, some solutions, such as [183], [184], [185], exploit standard RAN interfaces and data reports to conclude better decisions for media applications.

The following sections analyze how the described core technologies of 5G are applied to expand the network functions with core components for improved delivery of media streams, resulting with benefits in terms of enhanced quality and efficient resources utilization. Accordingly, Table XIII compiles and classifies all the research activities exploiting 5G to support performance-aware networking. The classification highlights the main features implemented, as well as secondary aspects, as sometimes the same approach is applicable to more than one solution. Some proposals are limited to architecture design and do not achieve a real implementation and experimentation. The implemented ones differ in terms of activation and processing approach, as they could operate in reactive or proactive manner and, in same cases, embed a processing algorithm (classic or ANN-based). All proposed solutions aim to have direct impact on the performance of the media streaming systems, ranging from QoS and QoE enhancement to more effective business costs and energy saving. However, most of them do not provide specific validation tests, especially in terms of HAS-centric QoE metrics, or insights on applicable cost models which include business aspects or evidence on energy footprint.

B. Media casting

For massive delivery of common data at once, synchronously, broadcast is still much more efficient that unicast communications widely employed by cellular networks. That is why 3GPP introduced Multimedia Broadcast/Multicast Service (MBMS) specification in Long-Term Evolution (LTE) release 9, which has been evolved towards further enhanced MBMS (FeMBMS) in release 14 to enable higher per cell bandwidth for MBMS services and simultaneous reception of both unicast and multicast services [186]. Furthermore, release 16 includes feedback for increased reliability [212].

In fact, as this technology is tied to the RAN system, it has sense in some use cases as firmware/software updates, clock synchronization, alarms and massive media contents to be turned in the network edge from unicast communications to broadcast signals. This would need the support from MEC systems which will turn popular streams into broadcast flows to expand the capacity of a cell. This is feasible as manifests of HAS technologies, such as HLS or DASH, even for encrypted contents keep the manifests unencrypted allowing a simple processing to parsed them by intermediaries, such as CDNs or MEC systems, for efficient and smart media delivery.

This architecture brings three major benefits by means of attracting all the ongoing live sessions to consume the broadcast dataflow, instead of establishing concurrent unicast sessions:

1) Efficiency at the radio link, as the broadcast stream reduces radio link usage. Data traffic is independent of volume of users since everyone is consuming the same broadcast signal.
2) Optimal fidelity, as the network is able to deliver to all the audience the maximum resolution (bitrate representation).
3) Enhanced QoE, as the media players sharing the radio-link do not have to struggle with independent adaptive mechanisms executed in each player competing for the available bandwidth. This means no bitrate or resolution changes to track time-varying network conditions and no freezes to refill the buffer.

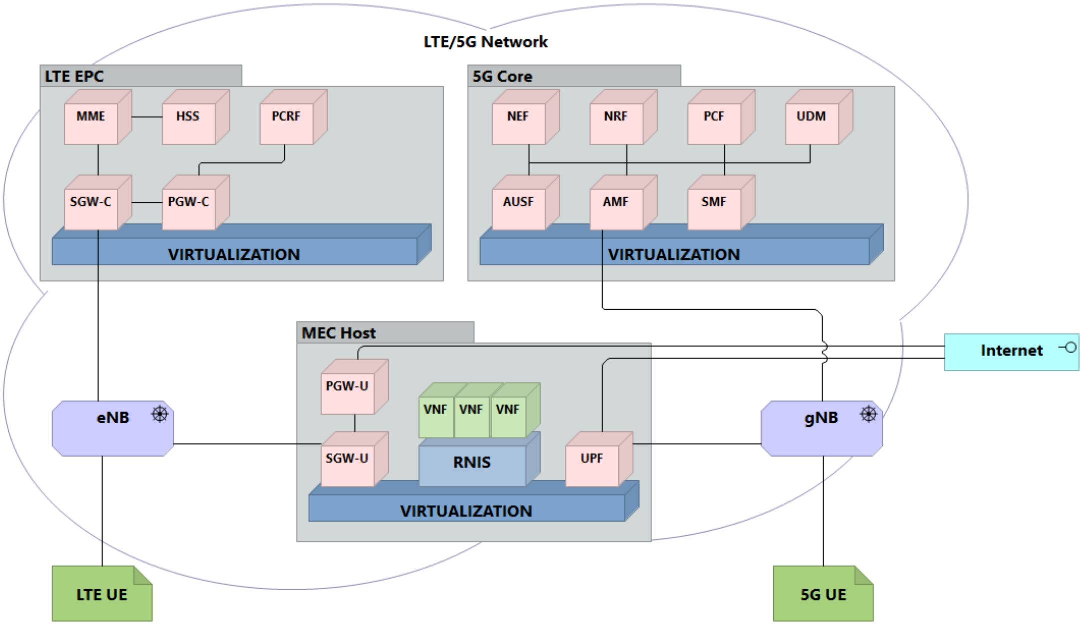


Fig. 8. MEC architecture and connection with RAN and CN. Fixed error: PGW-U/UPF

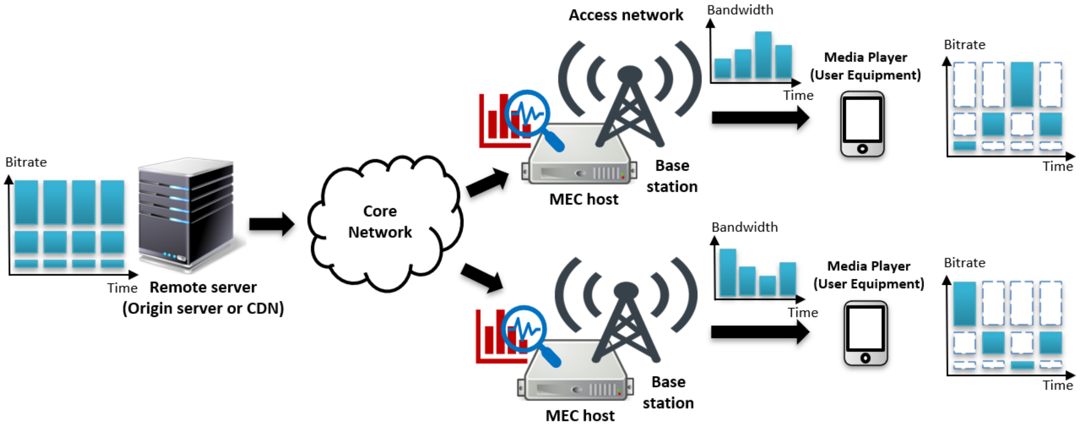


Fig. 9. MEC-powered media streaming.

This approach is possible thanks to the application of virtualization and softwarization paradigms to RAN technologies, making vRAN and the containerization of some RAN network functions such as FeMBMS feasible [178].

Specifically, broadcast communications are gaining relevance in the vehicular communications field as they allow synchronous provisioning of common awareness to vehicles, pedestrians and Road-Side Units (RSU) in a surrounding area. Common awareness can be essential for Cooperative, Connected and Automated Mobility (CCAM) applications related to safety of autonomous driving [213]. In these applications media flows are important as the vehicles gets fitted with more camera-like sensors capturing the environment and exchanging the raw/compressed data or processed insights/summaries from on-board computer vision systems [214].

C. Media transcoding

Media services have become a fundamental service in 5G networks. There, as summarized in Table III, HAS technologies, such as DASH or HLS, are widely employed and need the provision of several representations meaning different resolutions and bitrates [190]. Thus, VNF-based transcoders are being developed under international funding initiatives aiming to empower different use cases, e.g., live 3D media streaming [187] or automotive [215]. Here, the generation of representations at edge servers is gaining relevance to get higher efficiency by distributing the higher fidelity through the core and generating variants at the edge. This would reduce overheads in the core to send all the possible media variants. To this end, the media transcoding at the edge is essential [191], stressing the fronthaul capacity and requiring Cloud-RANs (C-RANs) or MEC systems in order to minimize the network delivery cost. Furthermore, the capillarity of the MEC systems brings a better adaptation to the local needs when transcoding to produce variants.

However, transcoding is a heavy process which needs a smart mechanism to gain efficiency. Transcoding at resource-constrained MEC server means a challenge for delay-sensitive services. Here, different works deal with the optimal position

TABLE XIII
PERFORMANCE DRIVEN NETWORKING FOR MEDIA STREAMS USING 5G TECHNOLOGIES.

Main Feature	Secondary Feature	Activation	Processing approach	References	Network features	Description
Casting	-	Not Applicable	Not Applicable	[186]	FeMBMS	Design of 3GPP architecture for media multicast
Casting	-	Reactive	Not Applicable	[178]	FeMBMS, VNF, SDR	Virtualization of FeMBMS with SDR setup
Transcoding	-	Not Applicable	Not Applicable	[187]	NFV, VNF, 5G Core	Design of centralized virtual transcoder solution at 5G Core
Transcoding	-	Reactive	ANN	[188]	VNF, MEC	On-the-fly transcoder at the network edge
Transcoding	Caching	Proactive	Classic	[189]	L1 MC-NOMA, MEC	Solution empowered by mulitcarrier non-orthogonal multiple access
Transcoding	Caching	Reactive	Classic	[190]	MEC, VNF	Transcoding and cache location in virtualized edge infrastructures
Transcoding	Caching	Reactive / Proactive	Classic	[191]	MEC, VNF	Transcoding and cache location when content popularity is known (proactive) or not (reactive)
Transcoding	Caching	Proactive	Classic	[192], [193]	MEC, VNF	Transcoding and cache location based on known content popularity
CDN Brokering	-	Reactive	Not Applicable	[194], [195], [196]	L7	Proprietary solution for selection of CDN vendor at startup
CDN Brokering	-	Reactive	Classic	[197], [198], [199]	L3 DNS	Performance-driven solution based on DNS resolution
CDN Brokering	-	Not Applicable	Not Applicable	[200], [201], [202], [203]	L3 DNS	Design of CDN-ISP collaborative solutions
CDN Brokering	-	Proactive	ANN	[95]	L7, L3	Solution for proactive CDN selection employing ANN algorithm to forecast network metrics
CDN Brokering	-	Reactive	Not Applicable	[204], [205], [206]	L7	Cloud solution for cost-effective CDN switching
Caching	CDN Brokering	Reactive / Proactive	Classic	[207]	L7, L3, MEC	Statistical solution for CDN selection (reactive) and content caching (proactive)
Caching	-	Not Applicable	Not Applicable	[208]	VNF, Orchestration	Design of virtual CDNs for media distribution
Caching	-	Proactive	Classic	[209]	MEC, SDR	Solution at edge exploiting radio network information
Caching	Fair QoE	Reactive	Classic	[184]	MEC, SDR	Solution at edge exploiting radio network information
Caching	Fair QoE	Reactive	Classic	[210]	MEC, SDR	Solution at edge exploiting radio network information and content popularity
Caching	-	Proactive	ANN	[211]	MEC	Solution for proactive caching employing ANN technologies to predict popularity

of transcoding systems in different edge hosts to respond to a distributed demand more efficiently and quickly, where players use a specific base station as a gateway linked to host and an edge server. To overcome this challenge, a mechanism for optimal request forwarding which respects the resources limitations and minimize serving latency is required [189]. In [192], different short/long-term decisions are concluded to deal with the time-varying conditions in terms of demand and network dynamics.

Beyond the planning of such transcoding process, other approaches consider different algorithms for reactive or proactive planning [191]. In this case, the dynamics have a big impact on the reaction time and forecast range. These aspects are minimized using a segment duration in the HAS stream with favor steady short-term conditions as changes comes in a segment duration-basis.

These works focus on enhancing QoS metrics while managing capacity of each processing asset. However, they do not consider heterogeneous SLAs and cost penalties to apply trade-off policies. As the GPU assets are required for HW-accelerated transcoding to ensure parallelization of transcoding threads and they have a big impact on infrastructure costs, this aspect should be a primary feature to evaluate.

It is important to underline that these solutions are often linked to caching strategies as both can be executed at the edge to better match the local conditions, patterns and demand features. Therefore, they design a joint strategy for transcoding processing and caching [190], [191], [189], [192], [193]. In [188], the authors only transcode the content on-the-fly if the content is not cached.

D. Content caching

1) CDN brokering: Caching is the most employed network function to improve the performance when accessing online contents and, in particular, media streaming ones. In this context, a CDN is the most popular network solution aiming to provide caching capabilities. It consists of a geographically distributed network of proxy servers and data centers to provide high availability of the contents. Caching mechanisms are key inside a CDN, as CDN proxy servers work by selectively storing the content such that the users can quickly access it from nearby locations. The employment of CDN service by the CPs increased in the last years, as the number of CDN vendors increased. Furthermore, major CPs also moved to multi-CDN strategies to provide a more reliable service while streaming their contents. Thus, an improved service also generates more satisfaction among the customers.

Nevertheless, how the different CDNs are employed can differ from a CP to another. Static selection of the CDN when a streaming session starts is the easiest and widely employed solution among the CPs. In 2012, this strategy was used by Netflix [194] and Hulu [195], with big similarities [196]. In both cases, they were using three different CDN vendors. They used to map the player device to a CDN depending on to its location or the subscriber when the streaming session starts. Moreover, the CDN is never changed during the streaming session, even when the performances decrease. Other solutions include client-side CDN selection [197] or Domain Name System (DNS)-based solutions [198]. Client has a privileged position to measure end-to-end QoS metrics (network bandwidth and latency) when choosing the CDN, but it has the advantage to produce an uncoordinated decision as each client selects the CDN independently from the others. A DNS-based solution means resolving a fixed hostname owned by the CP into different IP addresses referring to several CDNs. Depending on the DNS resolution, the client receives the content from the appropriate CDN. In any case, a sub-optimal CDN server selection could lead to performance decreasing [199], affecting the user's satisfaction.

In the last years, other network caching solutions are also raising to empower the delivery. The same Netflix changed its streaming strategies. It developed and deployed an in-house CDN, called Open Connect [200], to reduce the dependency from CDN vendors and streaming costs. Moreover, Open Connect is meant to be run also inside the ISP infrastructure, i.e., closer to the user, to guarantee better performances in terms of network bandwidth and latency [216]. The use of Open Connect also helps Netflix and other CPs having in-house solutions to better control the resources enabled for the streaming session and to reduce the costs. Anyway, it requires a large investment to have such a solution and it could not be affordable by small CPs.

The Streaming Video Alliance (SVA) is a joint initiative which works on different aspects of media streaming and aim to standardize the employed protocols and technologies. Its membership includes some of the major world-wide agents in content production and streaming. Among its activities, the SVA Open Caching Working Group [201] oversees identifying the critical components of a non-proprietary caching system and establishing the basic guidelines for its implementation inside the ISP infrastructure. Thus, it wants to promote an architecture similar to Netflix' Open Connect, but with the advantage to be standardized.

Other collaborations between CDN and ISP are proposed in literature. In [202], ISP provides the CDN provider with information concerning geographical user distribution and allows the CDN provider the possibility to allocate server resources inside the ISP network. The authors of [203] use a redirection center instance inside the ISP network which intercepts the client requests and selects the appropriate CDN server. The process is transparent to the client as the redirection center employs a CDN surrogate to store the content and instructs an OpenFlow controller to migrate the traffic to the CDN surrogate. Beyond the employment of multi-CDN solutions, there are still possibilities of improvements. CDN Brokering [217] is proposed to make more effective CDN utilization in a multi-CDN environment. It redirects clients dynamically between two or more CDNs.

