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

ROBERTO VIOLA, ÁNGEL MARTÍN, and MIKEL ZORRILLA, Fundación Vicomtech, Basque Research and Technology Alliance, Spain
JON MONTALBÁN, Department of Electronic Technology, University of the Basque Country, Spain
PABLO ANGUEIRA, Department of Communications Engineering, University of the Basque Country, Spain
GABRIEL-MIRO MUNTEAN, Performance Engineering Laboratory, School of Electronic Engineering, Dublin City University (DCU), Ireland

Media services must ensure an enhanced user's perceived quality during content playback to attract and retain audiences, especially while the streams are distributed remotely via networks. Thus, media streaming services rely heavily on good and predictable network performance when delivered to a large number of people. Furthermore, as the quality of media content gets high, the network performance demands are also increasing, and meeting them is challenging. Network functions devoted to improving media streaming services become essential to 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 media streaming.

CCS Concepts: • **Information systems** → **Multimedia streaming**; **Computing platforms**; • **Networks** → **Network management**.

Additional Key Words and Phrases: Media streaming, network functions, network virtualization

ACM Reference Format:
Roberto Viola, Ángel Martín, Mikel Zorrilla, Jon Montalbán, Pablo Angueira, and Gabriel-Miro Muntean. 2022. A Survey on Virtual Network Functions for Media Streaming: Solutions and Future Challenges. 1, 1 (October 2022), 35 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In recent years, media streaming traffic has been experiencing a growing trend. Wireless and mobile devices are becoming the primary producers and consumers of rich media content. 5G networks must cope with these new traffic demands by supporting higher bandwidth and reduced latency. It is estimated that 5G connections will handle nearly three times more traffic than current LTE connections by 2023 [61]. New applications involving video streams are gaining relevance and are attracting an increased audience, including vertical industries where media has a residual

Authors' addresses: Roberto Viola, rviola@vicomtech.org; Ángel Martín, amartin@vicomtech.org; Mikel Zorrilla, mzorrilla@vicomtech.org, Fundación Vicomtech, Basque Research and Technology Alliance, Paseo Mikeletegi 57, San Sebastián, Spain, 20009; Jon Montalbán, jon.montalban@ehu.eus, Department of Electronic Technology, University of the Basque Country, Bilbao, Spain, 48013; Pablo Angueira, pablo.angueira@ehu.eus, Department of Communications Engineering, University of the Basque Country, Bilbao, Spain, 48013; Gabriel-Miro Muntean, gabriel.muntean@dcu.ie, Performance Engineering Laboratory, School of Electronic Engineering, Dublin City University (DCU), Dublin, Ireland.

Manuscript submitted to ACM

presence. Examples of application areas that can benefit from advanced media streaming include the Industrial Internet of Things (IIoT), medical equipment, and connected and autonomous vehicles. Moreover, 360-degree and 3D video formats enable support for new services beyond entertainment, i.e., professional applications. They build novel interaction experiences and data navigation on top of several technologies, such as eXtended Reality (XR), Virtual Reality (VR), and Augmented Reality (AR) [82]. 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 significant increase in rich media streaming traffic, fueled mainly by the global COVID-19 pandemic. This pandemic is transforming users' habits to access the Internet [86, 140] and media content consumption [83, 126]. 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. Therefore, it is implementing an Agenda for Action to push an emergency response to the pandemic, aiming at Internet access extension and boosting its capacity [5, 6].

All the factors mentioned above inevitably influence the evolution of all services, especially affecting 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 traffic crosses networks working on a best-effort basis where no details regarding packet delivery (e.g., time) are guaranteed. Therefore, best-effort networked-transmitted media traffic may result in a lower user's Quality of Experience (QoE). A practical example of this QoE degradation is stalls or artifacts during media playback on player devices. Employing Content Delivery Networks (CDNs) is the most common solution to prevent adverse 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 service coverage. CDN price is decreasing, but the overall cost for the content provider is increasing, as the traffic from/to CDN is growing. [172].