CDN brokers work as switching services that dynamically and seamlessly select the optimal CDN to use at any time. To achieve this, CDN brokers collect and analyze in real time the performance of the available CDNs to select the best one. Thus, network analytics have a prominent role in CDN selection, in contrast with traditional multi-CDN strategies where the same CDN is kept during the streaming session. The approach from [95] applies ANN technologies to forecast dynamic demand and changeable performance to make decisions including cost-performance trade-offs. In this context, a representative example is Eurovision Flow [204], proposed by the European Broadcasting Union (EBU). Similar solutions are also provided by Citrix [205] and Haivision [206].

2) Edge caching: In [207], a MEC proxy retrieves media streaming metrics of video players at the access point and CDNs performance metrics to enhance DASH media streaming. The MEC proxy evaluates the performance of different CDNs and switches players' sessions when a CDN is under-performing and cannot support the demanded traffic. Moreover, it features a local edge caching to reduce network traffic. Recurrent content is downloaded and cached once for every player. In [208], a similar MEC cache is proposed for empowering the delivery.

With a deeper integration with RAN interfaces, in [209] and [184] the MEC cache is improved by exploiting RNI. The media segments and representations are selectively cached depending on the network state. In [210], both RNI and knowledge of segment popularity are employed to decide the segments to cache. Moving from a reactive to a proactive approach, the authors of [211] empower the edge cache with neural collaborative filtering to predict content popularity. The predictions are exploited to proactively cache the content at the MEC, as more content popularity means higher probability to be requested by the users.

VIII. Challenges of Virtual Network Functions for Media Streaming

VNF solutions play a significant role in the successful deployment of 5G networks. This is backed by evidence, especially for supporting rich media applications such as multimedia streaming, as described in section VII. However, VNF applications still require some challenges and open issues to be addressed, as shown in Figure 10. This section discusses and classifies these challenges around some key features studied in relation to 5G networks and presents the open issues in the context of the 6G networks' roadmap.

A. Self-Organizing Networks

Agile deployment and life-cycle management of VNFs exploiting a NFV MANO architecture are essential features to satisfy the expectations of smart 5G networks, but further research is still ongoing to increase network automation. In this context, the Self-Organizing Network (SON) paradigm

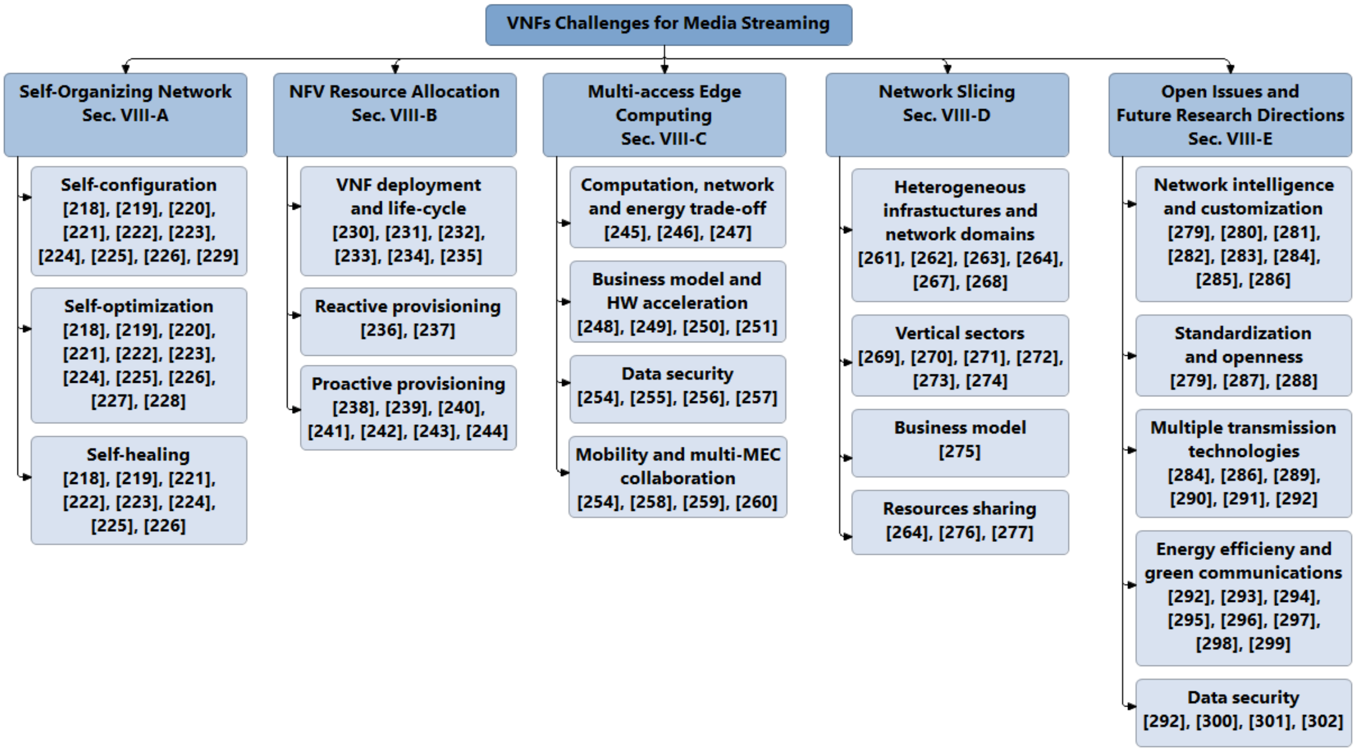

VNFs Challenges for Media Streaming
Self-Organizing Network Sec. VIII-A
Self-configuration [218], [219], [220], [221], [222], [223], [224], [225], [226], [229]
Self-optimization [218], [219], [220], [221], [222], [223], [224], [225], [226], [227], [228]
Self-healing [218], [219], [221], [222], [223], [224], [225], [226]
NFV Resource Allocation Sec. VIII-B
VNF deployment and life-cycle [230], [231], [232], [233], [234], [235]
Reactive provisioning [236], [237]
Proactive provisioning [238], [239], [240], [241], [242], [243], [244]
Multi-access Edge Computing Sec. VIII-C
Computation, network and energy trade-off [245], [246], [247]
Business model and HW acceleration [248], [249], [250], [251]
Data security [254], [255], [256], [257]
Mobility and multi-MEC collaboration [254], [258], [259], [260]
Network Slicing Sec. VIII-D
Heterogeneous infrastuctures and network domains [261], [262], [263], [264], [267], [268]
Vertical sectors [269], [270], [271], [272], [273], [274]
Business model [275]
Resources sharing [264], [276], [277]
Open Issues and Future Research Directions Sec. VIII-E
Network intelligence and customization [279], [280], [281], [282], [283], [284], [285], [286]
Standardization and openness [279], [287], [288]
Multiple transmission technologies [284], [286], [289], [290], [291], [292]
Energy efficieny and green communications [292], [293], [294], [295], [296], [297], [298], [299]
Data security [292], [300], [301], [302]

Fig. 10. Virtual Network Functions Challenges for Media Streaming.

TABLE XIV
SON CATEGORIES AND USE CASES.

Self-configuration	Self-optimization	Self-healing
• IP address & connectivity • neighbour & context discovery • radio access parameters • policy management	• load balancing • resource selection • caching infrastructure • coverage & capacity • radio interference management • mobility & handover	• fault detection • fault classification • countermeasures operations

[218] represents a next step to achieve a fully virtualized and automated network, as it empowers the network with specialized decision-making algorithms which monitor network resources and traffic patterns, and autonomously take actions to enforce or optimize network operations [219]. SON capabilities were initially meant to be included as add on features of LTE, as 3GPP Release 8 started defining LTE and already set the basis for SON concepts and requirements [220]. However, SON is expected to enhance 5G network management providing automation to cope with increasing network complexity [221].

Specifically, in the media streaming context, SON should provide the required network resources and guarantee target QoS or QoE scores when delivering media streams. More generally, SON turns static networks into dynamic ones by configuring network parameters, optimizing the allocated resources and fixing or preventing issues in real time.

A SON-enabled system can accomplish tasks belonging to three categories: self-configuration, self-optimization and self-healing [219]. Self-configuration techniques adjust network operational parameters to change network behavior and rules, according to specific business policies and node neighborhood context. Self-optimization strategies are dynamically applied to ensure that the network performance is near optimal. They include real-time network monitoring and performance metrics processing to proactively apply enhancement operational parameters. Self-optimization techniques can be applied in many areas: load balancing, resource selection, caching infrastructure, coverage and capacity, radio interference management, mobility and handover. Last, self-healing is necessary to generate a prompt reaction when faults, failures or any operational range violations in the network occur. The objective is to continuously monitor the system and ensure a fast and seamless recovery, whatever reason causes the failure. In case of a failure event, self-healing functions detect (fault detection) and diagnose (fault classification) it. Then, according to applicable policies and current setup, the appropriate countermeasure are applied to reestablish the desired network performance.

All these SON flavours need actionable data to process decision making algorithms. It is therefore very important to collect and exploit network data. Current networks are ready to probe and provide a huge amount of data. However, it is clear that specialized intelligence needs to be deployed within the network to infer valuable and useful information from the collected data [222]. Such information helps taking automatic actions to reach, recover or even improve the network performance. In the context of media streaming, it means that the SON paradigm has the potential to increase the QoS/QoE, while decreasing the business costs and energy consumption to maintain the network. In this context, the use of ML techniques will become prominent, even if the selection of the right algorithm is not trivial and depends on the considered use case [223], [221]. Table XIV shows the most common use cases belonging to the three SON categories, as seen from the network operator's perspective. Some SON applications are already provided by network vendors included in their commercial hardware equipment. Some examples are HCL's SON [224], Nokia's EdenNet [225] and Ericsson's SON Optimization Manager [226].

In any case, SON systems need to have a wider view of the delivered traffic beyond the metrics from the network functions and including service domain. It means that operated SON policies are usually steered by network statistics rather than application characteristics. Nevertheless, the communication dynamics of applications delivered on top of the network have an impact on network performance. Thus, the authors of [227] propose to design an application-driven SON in order to widen the view with both network performance and user's QoE metrics. When considering media streaming applications, data are available from network functions in the path and from playback devices. Thus, data exploitation inside a SON-enabled system needs further investigation, as the multi-domain data exploitation is still underexplored. Few solutions are available in the literature that apply a SON paradigm to media streaming scenarios. The authors of [228] propose a SON-enabled media transcoder to be deployed within the network. In [229] the authors introduce a self-organizing Unmanned Aerial Vehicle (UAV)-based communication framework for media streaming.

B. NFV Resource Allocation

The deployment of a VNF over a distributed platform requires the allocation of network and computing assets to be provisioned to host the VNF. Network and computing resource allocation is a challenging feature whose interest is raising and focusing on VNFs deployment and life-cycle management [230]. In this context, the NVFO is in charge of selecting the appropriate resources, among the available ones at the NVFI, when deploying a VNF, which is usually referred to as the NFV resource allocation (NFV-RA) problem. The NFV-RA includes three stages [230]: VNF Chain Composition (VNF-CC), VNF Forwarding Graph embedding (VNF-FGE) and VNF Scheduling (VNF-SCH).

VNF-CC deals with the composition of several VNFs to be deployed jointly by the NFVO. How the traffic flows between VNFs is also described trough the definition of VNF Forwarding Graphs (VNF-FG). Thus, any network service can be considered as composed of a set of VNFs and VNF-FGs. Each VNF executes a small function of the entire application or service [231]. VNF-FGE focuses on how to embed the VNFs and VNF-FGs into the infrastructure. It aims to find suitable resources and locations where to allocate the VNFs in NFVI. At this stage, resource selection and optimization must be accomplished with regard to the specific constraints defined by SLA [232]. Finally, VNF-SCH determines how to schedule the processing operations of the deployed VNFs [233].

When the VNFs are already deployed and running, the required resources vary during their life-cycle, as they depend on user demand of the running function provided by the VNFs. Allocated resources could be optimized to fit with the variable demand by the user. Increasing or decreasing the allocated resources means the VNFs also need to dynamically scale up/down. Then, an efficient orchestration and automation of the VNFs requires supporting this dynamic allocation of resources. This assumption was already envisioned when designing the NFV MANO architecture [234], where mechanisms to scale are essential to enable a flexible management of the running services.

However, the easiest and fastest approach consists of employing an over-provisioning strategy, where the amount of allocated resources for each VNF is larger than what is required. In case of experiencing an increasing demand, the VNF can manage overheads without any intervention as long as the allocated resources are not exceeded. This approach is operationally effective, but inefficient in terms of OPEX and energy consumption generated by the allocated resources which are not actually employed. This means that this approach is not cost-effective, as it is clear that adjusting the resources allocated for the VNF to the actual demand would avoid over-provisioning and reduce costs. Employment of dynamic provisioning strategies results in OPEX reductions for network operators and/or service providers [14].

Enabling dynamic resource allocation for VNFs allows scaling up and down and therefore coping with network traffic fluctuations and changeable demands from connected users. Dynamism, scalability and automation are important features for resource management [235]. Changes in resource allocation should be applied according to real-time network traffic and service demands. Dynamic resource allocation can be performed in reactive or proactive manners. Simple solutions involve a reactive provisioning approach which means changing the allocated resources to react when traffic and/or demand change. In [236] the authors design an online algorithm for VNF scaling in cloud data centers. The authors of [237] aim to minimize the OPEX by considering the trade-off between bandwidth and host resource consumption under diverse workload variations. All these reactive solutions have as an advantage a simple design, as there is no need for any complex algorithms for provisioning. On the other side, such an approach does not prevent any network issues or service faults from happening, affecting the services.

A more sophisticated approach consists of proactive provisioning where the future traffic and/or demand is predicted. Being able to foresee the amount of resources to be allocated constitutes a great benefit, as it enables to avoid network issues or service faults by proactively resizing the employed resources and scaling the deployed VNFs. In such a context, the problem of service demand prediction constitutes a mayor challenge. Most of the literature on demand prediction employs ANN algorithms [238], [239]. However, the application of such algorithms in practical solutions is limited, being mostly theoretical.

Among the most innovative solutions proposed, [240] describes a novel FTRL online algorithm for VNF provisioning which handles workload fluctuations. The solution in [241] employs an ANN algorithm to predict future resource requirements for each VNF contributing to a network service. The authors of [242] propose the POLAR algorithm, which combines online learning and online optimization of proactive provision resources with VNFs provisioning, while the VNFs chaining in a network service is ignored. In [243] a proactive failure recovery is proposed when considering VNF deployed at distributed edge computing nodes. In [11] a proactive VNF chaining aims to find the optimal number of VNFs and their location inside a CDN in order to minimize costs. Finally, the authors of [244] propose a multi-layer resource allocation solution, which aims to proactively provide resources to the VNFs deployed in several VIMs and network resources between VIMs.

C. Multi-access Edge Computing (MEC)

MEC represents a novel technological solution integrated in 5G networks to bring computation closer to the user. MEC infrastructures create new potential revenue flows to network operators opening their edge infrastructures to host specialized services at network edge. There are many aspects which require investigation to achieve a complete integration of MEC into the current network architecture and services. However some avenues are already seen as highly beneficial for MEC deployment and use. For instance, media streaming is a key application of MEC solutions, as ETSI considers it as one of MEC core use cases [182]. MEC platforms can host edge services to empower media streaming applications, which traditionally were based on server-client communications. As explained in the previous section, MEC and VNFs enable the deployment of innovative media-related services such as media casting, media transcoding and content caching.

More specifically, MEC resources are exploited by both the server and clients to offload computation tasks [245], [246]. Offloading server tasks targets reducing network traffic and latency, as the processing is performed close to UEs. MEC resources are shared between different service providers, but how the resources are distributed among different service providers is still undefined. The authors of [245] propose to allocate MEC resources proportionally to the demanded resources and payment of each service provider. If an UE offloads tasks to the MEC host, it reduces not only the device computation load, but also its power consumption, as computing-intensive tasks heavily impact on the battery duration. In [246], a video telephony application employs MEC to encode the content. It reduces processing operations at the UE, but increases network traffic since uncompressed raw content is sent to the base station. The authors focus on power consumption, but they do not consider operational costs generated by using the MEC platform. In general, how to balance network traffic, power consumption and operational costs trade-off needs to be studied. In [247], optimization of the allocation of both computing and network resources is discussed, while taking into account the energy efficiency. Even in this case, operational costs are not considered in the optimization problem. In general, business aspects raise complex discussions due to the lack of a clear business model [248]. MEC needs a business model equivalent to the one applicable in cloud computing infrastructures. However, unlike cloud computing, the decentralized location and utilization of shared resources between services makes the cost model more complex. Resource accounting and monitoring have to also be determined in order to create a complete business model. The debate on the business model is even more intricate if we consider hardware-acceleration assets, such as GPUs, required to accomplish critical tasks where general-purpose hardware (CPU) has limitations [249]. Some works suggest to employ Field-Programmable Gate Array (FPGA) approaches instead of GPU solutions due to their reduced price and power consumption [250], [251], but this possibility is again underexplored.

Regarding to accessible information at MEC, the API to communicate with RNIS [181] has been recently standardized and its development is on going [252], [253]. It means that services running at the MEC host cannot be further optimized. When RNIS implementations will be available, edge services could embed more complex and precise algorithms (classic or ML models), aiming to exploit RNI in order to improve their operations and the performance of the overall system. However, improved capabilities due to RNI exploitation raise some security concerns on how to manage information at MEC hosts, an aspect that needs further investigation [254], [255]. In order to exploit a MEC decentralized approach, the deployment of location-aware services is necessary. Thus, mechanisms for user privacy protection and anonymity are needed. Moreover, modification of the networks to introduce MEC capabilities opens the door for potential attacks, including DDoS attacks, malware injection, authentication and authorization attacks [256], [257].

Mobility remains another major concern and is becoming critical, as the explosion in availability and type of mobile devices (e.g., smartphone and tablets) involves an increasing number of UEs to be served. The same way the connectivity is guaranteed when moving from a cell to another in a cellular network, migration support for MEC services is also required. Consequently, the investigation on a multi-MEC cooperation should be addressed in order to guarantee seamless migration of sessions across MEC servers [254], [258].

From the perspective of media services, user QoE plays an important role and a wide MEC deployment definitely should target it, especially as transcoding and caching capabilities would be provided closer to UEs. How to balance the cost of MEC-based caching and transcoding and provision of high user QoE is an important direction for future research [254]. Moreover, it becomes relevant the ability to find suitable locations where MEC instances should be deployed, as it may affect the fulfilment of the demanded requirements. It is especially true for low latency multimedia services, where the distance between the MEC host and UE affects the overall delay [259]. Finally, content caching mechanisms in the network have been studied both at the core and at the edge, but a convergent solution has not identified yet. Caching solutions that integrate both core and edge caching could result in better network performance in terms of the energy consumption, network throughput, latency, and user QoE [260].

D. Network Slicing

While NFV-RA is limited to provide NFVI resources deployed for a specific section of the network (CN or RAN/MEC), network slice has a wider scope, as it is able to provide network and computing resources even across different networks. Network slicing [261], [262] is introduced in 5G networks as a solution involving several virtual/logical networks (slices) on top of a common physical network, where each virtual/logical network delivers the traffic generated by a specific service [263], [264]. It can be considered that a network slice is associated with a set of network resources and VNFs, which can be provided by that slice. In this context, NFV MANO and SDN play an important role, especially in the deployment and management of network slices [265], [266]. NFV MANO enables life cycle management and orchestration of the VNFs, while SDN allows for the configuration and control of the routing and forwarding planes of the underlying network infrastructure, providing communication between the deployed VNFs. This results in a logical network of resources and VNFs built over a common underlying physical infrastructure, separated into diverse network slices. Each network slice provides the service as an end-to-end connectivity, meaning that network slicing provisioning refers to three different aspects: at the air interface, in the RAN and in the CN [267], [268].

Network slicing at the air interface refers to partitioning physical radio resources (physical layer or L1) into subsets of several physical resources, each one for a different network slice, then mapping into logical resources to be provided to the Medium Access Control (MAC) sublayer at the datalink layer (or L2) and higher layers.

In the RAN, network slicing changes RAN operations, including MEC-operated ones, such as device association and access control, from a cell-specific perspective to a slice-specific one. Thus, the RAN operations are service-oriented instead of physical cell-oriented. Configuration of control and user planes is tailored and/or tuned considering the requirements of each slice individually. Then, factors such as QoS requirements, traffic load or type of service/traffic are prominent when operating the RAN.

Finally, network slicing in the CN enables the definition of vertical networks, where each one aims to support a service belonging to a specific vertical industry. NFV MANO and SDN have a higher impact in this aspect of the network, where each vertical industry should be able to run its VNF-specific solutions. CN needs flexible management to enable resource scalability and migration when required by the network traffic associated with a service.

A videoconferencing system is deployed in [269] through the deployment of two different slices to split audio and video transmissions, as they have different requirements in terms of network throughput. In [270], the authors focus on the eHealth vertical, where services are typically media-rich and mission-critical and are high QoS demanding. Then, a MEC-based application, empowered with end-to-end network slicing, is designed and developed to enable in-ambulance applications. The application is accessed by paramedics in the ambulance and sends audiovisual data to the hospital/doctor. The same vertical is addressed by [271] to enable a real-time communication between hospital staff and patients. In [272] and [273], applications of network slicing for Vehicle-to-Everything (V2X) services are investigated. Different uses cases are considered in a vehicle, including related to safety and traffic efficiency, autonomous or tele-operated driving, media & entertainment and remote diagnostics. Each use case means different requirements in terms of latency, throughput and communication reliability. Consequently, different network slices with different configurations are required on top of the same physical network of resources. The authors of [274] present several use cases belonging to different verticals, such as protection and smart metering in the smart grid sector, car and passenger data exchange in an intelligent transportation system and best-effort data delivery in a multimedia system. Each use case and vertical sector requires different capabilities in terms of latency and throughput. The different types of traffic are prioritized by splitting them into specialized network slices.

Network slicing-related research has increased importance in the current 5G network context. Ongoing challenges include solutions to allow wide employment and operation of slices for different industry verticals. Most of slicing operations relate to the exploitation of resources provided by the network operator, but the effects of changes in network operator's business models for operating network slicing are unknown [275]. The increase in the number of devices belonging to different verticals and their mobility management in the presence of different technologies (LTE, 5G, Wi-Fi) also need further investigation [242]. An end-to-end network slice implies that slice segments potentially stretch across different administrative domains. There are two requirements in order to achieve a unified control of the network slice. First, an exchange point that performs the resource negotiation between different administrative domains is necessary to enable multi-domain slices. Then, standardized APIs should make transparent the underlying domains and simplify the negotiations to provide the control on the slice [262]. Finally, network slicing leverages algorithms to accommodate applications with widely diverse requirements over the same physical network. Thus, complex algorithms are necessary for deciding how to efficiently allocate, manage, and control the physical resources to be shared across diverse slices [276]. Concerning these algorithms, the application of ML in network systems is capturing increased research attention lately and this trend is expected to continue in the future [277].

E. Open Issues and Future Research Directions

The benefits of virtualization for media streaming communications will increasingly evident in the next few years, as the 5G coverage will be extended. Complementary technologies such as MEC, SON and network slicing are still not fully integrated. Further efforts in integrating all these new paradigms and/or architectures are envisioned to provide a more efficient and intelligent network [278].

ML-powered network intelligence to manage NFV and VNFs is only partially achieved in 5G networks, but it will be also a key factor for the future 6G networks [279]. The concept of Intent-Based Networks (IBN) [280] means employing ML solutions to transform business intents into network configuration, operation, and maintenance strategies. In order to meet the massive service demands and overcome limitations due to time-varying network traffic, the network can continuously learn and adapt to the time-varying network environment based on the massive collected network data in real-time. An intelligent-native network exploits ML algorithms to improve its capabilities and reduce the business costs for service deployment and management [281], [282]. The advantages of an intelligent-native network are two-fold. First, the network can analyze user's behavior in real-time and autonomously learn his needs to predict its future behavior. Then, user's information can be employed for network customization to achieve a user-centric network [283]. Second, the network can met changing requirements of a network service during its life-cycle by autonomously matching the requirements to the corresponding network communication, computing and caching assets. This is also valid for new emerging services. Holographic (AR and VR) and haptic communications are meant to be wider available thanks to the future 6G network [284]. Moreover, the global COVID-19 pandemic is accelerating the digital transformation of multiple and heterogeneous verticals, such as development of new services for smart cities and innovation in the eHealth including telemedicine, medical and thermal imaging, and robotics for medicine practice [285], [286].

Openness is also an important aspect to achieve flexible network and services [279]. Having open network platform and interfaces (O-RAN, NFV MANO, SDN, etc.) allows interconnection and interoperability of different vendors, which is essential for sharing a physical infrastructure. Thus, agents of diverse vertical industries may deploy their private physical infrastructure and manage it though NFV MANO solutions and SDN controllers independent from public networks operated by mobile network operators [287]. Standardization process will continue in the next years to fulfil the remaining gaps and guarantee interoperability of heterogeneous implementations of open network solutions [288].

The cooperation of different physical networks will also attract attention. Multiple Radio Access Technology (multi-RAT) aims to employ different access network to improve the overall connectivity [289]. Its application to improve media streaming is already being investigated [290], [291], but new transmission solutions based on space, UAV-based and underwater communications will be integrated with terrestrial ones [284], [286]. Flexibility to operate the network at any level (spectrum/band, physical and MAC, etc.), despite the different involved technologies, will be imperative [292].

Energy efficiency and green communications [293] are envisioned to enable more sustainable networking [294]. Energy efficiency concerns are also relevant for media streaming services [295], [296]. Here, low-power wireless devices could harvest energy from the available high-power radio waves [292]. Thus, battery-free implementations will be an interesting topic to be further explored in different use cases, e.g., IoT [297] and media streaming communications [298], [299].

Finally, the growth of network and media traffic will have consequences for security. Critical media use cases, e.g., eHealth applications [300] and autonomous driving systems [301], need to be secured with security mechanisms which will complement the conventional cryptography-based ones. Increasing security will be assured with the design of cross-layer algorithms to protect the transferred information [292], [302].

IX. International Initiatives

Employing VNFs for media streaming is a research topic that has attracted the attention of international organizations and international funding programs for many years now. Recently, the European Commission has funded numerous research projects aiming at developing and implementing VNFs for different research scenarios and vertical industries. Table XV summarizes the most relevant actions. The project list includes initiatives targeting generic architectural design (i.e., CogNET [303], SELFNET [306] and SliceNet [306]), activities building testbed environments and pilot environments for use case definition and testing (FLAME [312], SoftFIRE [310] and 5GTango) [314], projects targeting specific application verticals and developing required functionalities (5G-Media [317], 5Growth [319], 5GCity [321]) and finally international software communities to provide open-source platforms (OpenAirInterface Software Alliance [323], Mosaic5G [324] and O-RAN Alliance [326]).

Regarding architectural definition, SELFNET H2020 project designed and tested an autonomous network management framework capable of the automatic detection and mitigation of common failures in the network [307]. Among others, it proposed the smart integration of state-of-the-art technologies in NFV. One of the outcomes is presented in [327], where the SELFNET framework preserves the health of the network maximizing the QoE and minimizing the end-to-end energy consumption. SliceNet project addressed both management and control planes of network slicing to leverage QoS for sliced services [328]. The project proposed an integrated network management, control and orchestration framework and applied the concept to a variety of use cases. One of those cases, related to multimedia health services is described in [270], where demanding QoS requirements (i.e., latency) need to be fulfilled. The network intelligence topic is tackled by CogNET, a project that focused on realizing the well-known control loop MAPE (Monitor, Analyze, Plan and Execute) with Machine Learning techniques and policy-based mechanisms for a vision of softwarized 5G networks. COGNET validated its vision in different use cases that include SLA Enforcement and Mobile Quality Predictors [305], [304].

A second group of projects aimed at creating platforms and testbed environments where specific use cases, applications,

TABLE XV
MAJOR SDN/NFV RELATED RESEARCH ACTIVITIES.

Project	Time period	Area of concern	References
CogNet (Building an Intelligent System of Insights and Action for 5G Network Management)	2015-2018	Architecture	[303], [304], [305]
SELFNET (Framework for Self-Organized Network Management in Virtualized and Software Defined Networks)	2015-2018	Architecture	[306], [307]
SliceNet (End-to-End Cognitive Network Slicing and Slice Management Framework in Virtualised Multi-Domain, Multi-Tenant 5G Networks)	2017-2020	Architecture	[308], [270], [309]
SoftFIRE (Software Defined Networks and Network Function Virtualization Testbed within FIRE+)	2016-2018	TestBeds	[310], [311]
FLAME (Facility for Large-scale Adaptive Media Experimentation)	2017-2020	TestBeds	[312], [313]
5GTango (5G Development and validation platform for global industry-specific network services and Apps)	2017-2020	TestBeds	[314], [315], [316]
5G-Media (Programmable edge-to-cloud virtualization fabric for the 5G Media industry)	2017-2020	Application Verticals	[317], [58], [318]
5Growth (5G-enabled Growth in Vertical Industries)	2019-2021	Application Verticals	[319], [320]
5GCity (A Distributed Cloud and Radio Platform for 5G Neutral Hosts)	2017-2020	Application Verticals	[321], [322]
OpenAirInterface Software Alliance	2014-	Development Platforms	[323], [174]
Mosaic5G	2016 -	Development Platforms	[324], [325]
O-RAN Alliance	2018-	Development Platforms	[323], [179]

algorithms, and interoperability solutions could be designed and validated. FLAME stands out in this area as a facility for experimenting large scale experiments in the field of Adaptive Media. Since 2017, FLAME has hosted different proposals [313] to offload proactively video content to the edge of the network on an SDN/NFV environment. FLAME tests include augmented reality applications as well as smart video surveillance for aiding impaired citizens. SoftFIRE is another testbed environment to experiment VNF services and applications in SDN/NFV. SoftFIRE aims at assessing the level of maturity of solutions in programmability, interoperability and security and showing how they can support the full potential of these properties in a real-world case [311]. Finally, 5GTango puts the focus on network flexible programmability [315] by providing software development kits (SDKs) [316]. This project included qualification and verification mechanisms as well as a modular service platform to bridge the gap between business needs and network operational management systems. 5GTango was demonstrated in two vertical through specific pilots: advanced manufacturing and immersive media [315].