Beyond CDNs, more advanced solutions such as load balancers, transcoders, and transraters based on Network Function Virtualization (NFV) technologies [105] are investigated to support media streaming services. NFV allows the deployment of Virtual Network Functions (VNF) devoted to empowering network abilities when delivering media streaming traffic in an optimized and cost-effective manner [68, 117].

In order to take advantage of using VNFs in a media streaming context, three significant aspects must be considered during their design. First, VNFs should monitor objective operational parameters of the network, such as throughput or latency, representative of QoS of the media streaming dataflows, which directly influence user's satisfaction. However, QoS metrics do not perfectly map to user experience, as users' perceived quality is highly subjective. VNFs should also consider QoE, which compiles subjective evaluation elements, including rewards for playback quality and smoothness and penalties for image freezes and unstable or low quality [28, 125]. Secondly, VNFs should give the CP more control over the network. CP should be able to shape the network traffic and allocate resources by establishing business rules for VNF deployment and life cycle management. These rules allow balancing the trade-offs between network resources and business costs [107], i.e., Capital Expenditure (CAPEX) and Operational expenditures (OPEX), so they are highly relevant. Last, VNFs should also consider energy efficiency. The volume, complexity, and real-time nature of the media streaming traffic have an evident impact on the energy consumption of the network and devices managing the media content. An optimized streaming delivery should reduce energy consumption by turning on or off the VNFs, depending on the demand for network resources, at any time. By putting together real-time QoS and QoE metrics, business policies, and energy efficiency constraints, the user and the CP can benefit from optimized VNFs for media streaming. VNFs can provide the user with a target QoE and reduce business costs and energy consumption to what is strictly necessary to achieve it.

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

- The paper provides an extensive overview of different technologies and protocols in the context of VNFs for media streaming;
- The review analyses state-of-the-art network media functions by classifying them into major categories, i.e., media casting, transcoding, and content caching, and providing a comparison among them;
- The paper presents technologies considered by the telecommunications industry as key enablers for the next generation networks and discusses remaining challenges, including emerging aspects such as data security, energy efficiency, and business models for new network assets;
- The survey provides an overview of international initiatives in the media streaming field, including research activities and projects.

The rest of the paper is structured as follows. First, Section 2 presents the objective of this work in the context of related surveys. Section 3 contains an overview of media streaming technologies and protocols. Section 4 identifies and describes the VNFs employed to date to enhance the performance of media streaming services. The employment of VNFs in media streaming has been growing in the last few years as the attention increases on media distribution over the newly deployed 5G networks [123]. VNFs are intrinsically designed to follow the principles of modularity, interoperability, scalability, and flexibility. Media streaming can leverage VNFs to enable higher network capacity and stability, media traffic optimization, and other performance-related advantages. In this sense, several network solutions to enable performance-driven management of the resources are already being employed and/or investigated with the final aim of increasing media streaming performance. These solutions include efficient use of network resources and end device capabilities [80] during their involvement in the streaming service. To better describe these performance-driven VNFs for media streaming, we classify them into major categories and provide a comparison among them. Nevertheless, even if in the current deployment of 5G networks, the VNFs have a significant role, as 5G aims to have a fully virtualized network deployment, there are still several open issues and challenges that need to be addressed in the future. Thus, Section 5 presents the current challenges in the virtualization process of network functions inside 5G and beyond to assess the open issues and scientific research directions. It discusses the future of VNFs to enable an improved media streaming process and enhanced user experience. Finally, we highlight some valuable international initiatives in Section 6 and assert our conclusions in Section 7.

2 PAPER OBJECTIVES IN THE CONTEXT OF RELATED SURVEYS

Table 1. Summary of Previous Surveys on Virtual Network Functions and Media Streaming.