The third category encompasses some examples of projects designing the required building blocks that enable the applications for specific vertical sectors. 5GCity was an H2020 project aiming at designing, implementing and demonstrating a distributed cloud and radio platform for municipalities and infrastructures with neutral hosting capabilities. One of the main outcomes of the project was the 5GCity Orchestration Platform, which supported the NFV MANO model. In [322], the authors demonstrate that the virtualized platform was able to address different use cases related to media streaming such as real-time video acquisition and production at the edge, UHD Video Distribution and immersive services or mobile real-time transmission. 5G-MEDIA [317] exploits the principles of NFV and SDN to facilitate the development, deployment, and operation of VNF-based media services on 5G networks. Key in this project is the development of a platform for service virtualization that provides an advanced cognitive management environment for the provisioning of network services and media applications [58]. The use cases include tele-immersive gaming, mobile journalism and UHD content distribution [318]. 5Growth [319] supports diverse industry verticals developing the tools for interfacing those verticals with the 5G end-to-end platforms. The system provides the creation of network slices with closed-loop automation and SLA life-cycle service control. ML-driven solutions are also part of the project targets to optimize access, transport, core and cloud, edge and fog resources, across multiple technologies and domains [320].

Finally, OpenAirInterface Software Alliance [323], Mosaic5G [324], and O-RAN Alliance [326] are mixed academic and industrial communities to create ecosystems of open-source projects for studying, building, and sustaining open flexible and integrated 5G network. OpenAirInterface Software Alliance [323] provides 5G network tools extensively used by researchers from both industry and academia. This initiative gathers developers from around the world, who work together to build wireless cellular RAN and CN technologies [174]. Mosaic5G [324] develops a set of 5G software solutions and has already hosted experiments targeting low latency MEC services, orchestration solutions and programmable RANs [325]. O-RAN Alliance [326] is pushing the standardization and the development of the O-RAN. RAN industry is moving towards open, intelligent, virtualized and fully interoperable RAN [179].

X. WHITECONCLUSIONS

The popularity of media streaming services is constantly growing due to increasing number of users and diversity of rich media applications, e.g., online gaming, VR/AR applications, etc. The latest smart mobile devices also have an important role in the success of media streaming, as their processing and rendering capabilities support streaming content at very high resolutions, e.g., Ultra-High-Definition (UHD) or 4K. Consequently, media streaming traffic accounts not only a very large share of the total Internet traffic, but, more importantly, also an increasing one.

To cope with this increasing media traffic and high dynamics of network performance and user mobility, improved network capabilities are required to maintain high QoS and QoE performance, while also achieving the best trade-off with business costs and energy efficiency. 5G networking is bringing new possibilities to deploy smart network functions, which monitor both the media streaming service through live and objective metrics and boost it in real time. Under the 5G umbrella, NFV and SDN will have a prominent role in the virtualization of network functions and their management and orchestration.

In this context, this work provided a state-of-the-art on VNFs applied to media streaming. To this end, we considered the factors that concur to the design and implementation of a stable VNF. Monitoring and collecting performance metrics enable their exploitation as source of information for the VNF life-cycle deployment and management, as well as to evaluate the effects of the capabilities provided by the VNF on the media streaming session. Moreover, network traffic monitoring and analysis allow to create models to approximate the behavior of the network and predict future network events to take actions in a proactive manner. Thus, any network malfunction or issue that affects the media steaming session can be prevented.

Several VNF solutions to improve media streaming are presented. Solutions including media casting, media transcoding and content caching can be employed at any segment of the network. Thanks to the NFV MANO architecture, the deployment of VNFs is not limited to the Network Core, but they can be also run at MEC infrastructures. Capillarity of the MEC allows computing operations close to the base stations and reduces the latency when dealing with live streaming services.

Finally, research challenges and open issues have been presented in the realm of VNFs applied to media streaming services. The achievement of dynamic resource allocation, complete MEC integration and network slicing are the main venues where the research will focus in the next few years. Long-term research directions will also address a strong employment of ML to foster network capabilities and the utilization of open network solutions and/or new access technologies, also combining them to increase the capacity. Green communications and security will also be major concerns, as the future networks should reduce their impact on the environment and guarantee the security of the processed information. In conclusion, VNFs represent an important enabler to improve the media streaming services, but despite the research done under international initiatives that are pushing 5G and network virtualization, several research challenges still exist and provide opportunities for further research activities.

Acknowledgment

This work was fully supported by Open-VERSO project (Red Cervera programme, Spanish government's Centre for the Development of Industrial Technology) and by the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 870610 for the TRACTION project. The support of the Science Foundation Ireland (SFI) Research Centres Programme Grant Numbers 12/RC/2289_P2 (Insight) and 16/SP/3804 (ENABLE) is also gratefully acknowledged.

References

[1] Cisco. (2020) Cisco annual internet report (2018–2023) white paper. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[2] A. Lutu, D. Perino, M. Bagnulo, E. Frias-Martinez, and J. Khangosstar, "A characterization of the covid-19 pandemic impact on a mobile network operator traffic," in *Proceedings of the ACM Internet Measurement Conference*, 2020, pp. 19–33.

[3] A. Feldmann, O. Gasser, F. Lichtblau, E. Pujol, I. Poese, C. Dietzel, D. Wagner, M. Wichtlhuber, J. Tapiador, N. Vallina-Rodriguez, O. Hohlfeld, and G. Smaragdakis, "The lockdown effect: Implications of the covid-19 pandemic on internet traffic," *Proceedings of the ACM Internet Measurement Conference*, p. 1–18, 2020.

[4] T. Favale, F. Soro, M. Trevisan, I. Drago, and M. Mellia, "Campus traffic and e-learning during covid-19 pandemic," *Computer Networks*, vol. 176, p. 107290, 2020.

[5] D. L. King, P. H. Delfabbro, J. Billieux, and M. N. Potenza, "Problematic online gaming and the covid-19 pandemic," *Journal of Behavioral Addictions*, vol. 9, no. 2, pp. 184–186, 2020.

[6] Broadband commission agenda for action for faster and better recovery. [Online]. Available: https://www.broadbandcommission.org/COVID19/Pages/default.aspx

[7] The affordability of ict services 2020. [Online]. Available: https://www.itu.int/en/ITU-D/Statistics/Documents/publications/prices2020/ITU_A4AI_Price_Briefing_2020.pdf

[8] D. Rayburn. (2020) Cdn/media pricing see's big drop for largest customers: Pricing down to $0.0006. [Online]. Available: https://www.streamingmediablog.com/2020/05/q1-cdn-pricing.html

[9] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90–97, 2015.

[10] N. T. Jahromi, S. Kianpisheh, and R. H. Glitho, "Online vnf placement and chaining for value-added services in content delivery networks," in *2018 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN)*. IEEE, 2018, pp. 19–24.

[11] M. Dieye, S. Ahvar, J. Sahoo, E. Ahvar, R. Glitho, H. Elbiaze, and N. Crespi, "Cpvnf: Cost-efficient proactive vnf placement and chaining for value-added services in content delivery networks," *IEEE Transactions on Network and Service Management*, vol. 15, no. 2, pp. 774–786, 2018.

[12] H. J. Kim and S. G. Choi, "A study on a qos/qoe correlation model for qoe evaluation on iptv service," in *2010 The 12th International Conference on Advanced Communication Technology (ICACT)*, vol. 2. IEEE, 2010, pp. 1377–1382.

[13] M. Alreshoodi and J. Woods, "Survey on qoe\qos correlation models for multimedia services," *arXiv preprint arXiv:1306.0221*, 2013.

[14] E. Hernandez-Valencia, S. Izzo, and B. Polonsky, "How will nfv/sdn transform service provider opex?" *IEEE Network*, vol. 29, no. 3, pp. 60–67, 2015.

[15] A. Adas, "Traffic models in broadband networks," *IEEE communications Magazine*, vol. 35, no.7, pp. 82–89, 1997.

[16] D. Chalmers and M. Sloman, "A survey of quality of service in mobile computing environments," *IEEE Communications surveys*, vol. 2, no. 2, pp. 2–10, 1999.

[17] J. Jin and K. Nahrstedt, "Qos specification languages for distributed multimedia applications: A survey and taxonomy," *IEEE multimedia*, vol. 11, no. 3, pp. 74–87, 2004.

[18] H. Feng and Y. Shu, "Study on network traffic prediction techniques," *Proceedings. 2005 International Conference on Wireless Communications, Networking and Mobile Computing*, vol. 2, pp. 1041–1044, IEEE, 2005.

[19] B. Chandrasekaran, "Survey of network traffic models," *Waschington University in St. Louis CSE*, 2009.

[20] A. M. Mohammed and A. F. Agamy, "A survey on the common network traffic sources models," *International Journal of Computer Networks (IJCN)*, vol. 3, no. 2, pp. 103–115, 2011.

[21] M. A. Hoque, M. Siekkinen, and J. K. Nurminen, "Energy efficient multimedia streaming to mobile devices—a survey," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 579–597, 2012.

[22] S. Baraković and L. Skorin-Kapov, "Survey and challenges of qoe management issues in wireless networks," *Journal of Computer Networks and Communications*, vol. 2013, 2013.

[23] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A survey on quality of experience of http adaptive streaming," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 469–492, 2014.

[24] P. Juluri, V. Tamarapalli, and D. Medhi, "Measurement of quality of experience of video-on-demand services: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 401–418, 2015.

[25] G.-M. Su, X. Su, Y. Bai, M. Wang, A. V. Vasilakos, and H. Wang, "Qoe in video streaming over wireless networks: perspectives and research challenges," *Wireless networks*, vol. 22, no. 5, pp. 1571–1593, 2016.

[26] T. Zhao, Q. Liu, and C. W. Chen, "Qoe in video transmission: A user experience-driven strategy," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 1, pp. 285–302, 2016.

[27] Z. Akhtar and T. H. Falk, "Audio-visual multimedia quality assessment: A comprehensive survey," *IEEE access*, vol. 5, pp. 21 090–21 117, 2017.

[28] S. Petrangeli, J. V. D. Hooft, T. Wauters, and F. D. Turck, "Quality of experience-centric management of adaptive video streaming services: Status and challenges," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 2s, pp. 1–29, 2018.

[29] L. Skorin-Kapov, M. Varela, T. Hoßfeld, and K.-T. Chen, "A survey of emerging concepts and challenges for qoe management of multimedia services," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 2s, pp. 1–29, 2018.

[30] A. A. Barakabitze, N. Barman, A. Ahmad, S. Zadtootaghaj, L. Sun, M. G. Martini, and L. Atzori, "Qoe management of multimedia streaming services in future networks: a tutorial and survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 526–565, 2019.

[31] N. Barman and M. G. Martini, "Qoe modeling for http adaptive video streaming–a survey and open challenges," *Ieee Access*, vol. 7, pp. 30 831–30 859, 2019.

[32] C. Zhang, H. P. Joshi, G. F. Riley, and S. A. Wright, "Towards a virtual network function research agenda: A systematic literature review of vnf design considerations," *Journal of Network and Computer Applications*, vol. 146. p. 102417, 2019.

[33] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A survey on 5g usage scenarios and traffic models," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 905–929, IEEE, 2020.

[34] H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson *et al.*, "Rtp: A transport protocol for real-time applications," *rfc 1889, January*, 1996.

[35] J. Postel *et al.*, "User datagram protocol," *STD 6, RFC 768, August*, 1980.

[36] ——, "Transmission control protocol," *STD 7, RFC 793, September*, 1981.

[37] L. Popa, A. Ghodsi, and I. Stoica, "Http as the narrow waist of the future internet," *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, pp. 1–6, 2010.

[38] H. Schulzrinne, A. Rao, and R. Lanphier, "Real time streaming protocol (rtsp)," *rfc 2326, April*, 1998.

[39] M. Thornburgh, "Adobe's rtmfp profile for flash communication," *Internet Engineering Task Force (IETF)*, 2014.

[40] M. P. Sharabayko *et al.*, "The srt protocol," *draft-sharabayko-srt-00*, 2021.

[41] C. Holmberg, S. Hakansson, and G. Eriksson, "Web real-time communication use cases and requirements," *Request for Comments (RFC)*, vol. 7478, 2015.

[42] D. Wing, P. Matthews, R. Mahy, and J. Rosenberg, "Session traversal utilities for nat (stun)," *RFC5389, October*, 2008.

[43] R. Mahy, P. Matthews, and J. Rosenberg, "Traversal using relays around nat (turn): Relay extensions to session traversal utilities for nat (stun)," RFC 5766 (Proposed Standard), Internet Engineering Task Force, Tech. Rep., 2010.

[44] R. Pantos and W. May, "Http live streaming," *rfc 8216, August*, 2017.

[45] I. Sodagar, "The mpeg-dash standard for multimedia streaming over the internet," *IEEE multimedia*, vol. 18, no. 4, pp. 62–67, 2011.

[46] K. Hughes and D. Singer, "Information technology–multimedia application format (mpeg-a)–part 19: Common media application format (cmaf) for segmented media," *ISO/IEC*, vol. 19, p. 23000, 2017.

[47] K. Durak, M. N. Akcay, Y. K. Erinc, B. Pekel, and A. C. Begen, "Evaluating the performance of apple's low-latency hls," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2020, pp. 1–6.

[48] N. Bouzakaria, C. Concolato, and J. Le Feuvre, "Overhead and performance of low latency live streaming using mpeg-dash," in *IISA 2014, The 5th International Conference on Information, Intelligence, Systems and Applications*. IEEE, 2014, pp. 92–97.

[49] P. Chown, "Advanced encryption standard (aes) ciphersuites for transport layer security (tls)," RFC 3268, June, Tech. Rep., 2002.

[50] D. Bhat, A. Rizk, and M. Zink, "Not so quic: A performance study of dash over quic," *Proceedings of the 27th workshop on network and operating systems support for digital audio and video*, pp. 13–18, 2017.

[51] W3C. (2020) Quic api for peer-to-peer connections. [Online]. Available: https://w3c.github.io/webrtc-quic/

[52] A. Langley, A. Riddoch, A. Wilk, A. Vicente, C. Krasic, D. Zhang, F. Yang, F. Kouranov, I. Swett, J. Iyengar *et al.*, "The quic transport protocol: Design and internet-scale deployment," in *Proceedings of the conference of the ACM special interest group on data communication*, 2017, pp. 183–196.

[53] L. Ong, J. Yoakum *et al.*, "An introduction to the stream control transmission protocol (sctp)," RFC 3286 (Informational), May, Tech. Rep., 2002.

[54] A. Ford, C. Raiciu, M. Handley, S. Barre, J. Iyengar *et al.*, "Architectural guidelines for multipath tcp development," *IETF, Informational RFC*, vol. 6182, pp. 2070–1721, 2011.

[55] Q. De Coninck and O. Bonaventure, "Multipath extensions for quic (mp-quic)," *draft-deconinck-quic-multipath-06*, 2020.

[56] K. Evensen, T. Kupka, H. Riiser, P. Ni, R. Eg, C. Griwodz, and P. Halvorsen, "Adaptive media streaming to mobile devices: challenges, enhancements, and recommendations," *Advances in Multimedia*, 2014.

[57] M. Keltsch, S. Prokesch, O. P. Gordo, J. Serrano, T. K. Phan, and I. Fritzsch, "Remote production and mobile contribution over 5g networks: scenarios, requirements and approaches for broadcast quality media streaming," *2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–7, 2018.