Survey	Scope and topics	Network virtualization	Domain	Research focus	Year
Skorin-Kapov et al. [190]	QoE assessment and management for HAS, MEC monitoring	SDN/NFV, MEC	Media streaming	QoE-driven architecture	2018
Barakabitze et al. [33]	QoE assessment and management in SDN/NFV, QoE-driven HAS over MEC	SDN/NFV, MEC, Cloud/Fog	Media streaming	SDN routing solutions	2019
Zhang et al. [223]	VNF design considerations	VNF, Cloud/Edge	Agnostic	Virtualization solutions	2019
Fei et al. [85]	NFV/VNF current limitations and future directions	NFV, VNF	Agnostic	Virtualization concepts	2020
This work	NFV architecture, performance-driven VNF	NFV, VNF, MEC	Media streaming	Network media functions	2022

This survey aims 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.

Several surveys on network virtualization have been published in the last few years, in line with the increased interest in virtualization. Zhang et al. [223] provide generic considerations when designing VNFs, while Fei et al. [85] analyze the limitations and identify future research directions. Both surveys focus on the VNF design and utilization to improve current networks, but they do not consider media streaming use cases as they remain domain-agnostic.

Limited to media streaming domain, Skorin-Kapov et al. [190] and Barakabitze et al. [33] discuss using virtualized solutions to assess and improve the QoE. Nevertheless, the former is limited to theoretical concepts for achieving a QoE-driven network architecture and does not provide evidence of implementations. The latter is focused only on routing solutions based on Software-defined networking (SDN) and does not consider VNF approaches.

A comparison of our work with other surveys is shown in Table 1. Our survey describes the NFV architecture and reviews VNFs solutions for the media streaming domain. The VNFs solutions are analyzed and categorized into major categories. Therefore, our survey addresses a more specific scope, as it discusses the relation between the VNFs and media streaming and provides an extensive review of state-of-art solutions concerning media-specific operations, such as media casting, transcoding, and content caching.

A list of acronyms used throughout the paper is presented in Table 2.

3 MEDIA STREAMING OVERVIEW

Before analyzing systems and functions to enhance the media streaming services, it is important to understand the technologies involved in media streaming services to design and implement better VNFs. Media streaming refers to delivering media content (e.g., live television, video clip, etc.) from a streaming server to a streaming client over a particular network infrastructure. The media source can be either live or pre-recorded. In some cases, the CP is also the infrastructure owner employed to stream the content. However, diverse providers and operators have recently entered the market with different roles in the media streaming process, e.g., Akamai, Netflix, etc. Some of them have their own proprietary media streaming solutions. However, the first solutions were based on the Real-time Transport Protocol (RTP) [184] on top of the User Datagram Protocol (UDP) [170], where the Real-time Transport Control Protocol (RTCP) [184] was employed to monitor network metrics and update the rate control. The choice of UDP was based on its lower latency than the Transmission Control Protocol (TCP) [171], 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 demand for content. In OTT services, the CP streams its content over a public network, and an Internet service provider (ISP) controls the actual content delivery. HTTP adaptive streaming (HAS) [186] technologies were introduced to deliver OTT contents, where 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 the HTTP/TCP stack and use pull-based streaming protocols. Thus, they cross current network infrastructure components, such as Network Address Translation (NAT) and firewall devices [169];

Table 2. List of Acronyms used in the paper.

3GPP	3^{rd} Generation Partnership Project
5G	Fifth Generation
6G	Sixth Generation
AES	Advanced Encryption Standard
AES-CBC	AES block cipher mode
AES-CTR	AES counter mode
ANN	Artificial Neural Network
API	Application Programming Interface
AR	Augmented Reality
C-RAN	Cloud-RAN
CAPEX	Capital Expenditure
CDN	Content Delivery Network
CMAF	Common Media Application Format
CN	Core Network
COTS	Commercial off-the-shelf
CP	Content Provider
DASH	Dynamic Adaptive Streaming over HTTP
DNS	Domain Name System
ETSI	European Telecommunications Standards Institute
FeMBMS	Further enhanced MBMS
FTRL	Follow The Regularized Leader
HAS	HTTP Adaptive Streaming
HLS	HTTP Live Streaming
HTTP	HyperText Transfer Protocol
IaaS	Infrastructure as a Service
IBN	Intent-Based Network
IIoT	Industrial Internet of Things
ISP	Internet Service Provider
ITU	International Telecommunication Union
KPI	Key Performance Indicator
L1	Physical layer
L2	Data link layer
L3	Network layer
L4	Transport layer
L7	Application layer
LL CMAF	Low Latency CMAF
LL-DASH	Low Latency DASH
LL-HLS	Low Latency HLS
LTE	Long-Term Evolution
M3U8	M3U UTF-8 Playlist File
MANO	Management and Orchestration
MBMS	Multimedia Broadcast/Multicast Service
MEC	Multi-access Edge Computing
ML	Machine Learning
MPD	Media Presentation Description
MPTCP	Multipath TCP
Multi-RAT	Multiple Radio Access Technology
NAT	Network Address Translation
NFV	Network Function Virtualization
NFVI	NFV Infrastructure
NFVO	NFV Orchestrator
NS	Network Service
O-RAN	Open RAN
ONAP	Open Network Automation Platform
OPEX	Operational Expenditure
OSM	Open Source MANO
OTT	Over-the-top
PoP	Point of presence
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RNI	Radio Network Information
RNIS	RNI Service
RTCP	Real-time Transport Control Protocol
RTMP	Real-time Messaging Protocol
RTP	Real-time Transport Protocol
RTSP	Real Time Streaming Protocol
SCTP	Stream Control Transmission Protocol
SDN	Software-defined networking
SDR	Software-defined radio
SLA	Service Level Agreement
SON	Self-Organizing Network
SRT	Secure Reliable Transport
STUN	Session Traversal Utilities for NAT
SVA	Streaming Video Alliance
TCP	Transmission Control Protocol
TURN	Traversal Using Relays around NAT
UAV	Unmanned Aerial Vehicle
UDP	User Datagram Protocol
UE	User Equipment
UHD	Ultra-High-Definition
UNESCO	United Nations Educational, Scientific and Cultural Organization
VIM	Virtual Infrastructure Manager
VNF	Virtual Network Function
VNF-CC	VNF Chain Composition
VNF-FG	VNF Forwarding Graph
VNF-FGE	VNF Forwarding Graph Embedding
VNF-PC	VNF Placement and Chaining
VNF-SCH	VNF Scheduling
VNFI	VNF Instance
VNFM	VNF Manager
VOD	Video-on-Demand
VR	Virtual Reality
vRAN	Virtual RAN
WebRTC	Web Real-Time Communication
XR	eXtended Reality

- Reuse and scalability: HAS-based media services can reuse existing CDN systems and caching infrastructures without modifications to reach broad audiences;
- User mobility and device heterogeneity: The dynamic content adaptation-enabled player mechanism is accommodated by all the latest heterogeneous UEs, i.e., smartphones and tablets, which support user mobility.

Figure 1 illustrates the HAS-based adaptive streaming principle. HAS works in pull mode, meaning the client pulls the data from a standard HTTP server, which 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 in bitrate and/or resolution and are split into segments of

Table 3. 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 (≤1sec)	RTCP	encoder	no	no
RTSP	UDP	SDP	packets lost & artifacts	very low (≤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 (≤1sec)	SRT control messages	encoder	no	AES-128 / 265 CTR
WebRTC	UDP, QUIC-ready	SDP	packets lost & artifacts	very low (≤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

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 on the available representations, including HTTP URLs indicating where to download the segments of each representation. During a typical HAS session, the client constantly measures specific parameters, such as available network bandwidth and playback buffer level. When it requests content, the client 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 the segment of an appropriate representation from the server.