[58] F. Alvarez, D. Breitgand, D. Griffin, P. Andriani, S. Rizou, N. Zioulis, F. Moscatelli, J. Serrano, M. Keltsch, P. Trakadas, T. K. Phan, A. Weit, U. Acar, O. Prieto, F. Iadanza, G. Carrozzo, H. Koumaras, D. Zarpalas, and D. Jimenez, "An Edge-to-Cloud Virtualized Multimedia Service Platform for 5G Networks," *IEEE Transactions on Broadcasting*, vol. 65, no. 2, pp. 369–380, Jun. 2019.

[59] C. Zhang, H. P. Joshi, G. F. Riley, and S. A. Wright, "Towards a virtual network function research agenda: A systematic literature review of vnf design considerations," *Journal of Network and Computer Applications*, vol. 146, p. 102417, 2019.

[60] ITU. (2016) Recommendation itu-t p.800.1: Mean opinion score terminology. [Online]. Available: https://www.itu.int/rec/T-REC-P.800.1

[61] ——. (2016) Recommendation itu-t p.800.2: Mean opinion score interpretation and reporting. [Online]. Available: https://www.itu.int/rec/T-REC-P.800.2

[62] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[63] A. Rehman, K. Zeng, and Z. Wang, "Display device-adapted video quality-of-experience assessment," *Human Vision and Electronic Imaging XX*, vol. 9394, p. 939406, International Society for Optics and Photonics, 2015.

[64] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. (2016) Toward a practical perceptual video quality metric. [Online]. Available: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652

[65] A. Aaron, Z. Li, M. Manohara, J. De Cock, and D. Ronca. (2015) Per-title encode optimization. [Online]. Available: https://netflixtechblog.com/per-title-encode-optimization-7e99442b62a2

[66] J. De Vriendt, D. De Vleeschauwer, and D. Robinson, "Model for estimating qoe of video delivered using http adaptive streaming," *2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013)*, pp. 1288–1293, 2013.

[67] X. Yin, V. Sekar, and B. Sinopoli, "Toward a principled framework to design dynamic adaptive streaming algorithms over http," *Proceedings of the 13th ACM Workshop on Hot Topics in Networks*, pp. 1–7, 2014.

[68] J. Xue, D. Q. Zhang, H. Yu, and C. W. Chen, "Assessing quality of experience for adaptive http video streaming," *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 1–6, IEEE, 2014.

[69] Y. Liu, S. Dey, F. Ulupinar, M. Luby, and Y. Mao, "Deriving and validating user experience model for dash video streaming," *IEEE Transactions on Broadcasting*, vol. 61, no. 4, pp. 651–665, IEEE, 2015.

[70] A. Bentaleb, A. C. Begen, and R. Zimmermann, "Sdndash: Improving qoe of http adaptive streaming using software defined networking," *Proceedings of the 24th ACM international conference on Multimedia*, pp. 1296–1305, 2016.

[71] Z. Duanmu, K. Zeng, K. Ma, A. Rehman, and Z. Wang, "A quality-of-experience index for streaming video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 154–66, 2016.

[72] Huawei. (2016) Video experience-based bearer network technical white paper. [Online]. Available: https://www.huawei.com/∼/media/CORPORATE/PDF/white%20paper/video-experience-based-bearer-network-technical-whitepaper

[73] ITU. (2017) Recommendation itu-t p.1203: Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport. [Online]. Available: https://www.itu.int/rec/T-REC-P.1203

[74] Z. Duanmu, W. Liu, D. Chen, Z. Li, Z. Wang, Y. Wang, and W. Gao, "A knowledge-driven quality-of-experience model for adaptive streaming videos," *arXiv preprint arXiv:1911.07944*, 2019.

[75] I. de Fez, R. Belda, and J. C. Guerri, "New objective qoe models for evaluating abr algorithms in dash," *Computer Communications*, vol. 158, pp. 126–140, 2020.

[76] ITU. (2020) Recommendation itu-t p.1204: Video quality assessment of streaming services over reliable transport for resolutions up to 4k. [Online]. Available: https://www.itu.int/rec/T-REC-P.1204

[77] W. Robitza, S. Göring, A. Raake, D. Lindegren, G. Heikkilä, J. Gustafsson, P. List, B. Feiten, U. Wüstenhagen, M. N. Garcia, and K. Yamagishi, "Http adaptive streaming qoe estimation with itu-t rec. p. 1203: open databases and software," *Proceedings of the 9th ACM Multimedia Systems Conference 2018*, pp. 466–471, 2018.

[78] R. K. Jain, D.-M. W. Chiu, W. R. Hawe *et al.*, "A quantitative measure of fairness and discrimination," *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*, 1984.

[79] T. Hoßfeld, L. Skorin-Kapov, P. E. Heegaard, and M. Varela, "Definition of qoe fairness in shared systems," *IEEE Communications Letters*, vol. 21, no. 1, pp. 184–187, 2016.

[80] H. Shi, R. V. Prasad, E. Onur, and I. G. M. M. Niemegeers, "Fairness in wireless networks:issues, measures and challenges," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 5–24, 2014.

[81] T. Lan, D. Kao, M. Chiang, and A. Sabharwal, "An axiomatic theory of fairness in network resource allocation," *2010 Proceedings IEEE INFOCOM*, pp. 1–9, San Diego, CA, USA, 2010.

[82] B. Radunovic and J. Y. L. Boudec, "A unified framework for max-min and min-max fairness with applications," *IEEE/ACM Transactions on Networking*, vol. 15, p. 1073–1083, 2007.

[83] J. Ozer. (2019) A video codec licensing update. [Online]. Available: https://www.streamingmedia.com/Articles/ReadArticle.aspx?ArticleID=129386

[84] ——. (2018) The future of hevc licensing is bleak, declares mpeg chairman. [Online]. Available: https://www.streamingmedia.com/Articles/ReadArticle.aspx?ArticleID=122983

[85] Mozilla. Web video codec guide. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/Media/Formats/Video_codecs

[86] J. Ozer. (2016) A cloud encoding pricing comparison. [Online]. Available: http://docs.hybrik.com/repo/cloud_pricing_comparison.pdf

[87] S. Basu. (2020) Cloud video encoding vs on-premise: Pros, cons and beyond. [Online]. Available: https://www.muvi.com/blogs/cloud-video-encoding-vs-on-premise.html

[88] A. Pellen. (2020) Cost comparison: On-premises vs cloud computing. [Online]. Available: https://www.harmonicinc.com/insights/blog/on-prem-vs-cloud

[89] F. Lambeau. (2021) Cloud-based per-title encoding workflows (with aws) – part 1: Establishing the architecture. [Online]. Available: https://bitmovin.com/cloud-based-per-title-encoding-aws-p1

[90] C. Timmerer. (2016) Mpeg-cmaf: Threat or opportunity? [Online]. Available: https://bitmovin.com/what-is-cmaf-threat-opportunity/

[91] S. Verbrugge, D. Colle, M. Pickavet, P. Demeester, S. Pasqualini, A. Iselt, and M. J. äger, "Methodology and input availability parameters for calculating opex and capex costs for realistic network scenarios," *Journal of Optical Networking*, vol. 5, no. 6, pp. 509–520, 2006.

[92] DaCast. (2020) 2019 live streaming cdn pricing comparison. [Online]. Available: https://www.dacast.com/blog/blog-live-streaming-cdn-pricing/

[93] CDNPerf. Cdn calculator. [Online]. Available: https://www.cdnperf.com/tools/cdn-calculator

[94] Wowza. Wowza streaming cloud plans. [Online]. Available: https://www.wowza.com/pricing/streaming-cloud-plans

[95] R. Viola, A. Martin, J. Morgade, S. Masneri, M. Zorrilla, P. Angueira, and J. Montalbán, "Predictive cdn selection for video delivery based on lstm network performance forecasts and cost-effective trade-offs," *IEEE Transactions on Broadcasting*, 2020.

[96] M. Kennedy, A. Ksentini, Y. Hadjadj-Aoul, and G. M. Muntean, "Adaptive energy optimization in multimedia-centric wireless devices: A survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 2, pp. 768–786, 2013.

[97] K. McClaning and T. Vito, "Radio receiver design," *Noble Publishing*, 2000.

[98] R. Trestian, O. Ormond, and G. M. Muntean, "Energy–quality–cost tradeoff in a multimedia-based heterogeneous wireless network environment," *IEEE Transactions on Broadcasting*, vol. 59, no. 2, pp. 340–357, 2013.

[99] L. Zou, A. Javed, and G. M. Muntean, "Smart mobile device power consumption measurement for video streaming in wireless environments: Wifi vs. lte," *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–6, Cagliari, Italy, 2017.

[100] C. A. Guérin, H. Nyberg, O. Perrin, S. Resnick, H. Rootzén, and C. Stărică, "Empirical testing of the infinite source poisson data traffic model," *Stochastic Models*, vol. 19, no. 2, pp. 151–200, 2003.

[101] T. Karagiannis, M. Molle, M. Faloutsos, , and A. Broido, "A nonstationary poisson view of internet traffic," *IEEE INFOCOM 2004*, vol. 3, pp. 1558–1569, IEEE, 2004.

[102] A. Bhattacharjee and S. Nandi, "Statistical analysis of network traffic inter-arrival," *The 12th International Conference on Advanced Communication Technology (ICACT)*, vol. 2, pp. 1052–1057, IEEE, 2010.

[103] W. Zhang and J. He, "Modeling end-to-end delay using pareto distribution," *Second International Conference on Internet Monitoring and Protection (ICIMP 2007)*, pp. 21–21, IEEE, 2007.

[104] M. A. Arfeen, K. Pawlikowski, D. McNickle, and A. Willig, "The role of the weibull distribution in internet traffic modeling," *Proceedings of the 2013 25th International Teletraffic Congress (ITC)*, pp. 1–8, IEEE, 2013.

[105] D. J. Daley and D. Vere-Jones, "An introduction to the theory of point processes: volume ii: general theory and structure," *Springer Science & Business Media*, 2007.

[106] Z. Liu, K. Wang, W. Li, Q. Xiao, and D. S. G. He, "Measurement and modeling study of iptv cdn network," *2009 IEEE International Conference on Network Infrastructure and Digital Content*, pp. 302–306, IEEE, 2009.

[107] A. Rao, A. Legout, Y. S. Lim, D. Towsley, C. Barakat, and W. Dabbous, "Network characteristics of video streaming traffic," *Proceedings of the Seventh COnference on emerging Networking EXperiments and Technologies*, p. 25, 2011.

[108] M. Zink, K. Suh, Y. Gu, , and J. Kurose, "Characteristics of youtube network traffic at a campus network–measurements, models, and implications," *Computer networks*, vol. 53 no. 4, pp. 501–514, 2009.

[109] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng, "Understanding user behavior in large-scale video-on-demand systems," *ACM SIGOPS Operating Systems Review*, vol. 40, no. 4, pp. 333–344, ACM, 2006.

[110] N. Liu, H. Cui, S. H. G. Chan, Z. Chen, and Y. Zhuang, "Dissecting user behaviors for a simultaneous live and vod iptv system," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 10, no.3, p. 23, 2014.

[111] R. N. Calheiros, E. Masoumi, R. Ranjan, and R. Buyya, "Workload prediction using arima model and its impact on cloud applications' qos," *IEEE Transactions on Cloud Computing*, vol. 3, no. 4, pp. 449–458, 2014.

[112] X. Dong, W. Fan, and J. Gu, "Predicting lte throughput using traffic time series," *ZTE Communications*, vol. 13, no. 4, pp. 61–64, 2015.

[113] A. Amin, L. Grunske, and A. Colman, "An automated approach to forecasting qos attributes based on linear and non-linear time series modelling," *Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering*, pp. 130–139, ACM, 2012.

[114] A. Amin, A. Colman, and L. Grunske, "An approach to forecasting qos attributes of web services based on arima and garch models," *2012 IEEE 19th International Conference on Web Services*, pp. 74–81, IEEE, 2012.

[115] C. Wang, Z. Lu, Z. Wu, J. Wu, and S. Huang, "Optimizing multi-cloud cdn deployment and scheduling strategies using big data analysis," *2017 IEEE International Conference on Services Computing (SCC)*, pp. 273–280, IEEE, 2017.

[116] M. Szmit, A. Szmit, S. Adamus, and S. Bugała, "Usage of holt-winters model and multilayer perceptron in network traffic modelling and anomaly detection," *Informatica*, vol. 36, no.4, 2012.

[117] A. A. Shahin, "Using multiple seasonal holt-winters exponential smoothing to predict cloud resource provisioning," *arXiv preprint arXiv:1701.03296*, 2017.

[118] M. Mirza, J. Sommers, P. Barford, and X. Zhu, "A machine learning approach to tcp throughput prediction," *IEEE/ACM Transactions on Networking (TON)*, vol. 18, no. 4, pp. 1026–1039, 2010.

[119] P. Bermolen and D. Rossi, "Support vector regression for link load prediction," *Computer Networks*, vol. 53, no. 2, pp. 191–201, 2009.

[120] V.-s. Feng and S. Y. Chang, "Determination of wireless networks parameters through parallel hierarchical support vector machines," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 3, pp. 505–512, 2011.

[121] X. Chen, F. Mériaux, and S. Valentin, "Predicting a user's next cell with supervised learning based on channel states," in *2013 IEEE 14th workshop on signal processing advances in wireless communications (SPAWC)*. IEEE, 2013, pp. 36–40.

[122] M. H. Zadeh and M. A. Seyyedi, "Qos monitoring for web services by time series forecasting," *2010 3rd International Conference on Computer Science and Information Technology*, vol. 5, pp. 659–663, IEEE, 2010.

[123] A. Khotanzad and N. Sadek, "Multi-scale high-speed network traffic prediction using combination of neural networks," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2, pp. 1071–1075, IEEE, 2003.

[124] S. Belhaj and M. Tagina, "Modelling and prediction of the internet end-to-end delay using recurrent neural networks," *Journal of Networks*, vol. 4, no. 6, pp. 528–535, 2009.

[125] H. D. Trinh, L. Giupponi, and P. Dini, "Mobile traffic prediction from raw data using lstm networks," *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1827–1832, IEEE, 2018.

[126] A. Azzouni and G. Pujolle, "Neutm: A neural network-based framework for traffic matrix prediction in sdn," *NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium*, pp. 1–5, IEEE, 2018.

[127] J.-M. Martinez-Caro and M.-D. Cano, "On the identification and prediction of stalling events to improve qoe in video streaming," *Electronics*, vol. 10, no. 6, p. 753, 2021.

[128] H. Cui, Y. Yao, K. Zhang, F. Sun, and Y. Liu, "Network traffic prediction based on hadoop," *2014 International Symposium on Wireless Personal Multimedia Communications (WPMC)*, pp. 29–33, IEEE, 2014.

[129] N. I. Sapankevych and R. Sankar, "Time series prediction using support vector machines: a survey," *IEEE Computational Intelligence Magazine*, vol. 4, no. 2, pp. 24–38, 2009.

[130] M. F. Iqbal, M. Zahid, D. Habib, and L. K. John, "Efficient prediction of network traffic for real-time applications," *Journal of Computer Networks and Communications*, 2019.

[131] Y. Kryftis, C. X. Mavromoustakis, G. Mastorakis, E. Pallis, J. M. Batalla, J. J. Rodrigues, C. Dobre, and G. Kormentzas, "Resource usage prediction algorithms for optimal selection of multimedia content delivery methods," *2015 IEEE international conference on communications (ICC)*, pp. 5903–5909, IEEE, 2015.

[132] G. Bontempi, S. B. Taieb, and Y. A. L. Borgne, "Machine learning strategies for time series forecasting," *European business intelligence summer school*, pp. 62–77, Springer, Berlin, Heidelberg, 2012.

[133] A. Azari, P. Papapetrou, S. Denic, and G. Peters, "Cellular traffic prediction and classification: a comparative evaluation of lstm and arima," *International Conference on Discovery Science*, 2019.

[134] ——, "User traffic prediction for proactive resource management: Learning-powered approaches," *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019.

[135] R. J. Hyndman and G. Athanasopoulos, "Forecasting: principles and practice," *OTexts*, 2018.

[136] C. C. Holt, "Forecasting seasonals and trends by exponentially weighted moving averages," *International journal of forecasting*, vol. 20, no. 1, pp. 5–10, 2004.

[137] K. R. Müller, A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, "Predicting time series with support vector machines," *International Conference on Artificial Neural Networks*, pp. 999–1004, Springer, Berlin, Heidelberg, 1997.

[138] R. Madan and P. S. Mangipudi, "Predicting computer network traffic: a time series forecasting approach using dwt, arima and rnn," *2018 Eleventh International Conference on Contemporary Computing (IC3)*, pp. 1–5, IEEE, 2018.

[139] C. N. Babu and B. E. Reddy, "Performance comparison of four new arima-ann prediction models on internet traffic data," *Journal of Telecommunications and Information Technology*, 2015.

[140] Prometheus. [Online]. Available: https://prometheus.io/

[141] S. N. Naqvi, S. Yfantidou, and E. Zimányi, "Time series databases and influxdb," *Studienarbeit, Université Libre de Bruxelles*, p. 12, 2017.

[142] Grafana. Grafana. [Online]. Available: https://grafana.com/

[143] Elastic. Elastic stack. [Online]. Available: https://www.elastic.co/elastic-stack

[144] Board. Board. [Online]. Available: https://www.board.com/en

[145] J. Hoelscher and A. Mortimer, "Using tableau to visualize data and drive decision-making," *Journal of Accounting Education*, vol. 44, pp. 49–59, 2018.

[146] Citrix. Citrix analytics. [Online]. Available: https://www.citrix.com/solutions/analytics/

[147] J. L. Ledford, J. Teixeira, and M. E. Tyler, "Google analytics," *John Wiley and Sons*, 2011.

[148] Akamai. Media analytics. [Online]. Available: https://www.akamai.com/us/en/products/media-delivery/media-analytics.jsp

[149] Conviva. Conviva streaming analytics. [Online]. Available: https://www.conviva.com/streaming-analytics/

[150] Amazon. Amazon kinesis. [Online]. Available: https://aws.amazon.com/kinesis/

[151] A. Varga and R. Hornig, "An overview of the omnet++ simulation environment," *Proceedings of the 1st international conference on Simulation tools and techniques for communications, networks and systems & workshops*, pp. 1–10, 2008.

[152] T. Issariyakul and E. Hossain, "Introduction to network simulator 2 (ns2)," *Introduction to network simulator NS2*, pp. 1–18, Springer, Boston, MA, 2009.

[153] T. R. Henderson, M. Lacage, G. F. Riley, C. Dowell, and J. Kopena, "Network simulations with the ns-3 simulator," *SIGCOMM demonstration*, vol. 14, no. 14, p.527, 2008.

[154] X. Chang, "Network simulations with opnet," *WSC'99. 1999 Winter Simulation Conference Proceedings.'Simulation-A Bridge to the Future'(Cat. No. 99CH37038)*, vol. 1, pp. 307–314, IEEE, 1999.

[155] M. A. I. V. N. on your Laptop (or other PC). [Online]. Available: http://mininet.org/

[156] Tetcos. Netsim. [Online]. Available: https://www.tetcos.com/netsim-std.html

[157] A. Keränen, J. Ott, and T. Kärkkäinen, "The one simulator for dtn protocol evaluation," *Proceedings of the 2nd international conference on simulation tools and techniques*, pp. 1–10, 2009.

[158] iperf - the ultimate speed test tool for tcp and udp and sctp. [Online]. Available: https://iperf.fr/

[159] M. Jemec. packeth–ethernet packet generator. [Online]. Available: http://packeth.sourceforge.net/packeth/Home.html

[160] R. Olsson, "Pktgen the linux packet generator," *Proceedings of the Linux Symposium*, vol. 2, pp. 11–24, Ottawa, Canada, 2005.

[161] P. Emmerich, S. Gallenmüller, D. Raumer, F. Wohlfart, and G. Carle, "Moongen: A scriptable high-speed packet generator," *Proceedings of the 2015 Internet Measurement Conference*, pp. 275–287, 2015.

[162] N. Bonelli, S. Giordano, G. Procissi, and S. Raffaello, "Brute: A high performance and extensibile traffic generator," *Int'l Symposium on Performance of Telecommunication Systems (SPECTS'05)*, vol. 1, pp. 222–227, 2005.

[163] J. Sommers and P. Barford, "Self-configuring network traffic generation," *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pp. 68–81, ACM, 2004.

[164] B. R. Patil, M. Moharir, P. K. Mohanty, G. Shobha, and S. Sajeev, "Ostinato-a powerful traffic generator," *2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*, pp. 1–5, IEEE, 2017.

[165] Cisco. Trex. [Online]. Available: https://trex-tgn.cisco.com/

[166] S. Avallone, S. Guadagno, D. Emma, A. Pescapè, and G. Ventre, "D-itg distributed internet traffic generator," *First International Conference on the Quantitative Evaluation of Systems*, pp. 316–317, IEEE, 2004.

[167] A. Botta, A. Dainotti, and A. Pescapé, "A tool for the generation of realistic network workload for emerging networking scenarios," *Computer Networks*, vol. 56, n. 15, pp. 3531–3547, 2012.

[168] HP. (2006) Seagull – open source tool for ims testing. [Online]. Available: http://gull.sourceforge.net/doc/WP_Seagull_Open_Source_tool_for_IMS_testing.pdf

[169] C. Liang, F. R. Yu, and X. Zhang, "Information-centric network function virtualization over 5g mobile wireless networks," *IEEE network*, vol. 29, no. 3, pp. 68–74, 2015.

[170] ETSI. (2014) Etsi gs nfv-man 001: Network functions virtualisation (nfv); management and orchestration. [Online]. Available: https://www.etsi.org/deliver/etsi_gs/nfv-man/001_099/001/01.01.01_60/gs_nfv-man001v010101p.pdf

[171] ——. Open source mano (osm). [Online]. Available: https://osm.etsi.org/

[172] L. Foundation. Open network automation platform (onap). [Online]. Available: https://www.onap.org/

[173] D. Sabella, V. Sukhomlinov, L. Trang, S. Kekki, P. Paglierani, R. Rossbach, X. Li, Y. Fang, D. Druta, F. Giust, and L. Cominardi, "Developing software for multi-access edge computing," *ETSI white paper 20*, pp. 1–38, 2019.

[174] N. Nikaein, M. K. Marina, S. Manickam, A. Dawson, R. Knopp, and C. Bonnet, "OpenAirInterface: A Flexible Platform for 5G Research," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 5, pp. 33–38, Oct. 2014.

[175] I. Gomez-Miguelez, A. Garcia-Saavedra, P. D. Sutton, P. Serrano, C. Cano, and D. J. Leith, "srslte: An open-source platform for lte evolution and experimentation," *Proceedings of the Tenth ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation, and Characterization*, pp. 25–32, 2016.

[176] V. G. Nguyen, A. Brunstrom, K. J. Grinnemo, and J. Taheri, "Sdn/nfv-based mobile packet core network architectures: A survey," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1567–602, 2017.

[177] A. F. Ocampo, T. Dreibholz, M. R. Fida, A. Elmokashfi, and H. Bryhni, "Integrating cloud-ran with packet core as vnf using open source mano and openairinterface," *Proceedings of the 45th IEEE Conference on Local Computer Networks (LCN)*, 2020.

[178] A. Gabilondo, J. Morgade, R. Viola, J. F. Mogollon, M. Zorrilla, P. Angueira, and J. Montalbán, "Realising a vran based fembms management and orchestration framework," *2020 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–7, 2020.

[179] M. Yang, Y. Li, D. Jin, L. Su, S. Ma, and L. Zeng, "Openran: A software-defined ran architecture via virtualization," *SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 4, p. 549–550, Aug. 2013. [Online]. Available: https://doi.org/10.1145/2534169.2491732

[180] L. Gavrilovska, V. Rakovic, and D. Denkovski, "From cloud ran to open ran," *Wireless Personal Communications*, pp. 1–17, 2020.

[181] ETSI. (2017) Etsi gs mec 012: Mobile edge computing (mec); radio network information api. [Online]. Available: https://www.etsi.org/deliver/etsi_gs/MEC/001_099/012/01.01.01_60/gs_MEC012v010101p.pdf

[182] ——. (2018) Etsi gs mec 002: Multi-access edge computing (mec): Phase 2: Use cases and requirements. [Online]. Available: https://www.etsi.org/deliver/etsi_gs/MEC/001_099/002/02.01.01_60/gs_MEC002v020101p.pdf

[183] F. Giannone, P. A. Frangoudis, A. Ksentini, and L. Valcarenghi, "Orchestrating heterogeneous mec-based applications for connected vehicles," *Computer Networks*, vol. 180, 2020.

[184] A. Martin, R. Viola, M. Zorrilla, J. Flórez, P. Angueira, and J. Montalbán, "Mec for fair, reliable and efficient media streaming in mobile networks," *IEEE Transactions on Broadcasting*, vol. 66, no. 2, pp. 264–78, 2019.

[185] Y. Li, P. A. Frangoudis, Y. Hadjadj-Aoul, and P. Bertin, "A mobile edge computing-based architecture for improved adaptive http video delivery," *2016 IEEE Conference on Standards for Communications and Networking (CSCN)*, pp. 1–6, IEEE, 2016.

[186] J. J. G. et al., "5g new radio for terrestrial broadcast: A forward-looking approach for nr-mbms," *IEEE Transactions on Broadcasting*, vol. 65, no. 2, pp. 356–368, 2019. doi: 10.1109/TBC.2019.2912117.

[187] A. Doumanoglou, N. Zioulis, D. Griffin, J. Serrano, T. K. Phan, D. Jiménez, D. Zarpalas, F. Alvarez, M. Rio, and P. Daras, "A system architecture for live immersive 3d-media transcoding over 5g networks," *2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 11–15, 2018.

[188] S. Dutta, T. Taleb, P. A. Frangoudis, and A. Ksentini, "On-the-fly qoe-aware transcoding in the mobile edge," *2016 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, IEEE, 2016.

[189] S. Rezvani, S. Parsaeefard, N. Mokari, M. R. Javan, and H. Yanikomeroglu, "Cooperative multi-bitrate video caching and transcoding in multicarrier noma-assisted heterogeneous virtualized mec networks," *IEEE Access*, vol. 7, pp. 93 511–93 536, 2019. doi: 10.1109/ACCESS.2019.2927903.

[190] C. Liu, H. Zhang, H. Ji, and X. Li, "Mec-assisted flexible transcoding strategy for adaptive bitrate video streaming in small cell networks," *China Communications*, vol. 18, no. 2, pp. 200–214, 2021.

[191] T. X. Tran and D. Pompili, "Adaptive bitrate video caching and processing in mobile-edge computing networks," *IEEE Transactions on Mobile Computing*, vol. 18, no. 9, pp. 1965–1978, 2019.

[192] Y. Wang, Y. Zhang, M. Sheng, and K. Guo, "On the interaction of video caching and retrieving in multi-server mobile-edge computing systems," *IEEE Wireless Communications Letters*, vol. 8, no. 5, pp. 1444–1447, 2019. doi: 10.1109/LWC.2019.2921759.

[193] Q. Jia, R. Xie, H. Lu, W. Zheng, and H. Luo, "Joint optimization scheme for caching, transcoding and bandwidth in 5g networks with mobile edge computing," *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, pp. 999–1004, 2019.

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

[195] V. K. Adhikari, Y. Guo, F. Hao, V. Hilt, and Z. L. Zhang, "A tale of three cdns: An active measurement study of hulu and its cdns," *2012 Proceedings IEEE INFOCOM Workshops*, pp. 7–12, IEEE, 2012.

[196] V. K. Adhikari, Y. Guo, F. Hao, V. Hilt, Z. L. Zhang, M. Varvello, and M. Steiner, "Measurement study of netflix, hulu, and a tale of three cdns," *IEEE/ACM Transactions on Networking*, vol. 23, no. 6, pp. 1984–1997, 2015.

[197] J. S. Otto, M. A. Sánchez, J. P. Rula, T. Stein, and F. E. Bustamante, "namehelp: Intelligent client-side dns resolution," in *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication*, 2012, pp. 287–288.

[198] R. Torres, A. Finamore, J. R. Kim, M. Mellia, M. M. Munafo, and S. Rao, "Dissecting video server selection strategies in the youtube cdn," *2011 31st International Conference on Distributed Computing Systems*, pp. 248–257, IEEE, 2011.

[199] U. Goel, M. P. Wittie, and M. Steiner, "Faster web through client-assisted cdn server selection," *2015 24th International conference on computer communication and networks (ICCCN)*, pp. 1–10, IEEE, 2015.

[200] T. Böttger, F. Cuadrado, G. Tyson, I. Castro, and S. Uhlig, "Open connect everywhere: A glimpse at the internet ecosystem through the lens of the netflix cdn," *ACM SIGCOMM Computer Communication Review*, vol. 48, no. 1, pp. 28–34, 2018.

[201] SVA. Open caching. [Online]. Available: https://www.streamingvideoalliance.org/working-group/open-caching/

[202] B. Frank, I. Poese, Y. Lin, G. Smaragdakis, A. Feldmann, B. Maggs, J. Rake, S. Uhlig, and R. Weber, "Pushing cdn-isp collaboration to the limit," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 3, pp. 34–44, 2013.

[203] M. Wichtlhuber, R. Reinecke, and D. Hausheer, "An sdn-based cdn/isp collaboration architecture for managing high-volume flows," *IEEE Transactions on Network and Service Management*, vol. 12, no. 1, pp. 48–60, 2015.

[204] EBU. Eurovision flow. [Online]. Available: https://tech.ebu.ch/docs/groups/flow/Eurovision%20Flow%20Brochure.pdf

[205] Citrix. Intelligent traffic management. [Online]. Available: https://www.citrix.com/products/citrix-intelligent-traffic-management/

[206] Haivision. Haivision lightflow multicdn. [Online]. Available: https://www.haivision.com/products/haivision-lightflow-multicdn/

[207] R. Viola, A. Martin, M. Zorrilla, and J. Montalban, "Mec proxy for efficient cache and reliable multi-cdn video distribution," *2018*

IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), pp. 1–7, IEEE, 2018.

[208] G. Carrozzo, F. Moscatelli, G. Solsona, O. P. Gordo, M. Keltsch, and M. Schmalohr, “Virtual cdns over 5g networks: scenarios and requirements for ultra-high definition media distribution,” *2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–5, IEEE, 2018.

[209] Y. Tan, C. Han, M. Luo, and X. Z. X. Zhang, “Radio network-aware edge caching for video delivery in mec-enabled cellular networks,” *2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pp. 179–184, IEEE, 2018.

[210] C. Ge, N. Wang, S. Skillman, G. Foster, and Y. Cao, “Qoe-driven dash video caching and adaptation at 5g mobile edge,” *Proceedings of the 3rd ACM Conference on Information-Centric Networking*, pp. 237–242, ACM, 2016.