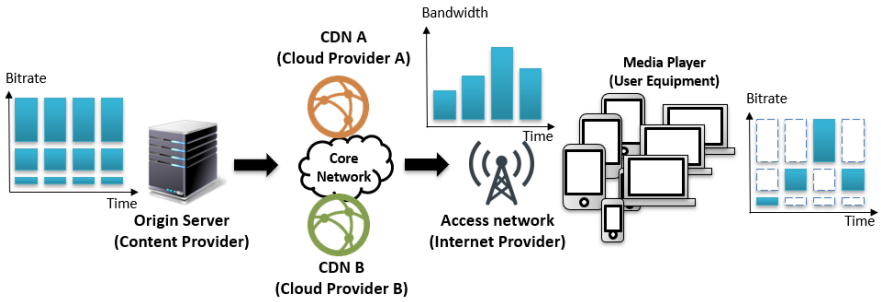

Fig. 1. HAS-based media streaming principle

Besides transport-layer protocols such as RTP, application-layer protocols are also employed in media streaming. Table 3 summarizes different aspects of interest when identifying the best protocol candidates for video streaming.

RTP and Real Time Streaming Protocol (RTSP) [185] perform low latency communications compatible with multicast media streaming. TCP-based Real-time Messaging Protocol (RTMP) [198] enables higher reliability than RTSP, but with higher latency. Secure Reliable Transport (SRT) [188] simplifies the delivery by enabling both push and pull modes of operation. Web Real-Time Communication (WebRTC) [109] enables media streaming through a web browser by exploiting Session Traversal Utilities for NAT (STUN) [215], and Traversal Using Relays around NAT (TURN) [142] protocols provided by third-party servers. SRT and WebRTC increase security by including mandatory encryption support, which is not always required for RTMP. HTTP Live Streaming (HLS) [166] and Dynamic Adaptive Streaming over HTTP (DASH) [191] increase latency due to an internal buffering to overcome network dynamics. In any case, violations of delivery timing could cause stalls and image freezes during the playback if the internal buffer gets empty. To minimize such issues, HAS allows dynamic adaptation mechanisms to track the variability of the network and select the appropriate bitrate. Thus, sudden networking problems are prevented by an alternative bitrate selection from the manifest. Common Media Application Format (CMAF) [114] 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 the integration of HLS and DASH streaming formats since the implementations of Low Latency HLS (LL-HLS) [71] and Low Latency DASH (LL-DASH) [49] still present some differences. Thus, LL-HLS and LL-DASH employ different approaches for HTTP transport and encryption schemes. For instance, a common feature of most HTTP-based solutions is the security by design where different encryption standards protect communications, such as Advanced Encryption Standard (AES) [60] 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 [40] and WebRTC [208], are already evolving and/or being tested with QUIC, a new transport protocol that 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 [133]. There are also proposals to use HAS-based media streaming with protocols such as Stream Control Transmission Protocol (SCTP) [162] and Multipath TCP (MPTCP) [87], which support multihoming and are 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 [67] protocol to combine the benefits of these approaches. However, no HAS-based media delivery solution has used it so far.

4 PERFORMANCE-DRIVEN NETWORK FUNCTIONS

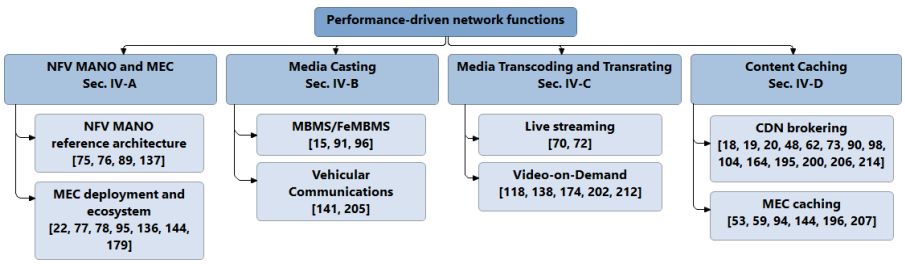

Fig. 2. Performance-driven Network Functions.

VNF brings the architecture, program model, and Application Programming Interfaces (APIs) to deploy specialized media functions as micro-services, exploiting cloud technologies for a smart and agile enhancement of media streaming sessions. This section presents an overview of VNF-based solutions designed to improve media streaming performance. These solutions employ knowledge from network studies and data acquired from live monitoring network traffic. Figure 2 illustrates the significant avenues that performance-driven VNF solutions take. First, we introduce NFV Management and Orchestration (NFV MANO) and Multi-access Edge Computing (MEC), as VNFs rely on these architectures introduced by European Telecommunications Standards Institute (ETSI) and embraced by 5G networks. Then, we discuss the state-of-the-art relevant media-related functions such as media casting, media transcoding, and content caching.

4.1 NFV Management and Orchestration and Multi-access Edge Computing

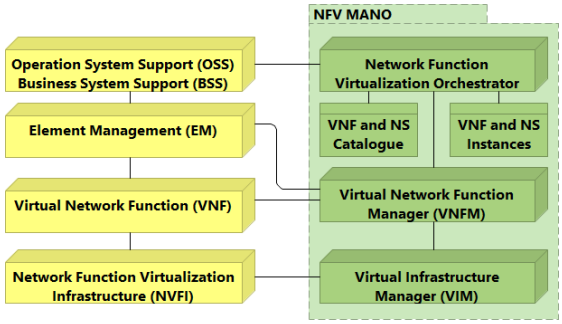

Fig. 3. ETSI NFV MANO architecture.

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 digital transformation. The concept that one network fits all is over. It is time to adapt the network according to applicable resource efficiency and delivery performance trade-offs. The goal is to allow the network management system to coordinate the network in an agile, programmable and efficient way. This vision is being fueled by the transformation of network functions into dynamically controllable and configurable software components. Network functions are virtualized by exploiting cloud technologies and their scalable mechanisms, and their orchestration is done on top of standardized software solutions. 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. The network would independently handle 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 [137] concepts and comes with an NFV MANO architecture [76], 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 deploying a network by providing hardware and software decoupling. Thus, general purpose hardware, referred to as commercial off-the-shelf (COTS) hardware, can be used to run every network function having a software implementation (VNF). Cloud vendors mainly employ this architecture to provide Infrastructure as a Service (IaaS) solutions. Thus, hosting for systems on top of hardware and connectivity setup is performed on demand. VNFs go 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 3. 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 managing 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 constitutes the NS.

Since 5G architecture allows for public and private network deployment, existing NFV MANO-compliant solutions encompass commercial and open-source alternatives for each of the three components. Some examples are Open Source MANO (OSM) [75], whose development is promoted by ETSI, and Open Network Automation Platform (ONAP) [89], 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 a key pillar for the next ones (beyond 5G), and MEC infrastructures [179]. MEC architectures enable context-aware applications. It opens computing infrastructures co-located with the base stations to host services close to the mobile users. It aims to exploit 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 [159] or srsLTE [99], ease the integration of MEC services with RAN systems. These open-source solutions implement the Mobile Packet Core (Evolved Packet Core for LTE, 5G Core for 5G) and the RAN on top of open-source hardware. They enable the deployment, management, and orchestration through NFV MANO of both the mobile packet core [156, 161] and RAN [91]. A RAN deployment through NFV and VNF is usually referred to as virtual RAN (vRAN). Furthermore, vRAN is also evolving towards the concept of Open RAN (O-RAN) [219], 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 ecosystem [93]. In this context, MEC is an NFV MANO-compliant platform that also comes with a specific API to access Radio Network Information (RNI) [77].

Figure 4 shows how network core and edge leveraging virtualization technologies are monitored and orchestrated according to business and technical policies. Different NFVIs can be managed by a unique NFV MANO system and interact with each other through Software Defined Networking (SDN)-enabled communications [46]. The NFV MANO is in charge of deploying VNFs at different NFVIs and managing their life cycle, while the SDN controller configures the forwarding rules of the network to forward the packets between VNFs. Then, a monitoring system collects the performance metrics coming from the different network sections (network core and edge) and processes them to check that the established policies are effective. When considering media streaming, any policy changes to regulate the trade-off between OPEX and QoE involve adjusting the network configuration. It means acting on the NFV MANO system to manage the life cycle of the VNFs and on the SDN controller to update the forwarding rules.