[211] Y. Chen, Y. Liu, J. Zhao, and Q. Zhu, “Mobile edge cache strategy based on neural collaborative filtering,” *IEEE Access*, vol. 8, pp. 18 475–18 482, 2020.

[212] 3GPP, “Overall description of lte-based 5g broadcast; version 16.0.0; technical report (tr) 36.976,” *3rd Generation Partnership Project (3GPP)*, 2020.

[213] H. Ma, S. Li, E. Zhang, Z. Lv, J. Hu, and X. Wei, “Cooperative autonomous driving oriented mec-aided 5g-v2x: Prototype system design, field tests and ai-based optimization tools,” *IEEE Access*, vol. 8, pp. 54 288–54 302, 2020. doi: 10.1109/ACCESS.2020.2981463.

[214] G. Velez, A. Martin, G. Pastor, and E. Mutafungwa, “5g beyond 3gpp release 15 for connected automated mobility in cross-border contexts,” *Sensors*, vol. 20, no. 22, 2020.

[215] 5GINFIRE. D2.2-5ginfire experimental infrastructure architecture and 5g automotive use case(update). [Online]. Available: https://bscw.5g-ppp.eu/pub/bscw.cgi/d302168/D2-2-5GINFIRE_Experimental_Infrastructure_Architecture_and_5G_Automotive_Use_Case-v1-0.pdf

[216] T. V. Doan, V. Bajpai, and S. Crawford, “A longitudinal view of netflix: Content delivery over ipv6 and content cache deployments,” in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 1073–1082.

[217] A. Biliris, C. Cranor, F. Douglis, M. Rabinovich, S. Sibal, O. Spatscheck, and W. Sturm, “Cdn brokering,” *Computer Communications*, vol. 25, no. 4, pp. 393–402, 2002.

[218] ETSI. (2020) Etsi ts 128 313: 5g;self-organizing networks (son) for 5g networks. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/128300_128399/128313/16.00.00_60/ts_128313v160000p.pdf

[219] O. G. Aliu, A. Imran, M. A. Imran, and B. Evans, “A survey of self organisation in future cellular networks,” *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 336–361, 2012.

[220] H. Hu, J. Zhang, X. Zheng, Y. Yang, and P. Wu, “Self-configuration and self-optimization for lte networks,” *IEEE Communications Magazine*, vol. 48, no. 2, pp. 94–100, 2010.

[221] J. Moysen and L. Giupponi, “From 4g to 5g: Self-organized network management meets machine learning,” *Computer Communications*, vol. 129, pp. 248–268, 2018.

[222] A. Imran, A. Zoha, and A. Abu-Dayya, “Challenges in 5g: how to empower son with big data for enabling 5g,” *IEEE network*, vol. 28, no. 6, pp. 27–33, 2014.

[223] P. V. Klaine, M. A. Imran, O. Onireti, and R. D. Souza, “A survey of machine learning techniques applied to self-organizing cellular networks,” *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2392–2431, 2017.

[224] HCL. Son. [Online]. Available: https://www.hcltech.com/ERX/Telecom-and-5G/SON

[225] Nokia. Edennet. [Online]. Available: https://www.nokia.com/networks/portfolio/self-organizing-networks

[226] Ericsson. Son optimization manager. [Online]. Available: https://www.ericsson.com/en/portfolio/digital-services/automated-network-operations/network-management/son-optimization-manager

[227] Y. Ouyang, Z. Li, L. Su, W. Lu, and Z. Lin, “Application behaviors driven self-organizing network (son) for 4g lte networks,” *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 1, pp. 3–14, 2018.

[228] J. I. Khan, S. S. Yang, Q. Gu, D. Patel, P. Mail, O. Komogortsev, W. Oh, and Z. Guo, “Resource adaptive netcentric systems: A case study with sonet-a self-organizing network embedded transcoder,” *Proceedings of the ninth ACM international conference on Multimedia*, pp. 617–620, 2001.

[229] C. Singhal and B. N. Chandana, “Aerial-son: Uav-based self-organizing network for video streaming in dense urban scenario,” *2021 International Conference on COMmunication Systems & NETworkS (COMSNETS)*, pp. 7–12, 2021.

[230] J. G. Herrera and J. F. Botero, “Resource allocation in nfv: A comprehensive survey,” *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 518–532, 2016.

[231] Y. Xie, Z. Liu, S. Wang, and Y. Wang, “Service function chaining resource allocation: A survey,” *arXiv preprint arXiv:1608.00095*, 2016.

[232] F. Schardong, I. Nunes, and A. Schaeffer-Filho, “Nfv resource allocation: A systematic review and taxonomy of vnf forwarding graph embedding,” *Computer Networks*, p. 107726, 2020.

[233] J. F. Riera, E. Escalona, J. Batalle, E. Grasa, and J. A. Garcia-Espin, “Virtual network function scheduling: Concept and challenges,” *2014 international conference on smart communications in network technologies (SaCoNeT)*, pp. 1–5, IEEE, 2014.

[234] O. Adamuz-Hinojosa, J. Ordonez-Lucena, P. Ameigeiras, J. J. Ramos-Munoz, D. Lopez, and J. Folgueira, “Automated network service scaling in nfv: Concepts, mechanisms and scaling workflow,” *IEEE Communications Magazine*, vol. 56, no. 7, pp. 162–169, 2018.

[235] R. Mijumbi, J. Serrat, J.-L. Gorricho, S. Latré, M. Charalambides, and D. Lopez, “Management and orchestration challenges in network functions virtualization,” *IEEE Communications Magazine*, vol. 54, no. 1, pp. 98–105, 2016.

[236] X. Wang, C. Wu, F. Le, A. Liu, Z. Li, and F. Lau, “Online vnf scaling in datacenters,” *2016 IEEE 9th International Conference on Cloud Computing (CLOUD)*, pp. 140–147, IEEE, 2016.

[237] M. Ghaznavi, A. Khan, N. Shahriar, K. Alsubhi, R. Ahmed, and R. Boutaba, “Elastic virtual network function placement,” *2015 IEEE 4th International Conference on Cloud Networking (CloudNet)*, pp. 255–260, IEEE, 2015.

[238] D. A. Tedjopurnomo, Z. Bao, B. Zheng, F. Choudhury, and A. K. Qin, “A survey on modern deep neural network for traffic prediction: Trends, methods and challenges,” *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[239] P. Sandhir and K. Mitchell, “A neural network demand prediction scheme for resource allocation in cellular wireless systems,” *2008 IEEE Region 5 Conference*, pp. 1–6, IEEE, 2008.

[240] X. Fei, F. Liu, H. Xu, and H. Jin, “Adaptive vnf scaling and flow routing with proactive demand prediction,” *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pp. 486–494, 2018.

[241] R. Mijumbi, S. Hasija, S. Davy, A. Davy, B. Jennings, and R. Boutaba, “A connectionist approach to dynamic resource management for virtualised network functions,” *2016 12th International Conference on Network and Service Management (CNSM)*, pp. 1–9, 2016.

[242] X. Zhang, C. Wu, Z. Li, and F. C. Lau, “Proactive vnf provisioning with multi-timescale cloud resources: Fusing online learning and online optimization,” *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, pp. 1–9, 2017.

[243] H. Huang and S. Guo, “Proactive failure recovery for nfv in distributed edge computing,” *IEEE Communications Magazine*, vol. 57, no. 5, pp. 131–137, 2019.

[244] F. J. Moreno-Muro, M. Garrich, C. San-Nicolás-Martínez, M. Hernández-Bastida, P. Pavón-Mariño, A. Bravalheri, A. S. Muqaddas, N. Uniyal, R. Nejabati, D. Simeonidou, and R. Casellas, “Joint vnf and multi-layer resource allocation with an open-source optimization-as-a-service integration,” *45th European Conference on Optical Communication (ECOC 2019)*, pp. 1–4, 2019.

[245] A. Ndikumana, S. Ullah, T. LeAnh, N. H. Tran, and C. S. Hong, “Collaborative cache allocation and computation offloading in mobile edge computing,” *2017 19th Asia-Pacific Network Operations and Management Symposium (APNOMS)*, pp. 366–369, IEEE, 2017.

[246] M. T. Beck, S. Feld, A. Fichtner, C. Linnhoff-Popien, and T. Schimper, “Me-volte: Network functions for energy-efficient video transcoding at the mobile edge,” *2015 18th International Conference on Intelligence in Next Generation Networks*, pp. 38–44, IEEE, 2015.

[247] Y. Sun, T. Wei, H. Li, T. Zhang, and W. Wu, “Energy-efficient multimedia task assignment and computing offloading for mobile edge computing networks,” *IEEE Access*, vol. 8, pp. 36 702–36 713, 2020.

[248] E. Ahmed, A. Ahmed, I. Yaqoob, J. Shuja, A. Gani, M. Imran, and M. Shoaib, “Bringing computation closer toward the user network: Is edge computing the solution?” *IEEE Communications Magazine*, vol. 55, no. 11, 138-144, 2017.

[249] S. Liu, L. Liu, J. Tang, B. Yu, Y. Wang, and W. Shi, “Edge computing for autonomous driving: Opportunities and challenges,” *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1697–1716, 2019.

[250] S. Biookaghazadeh, M. Zhao, and F. Ren, “Are fpgas suitable for edge computing?” in {*USENIX*} *Workshop on Hot Topics in Edge Computing (HotEdge 18)*, 2018.

[251] I. Colbert, J. Daly, K. Kreutz-Delgado, and S. Das, "A competitive edge: Can fpgas beat gpus at dcnn inference acceleration in resource-limited edge computing applications?" *arXiv preprint arXiv:2102.00294*, 2021.

[252] S. Arora, P. A. Frangoudis, and A. Ksentini, "Exposing radio network information in a mec-in-nfv environment: the rnisaas concept," *2019 IEEE Conference on Network Softwarization (NetSoft)*, pp. 306–310, 2019.

[253] L. Tomaszewski, S. Kukliński, and R. Kołakowski, "A new approach to 5g and mec integration," *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 15–24, 2020.

[254] X. Jiang, F. R. Yu, T. Song, and V. C. Leung, "A survey on multi-access edge computing applied to video streaming: Some research issues and challenges," *IEEE Communications Surveys & Tutorials*, 2021.

[255] G. Gür, P. Porambage, and M. Liyanage, "Convergence of icn and mec for 5g: Opportunities and challenges," *IEEE Communications Standards Magazine*, vol. 4, no. 4, pp. 64–71, 2020.

[256] R. Roman, J. Lopez, and M. Mambo, "Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges," *Future Generation Computer Systems*, vol. 78, pp. 680–698, 2018.

[257] Y. Xiao, Y. Jia, C. Liu, X. Cheng, J. Yu, and W. Lv, "Edge computing security: State of the art and challenges," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1608–1631, 2019.

[258] S. Shahzadi, M. Iqbal, T. Dagiuklas, and Z. U. Qayyum, "Multi-access edge computing: open issues, challenges and future perspectives," *Journal of Cloud Computing*, vol. 6, no. 1, pp. 1–3, 2017.

[259] J. Martín-Pérez, L. Cominardi, C. J. Bernardos, A. de la Oliva, and A. Azcorra, "Modeling mobile edge computing deployments for low latency multimedia services," *IEEE Transactions on Broadcasting*, vol. 65, no. 2, pp. 464–474, 2019.

[260] J. Yao, T. Han, and N. Ansari, "On mobile edge caching," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2525–2553, 2019.

[261] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5g: Survey and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94–100, 2017.

[262] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2429–2453, 2018.

[263] A. Nakao, P. Du, Y. Kiriha, F. Granelli, A. A. Gebremariam, T. Taleb, and M. Bagaa, "End-to-end network slicing for 5g mobile networks," *Journal of Information Processing*, vol. 25, pp. 153–163, 2017.

[264] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, and D. Aziz, "Network slicing to enable scalability and flexibility in 5g mobile networks," *IEEE Communications magazine*, vol. 55, no. 5, pp. 72–79, 2017.

[265] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira, "Network slicing for 5g with sdn/nfv: Concepts, architectures, and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80–7, 2017.

[266] S. Zhang, "An overview of network slicing for 5g," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 111–117, 2019.

[267] Z. Kotulski, T. Nowak, M. Sepczuk, M. Tunia, R. Artych, K. Bocianiak, T. Osko, and J. P. Wary, "On end-to-end approach for slice isolation in 5g networks. fundamental challenges," *2017 Federated conference on computer science and information systems (FedCSIS)*, pp. 783–792, 2017.

[268] Q. Li, G. Wu, A. Papathanassiou, and U. Mukherjee, "An end-to-end network slicing framework for 5g wireless communication systems," *arXiv preprint arXiv:1608.00572*, 2016.

[269] P. Alemany, L. Juan, A. Pol, A. Roman, P. Trakadas, P. Karkazis, M. Touloupou, E. Kapassa, D. Kyriazis, T. Soenen, and C. Parada, "Network slicing over a packet/optical network for vertical applications applied to multimedia real-time communications," *2019 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, pp. 1–2, IEEE, 2019.

[270] Q. Wang, J. Alcaraz-Calero, R. Ricart-Sanchez, M. B. Weiss, A. Gavras, N. Nikaein, X. Vasilakos, B. Giacomo, G. Pietro, M. Roddy, M. Healy, P. Walsh, T. Truong, Z. Bozakov, K. Koutsopoulos, P. Neves, C. Patachia-Sultanoiu, M. Iordache, E. Oproiu, I. G. B. Yahia, C. Angelo, C. Zotti, G. Celozzi, D. Morris, R. Figueiredo, D. Lorenz, S. Spadaro, G. Agapiou, A. Aleixo, and C. Lomba, "Enable Advanced QoS-Aware Network Slicing in 5G Networks for Slice-Based Media Use Cases," *IEEE Transactions on Broadcasting*, vol. 65, no. 2, pp. 444–453, Jun. 2019.

[271] A. H. Celdrán, M. G. Pérez, F. J. G. Clemente, F. Ippoliti, and G. M. Pérez, "Dynamic network slicing management of multimedia scenarios for future remote healthcare," *Multimedia Tools and Applications*, vol. 78, no. 17, pp. 24 707–24 737, 2019.

[272] C. Campolo, A. Molinaro, A. Iera, and F. Menichella, "5g network slicing for vehicle-to-everything services," *IEEE Wireless Communications*, vol. 24, no. 6, pp. 38–45, 2017.

[273] J. Mei, X. Wang, and K. Zheng, "Intelligent network slicing for v2x services toward 5g," *IEEE Network*, vol. 33, no. 6, pp. 196–204, 2019.

[274] F. Kurtz, C. Bektas, N. Dorsch, and C. Wietfeld, "Network slicing for critical communications in shared 5g infrastructures-an empirical evaluation," *2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft)*, pp. 393–399, IEEE, 2018.

[275] A. A. Barakabitze, A. Ahmad, R. Mijumbi, and A. Hines, "5g network slicing using sdn and nfv: A survey of taxonomy and architectures and future challenges," *Computer Networks*, vol. 167, p. 106984, 2020.

[276] S. Vassilaras, L. Gkatzikis, N. Liakopoulos, I. N. Stiakogiannakis, M. Qi, L. Shi, L. Liu, M. Debbah, and G. S. Paschos, "The algorithmic aspects of network slicing," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 112–119, 2017.