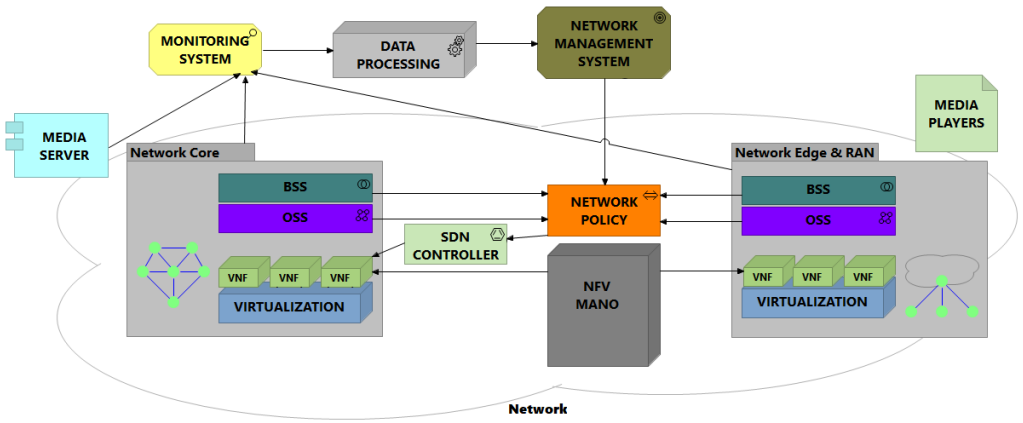

Fig. 4. ETSI NFV architecture applied to media streaming services.

While focusing on the edge architecture, Figure 5 illustrates the MEC components and their interactions with the rest of the RAN and Core Network (CN) building blocks. The MEC host manages the User-plane, while the Control-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, the 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 the 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.

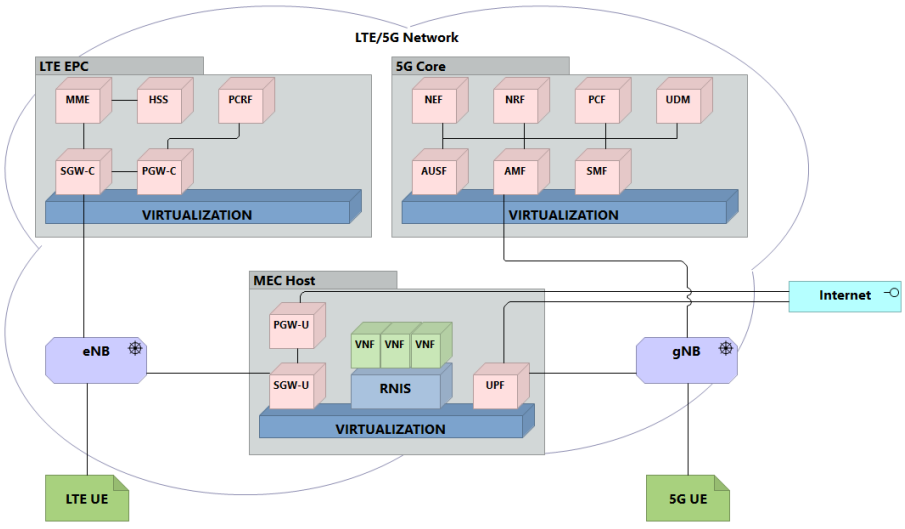

Fig. 5. MEC architecture and connection with RAN and CN.

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 that streaming rate matches network available bandwidth, resulting in higher quality at destination;

- content caching, including storing popular data to help improve high traffic conditions and managing alternative endpoints to balance the data requests.