[277] Z. M. Fadlullah, F. Tang, B. Mao, N. Kato, O. Akashi, T. Inoue, and K. Mizutani, "State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2432–24 355, 2017.

[278] C.-L. I, S. Kuklinskí, and T. Chen, "A perspective of o-ran integration with mec, son, and network slicing in the 5g era," *IEEE Network*, vol. 34, no. 6, pp. 3–4, 2020.

[279] Y. Zhou, L. Liu, L. Wang, N. Hui, X. Cui, J. Wu, Y. Peng, Y. Qi, and C. Xing, "Service aware 6g: an intelligent and open network based on convergence of communication, computing and caching," *Digital Communications and Networks*, 2020.

[280] Y. Wei, M. Peng, and Y. Liu, "Intent-based networks for 6g: Insights and challenges," *Digital Communications and Networks*, vol. 6, no. 3, pp. 270–280, 2020.

[281] J.-B. Monteil, J. Hribar, P. Barnard, Y. Li, and L. A. DaSilva, "Resource reservation within sliced 5g networks: A cost-reduction strategy for service providers," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2020, pp. 1–6.

[282] D. Chen, Y.-C. Liu, B. Kim, J. Xie, C. S. Hong, and Z. Han, "Edge computing resources reservation in vehicular networks: A meta-learning approach," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 5, pp. 5634–5646, 2020.

[283] X. Wang, J. Li, L. Wang, C. Yang, and Z. Han, "Intelligent user-centric network selection: A model-driven reinforcement learning framework," *IEEE Access*, vol. 7, pp. 21 645–21 661, 2019.

[284] J. R. Bhat and S. A. Alqahtani, "6g ecosystem: Current status and future perspective," *IEEE Access*, vol. 9, pp. 43 134–43 167, 2021.

[285] Z. Allam and D. S. Jones, "Future (post-covid) digital, smart and sustainable cities in the wake of 6g: Digital twins, immersive realities and new urban economies," *Land Use Policy*, vol. 101, p. 105201, 2021.

[286] M. W. Akhtar, S. A. Hassan, R. Ghaffar, H. Jung, S. Garg, and M. S. Hossain, "The shift to 6g communications: vision and requirements," *Human-centric Computing and Information Sciences*, vol. 10, no. 1, pp. 1–27, 2020.

[287] A. Rostami, "Private 5g networks for vertical industries: Deployment and operation models," in *2019 IEEE 2nd 5G World Forum (5GWF)*. IEEE, 2019, pp. 433–439.

[288] L. Bonati, M. Polese, S. D'Oro, S. Basagni, and T. Melodia, "Open, programmable, and virtualized 5g networks: State-of-the-art and the road ahead," *Computer Networks*, vol. 182, p. 107516, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128620311786

[289] O. Galinina, A. Pyattaev, S. Andreev, M. Dohler, and Y. Koucheryavy, "5g multi-rat lte-wifi ultra-dense small cells: Performance dynamics, architecture, and trends," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 6, pp. 1224–1240, 2015.

[290] P. Basaras, G. Iosifidis, S. Kucera, and H. Claussen, "Multicast optimization for video delivery in multi-rat networks," *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 4973–4985, 2020.

[291] S. Borst, A. Ö. Kaya, D. Calin, and H. Viswanathan, "Dynamic path selection in 5g multi-rat wireless networks," in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 2017, pp. 1–9.

[292] A. Yazar, S. D. Tusha, and H. Arslan, "6g vision: An ultra-flexible perspective," *ITU Journal on Future and Evolving Technologies*, vol. 1, no. 1, 2020.

[293] T. Huang, W. Yang, J. Wu, J. Ma, X. Zhang, and D. Zhang, "A survey on green 6g network: Architecture and technologies," *IEEE Access*, vol. 7, pp. 175 758–175 768, 2019.

[294] D. Renga and M. Meo, "From self-sustainable green mobile networks to enhanced interaction with the smart grid," in *2018 30th International Teletraffic Congress (ITC 30)*, vol. 1. IEEE, 2018, pp. 129–134.

[295] H. Abou-Zeid and H. S. Hassanein, "Predictive green wireless access: Exploiting mobility and application information," *IEEE wireless communications*, vol. 20, no. 5, pp. 92–99, 2013.

[296] A. Mehrabi, M. Siekkinen, and A. Ylä-Jääski, "Energy-aware qoe and backhaul traffic optimization in green edge adaptive mobile video streaming," *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 3, pp. 828–839, 2019.

[297] C. Xu, L. Yang, and P. Zhang, "Practical backscatter communication systems for battery-free internet of things: A tutorial and survey of recent research," *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 16–27, 2018.

[298] S. Naderiparizi, M. Hessar, V. Talla, S. Gollakota, and J. R. Smith, "Towards battery-free {HD} video streaming," in *15th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 18)*, 2018, pp. 233–247.

[299] A. Saffari, M. Hessar, S. Naderiparizi, and J. R. Smith, "Battery-free wireless video streaming camera system," in *2019 IEEE International Conference on RFID (RFID)*. IEEE, 2019, pp. 1–8.

[300] Y. Al-Issa, M. A. Ottom, and A. Tamrawi, "ehealth cloud security challenges: a survey," *Journal of healthcare engineering*, vol. 2019, 2019.

[301] J. Cui, L. S. Liew, G. Sabaliauskaite, and F. Zhou, "A review on safety failures, security attacks, and available countermeasures for autonomous vehicles," *Ad Hoc Networks*, vol. 90, p. 101823, 2019.

[302] M. Wang, T. Zhu, T. Zhang, J. Zhang, S. Yu, and W. Zhou, "Security and privacy in 6g networks: New areas and new challenges," *Digital Communications and Networks*, vol. 6, no. 3, pp. 281–291, 2020.

[303] "CogNet (Building an Intelligent System of Insights and Action for 5G Network Management)." [Online]. Available: http://www.cognet.5g-ppp.eu/

[304] I. G. Ben Yahia, J. Bendriss, A. Samba, and P. Dooze, "CogNitive 5G networks: Comprehensive operator use cases with machine learning for management operations," in *2017 20th Conference on Innovations in Clouds, Internet and Networks (ICIN)*. Paris: IEEE, Mar. 2017, pp. 252–259.

[305] H. Assem, L. Xu, T. S. Buda, and D. O'Sullivan, "Machine learning as a service for enabling Internet of Things and People," *Personal and Ubiquitous Computing*, vol. 20, no. 6, pp. 899–914, Nov. 2016.

[306] "SELFNET (Framework for Self-Organized Network Management in Virtualized and Software Defined Networks)." [Online]. Available: https://selfnet-5g.eu/

[307] A. H. Celdran, M. G. Perez, F. J. G. Clemente, and G. M. Perez, "Enabling Highly Dynamic Mobile Scenarios with Software Defined Networking," *IEEE Communications Magazine*, vol. 55, no. 4, pp. 108–113, Apr. 2017.

[308] "SliceNet (End-to-End Cognitive Network Slicing and Slice Management Framework in Virtualised Multi-Domain, Multi-Tenant 5G Networks)." [Online]. Available: https://slicenet.eu/

[309] P. Salva-Garcia, J. M. Alcaraz-Calero, Q. Wang, M. Arevalillo-Herraez, and J. Bernal Bernabe, "Scalable Virtual Network Video-Optimizer for Adaptive Real-Time Video Transmission in 5G Networks," *IEEE Transactions on Network and Service Management*, vol. 17, no. 2, pp. 1068–1081, Jun. 2020.

[310] "SoftFIRE (Software Defined Networks and Network Function Virtualization Testbed within FIRE+)." [Online]. Available: https://www.softfire.eu/

[311] D. Lake, G. Foster, S. Vural, Y. Rahulan, B.-H. Oh, N. Wang, and R. Tafazolli, "Virtualising and orchestrating a 5G evolved packet core network," in *2017 IEEE Conference on Network Softwarization (NetSoft)*, Jul. 2017, pp. 1–5.

[312] "FLAME (Facility for Large-scale Adaptive Media Experimentation)," Jan. 2017. [Online]. Available: https://www.ict-flame.eu/

[313] K. Haensge, D. Trossen, S. Robitzsch, M. Boniface, and S. Phillips, "Cloud-Native 5G Service Delivery Platform," in *2019 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, Nov. 2019, pp. 1–7.

[314] "5GTango (5G Development and validation platform for global industry-specific network services and Apps)." [Online]. Available: https://5gtango.eu/

[315] M. Peuster, S. Schneider, M. Zhao, G. Xilouris, P. Trakadas, F. Vicens, W. Tavernier, T. Soenen, R. Vilalta, G. Andreou, D. Kyriazis, and H. Karl, "Introducing Automated Verification and Validation for Virtualized Network Functions and Services," *IEEE Communications Magazine*, vol. 57, no. 5, pp. 96–102, May 2019.

[316] T. Soenen, W. Tavernier, M. Peuster, F. Vicens, G. Xilouris, S. Kolometsos, Michail-Alexandros Kourtis, and D. Colle, "Empowering network service developers: enhanced nfv devops and programmable mano," *IEEE Communications Magazine*, May 2019.

[317] "5G-Media (Programmable Edge-to-Cloud Virtualization Fabric for the 5G Media Industry)." [Online]. Available: http://www.5gmedia.eu/

[318] D. Breitgand, A. Weit, S. Rizou, D. Griffin, U. Acar, G. Carrozzo, N. Zioulis, P. Andriani, and F. Iadanza, "Towards Serverless NFV for 5G Media Applications," in *Proceedings of the 11th ACM International Systems and Storage Conference*, ser. SYSTOR '18. Haifa, Israel: Association for Computing Machinery, Jun. 2018, p. 118.

[319] "5Growth (5G-enabled Growth in Vertical Industries)." [Online]. Available: https://5growth.eu/

[320] X. Li, A. Garcia-Saavedra, X. Costa-Perez, C. J. Bernardos, C. Guimarães, K. Antevski, J. Mangues-Bafalluy, J. Baranda, E. Zeydan, D. Corujo *et al.*, "5growth: An end-to-end service platform for automated deployment and management of vertical services over 5g networks," *IEEE Communications Magazine*, vol. 59, no. 3, pp. 84–90, 2021.

[321] "5GCity – A distributed cloud & radio platform for 5G Neutral Hosts." [Online]. Available: https://www.5gcity.eu/

[322] C. Colman-Meixner, H. Khalili, K. Antoniou, M. S. Siddiqui, A. Papageorgiou, A. Albanese, P. Cruschelli, G. Carrozzo, L. Vignaroli, A. Ulisses, P. Santos, J. Colom, I. Neokosmidis, D. Pujals, R. Spada, A. Garcia, S. Figerola, R. Nejabati, and D. Simeonidou, "Deploying a Novel 5G-Enabled Architecture on City Infrastructure for Ultra-High Definition and Immersive Media Production and Broadcasting," *IEEE Transactions on Broadcasting*, vol. 65, no. 2, pp. 392–403, Jun. 2019.

[323] "OpenAirInterface – 5G software alliance for democratising wireless innovation." [Online]. Available: https://openairinterface.org/

[324] "Mosaic5G." [Online]. Available: https://mosaic5g.io/

[325] N. Nikaein, C.-Y. Chang, and K. Alexandris, "Mosaic5G: agile and flexible service platforms for 5G research," *ACM SIGCOMM Computer Communication Review*, vol. 48, no. 3, pp. 29–34, Sep. 2018.

[326] "O-ran alliance." [Online]. Available: https://www.o-ran.org/

[327] J. Nightingale, Q. Wang, J. M. A. Calero, E. Chirivella-Perez, M. Ulbricht, J. A. Alonso-Lopez, R. Preto, T. Batista, T. Teixeira, M. J. Barros *et al.*, "Qoe-driven, energy-aware video adaptation in 5g networks: The selfnet self-optimisation use case." *IJDSN*, vol. 12, no. 1, pp. 7 829 305–1, 2016.

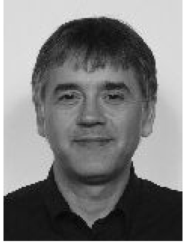
[328] C.-Y. Chang, N. Nikaein, O. Arouk, K. Katsalis, A. Ksentini, T. Turletti, and K. Samdanis, "Slice Orchestration for Multi-Service Disaggregated Ultra-Dense RANs," *IEEE Communications Magazine*, vol. 56, no. 8, pp. 70–77, Aug. 2018.

Roberto Viola is with the Department of Digital Media, Vicomtech. He received his Computer and Telecommunication Engineering degree in 2014 and an advanced degree in Telecommunication Engineering in 2016 from University of Cassino and Southern Lazio (Italy). He is a Research Associate at Vicomtech involved in projects dealing with multimedia services and network infrastructure. He is working on his PhD degree on video content distribution in 5G networks at the University of the Basque Country (UPV/EHU).

Angel Martin is with the Department of Digital Media, Vicomtech. He received his PhD degree (2018) from UPV/EHU and his engineering degree (2003) from University Carlos III. He developed in Prodys an standard MPEG-4 AVC/H.264 codec for DSP (2003-2005). He worked in media streaming and encoding research (2005-2008) in Telefonica. He worked in the fields of smart environments and ubiquitous and pervasive computing (2008-2010) in Innovalia. Currently, he is working in Vicomtech in multimedia services and 5G infrastructures projects.

Mikel Zorrilla is head of the Digital Media department, Vicomtech. He received his Telecommunication Engineering degree (2007) from Mondragon Unibertsitatea, and an advanced degree (2012) and PhD degree (2016) in Computer Science from UPV/EHU. He has participated in many international research projects, such as MediaScape or Hbb4All European Projects. Previously he has held positions at IK4-Ikerlan as an assistant researcher (2002-2006) and at Deusto Business School (2014) as an associate professor in media.

Jon Montalbán received the M.S. Degree (2009) and PhD (2014) in Telecommunications Engineering from the University of the Basque Country (UPV/EHU). Since 2009 he is part of the TSR (Radiocommunications and Signal Processing) research group at UPV/EHU, where he is a postdoctoral researcher involved in several projects in the Digital Terrestrial TV broadcasting. His current research interests include digital communications and digital signal processing for mobile reception of broadband wireless communications systems in 5G.



Pablo Angueira received the M.S. Degree (1997) and PhD (2002) in Telecommunications Engineering from UPV/EHU. He joined the Communications Engineering Department at UPV/EHU in 1998, where he is a Full Professor and part of the TSR staff (Signal Processing and Radiocommunications) research group, involved in digital broadcasting technologies with contributions to the ITU-R WG6 and WG3. His main research interests are network planning and spectrum management for digital terrestrial broadcast technologies, and broadcasting in 5G networks.

Gabriel-Miro Muntean (M'04–SM'17) is a Professor with the School of Electronic Engineering, Dublin City University (DCU), Ireland, and co-Director of the DCU Performance Engineering Laboratory. His research interests include quality, performance, and energy saving issues related to multimedia and multiple sensorial media delivery, technology-enhanced learning, and other data communications over heterogeneous networks. He has published over 400 papers in top-level international journals and conferences, authored 4 books and 23 book chapters, and edited 8 additional books. He is an Associate Editor of the IEEE Transactions on Broadcasting, the Multimedia Communications Area Editor of the IEEE Communications Surveys and Tutorials, and a Reviewer for important international journals, conferences, and funding agencies. He was the Project Coordinator for the EU-funded project NEWTON http://www.newtonproject.eu and is the DCU Coordinator for the EU project TRACTION https://www.traction-project.eu.