These network functions perform specialized operations within the media applications to improve network efficiency, reduce bandwidth overheads, favor idle resource allocation to other network flows, and enhance QoE with enforced KPIs according to SLAs.

ETSI includes several media streaming use cases to be considered for MEC deployment [78] to empower traditional media streaming applications. Typical applications are based on the interaction between the remote server (origin server or CDN) and the client, as shown in Figure 6. In this scenario, the MEC platform can host diverse VNFs, which exploit RNI to get a more comprehensive view of the local conditions to enhance media streaming service. In this line, some solutions, such as [95, 136, 144], exploit standard RAN interfaces and data reports to conclude better decisions for media applications.

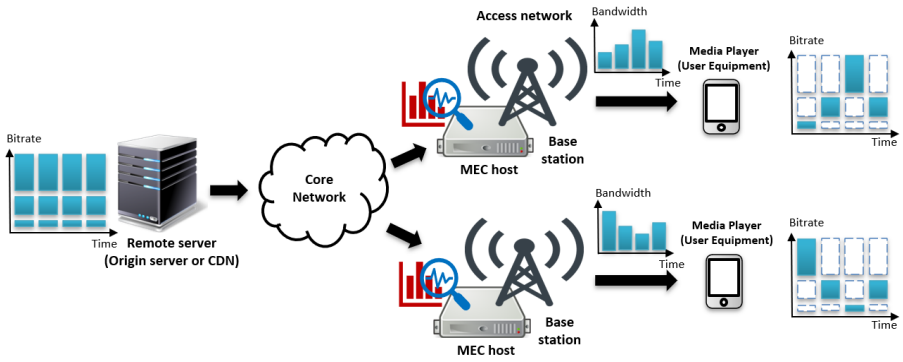

Fig. 6. MEC-powered media streaming.

The following sections analyze how the described core technologies of 5G are applied to expand the network functions with core components for improved media stream delivery, resulting in benefits in terms of enhanced quality and efficient resource utilization. Accordingly, Table 4 compiles and classifies all the research activities exploiting 5G to support performance-aware networking. The classification highlights the main features implemented and 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 real implementation and experimentation. The implemented ones differ in activation and processing approach, as they could operate in a reactive or proactive manner and, in some cases, embed a processing algorithm using classic data analysis or Artificial Neural Network (ANN)-based approaches.

In any case, all the proposed solutions directly impact the performance of the media streaming systems. Performance increase can be provided in terms of QoS and QoE enhancement, more effective business costs (CAPEX and OPEX), and energy saving. However, most of the proposed solutions do not provide specific validation tests for all the performance aspects (QoS/QoE, CAPEX/OPEX, and energy). They mainly limit their analysis to HAS-centric QoE metrics and do not provide insights on applicable cost models, including business aspects or energy footprint evidence. Furthermore, when considering business aspects only, providing cost models is also difficult due to the lack of business models to use new network assets, i.e., NFVI and MEC. Clear business models are required from network operators to improve the evaluation of business costs [22].

Table 4. Performance driven networking for media streams using 5G technologies.

Main feature	Second feature	Activation	Processing approach	Reference	Network feature	Description	Limitations
Casting	-	Not applicable	Not applicable	[96]	FeMBMS	Design of 3GPP architecture for media multicast	To be validated in field test
Casting	-	Reactive	Not applicable	[91]	FeMBMS, VNF, SDR	Virtualization of FeMBMS with SDR setup	No optimization or overhead evaluation
Transcoding	-	Not applicable	Not applicable	[70]	NFV, VNF, 5G Core	Design of centralized virtual transcoder solution at 5G Core	Lack of mathematical analysis
Transcoding	-	Reactive	ANN	[72]	VNF, MEC	On-the-fly transcoder at the network edge	Cost performance trade-off not conducted
Transcoding	Caching	Proactive	Classic	[174]	L1 MC-NOMA, MEC	Solution empowered by mulitcarrier non-orthogonal multiple access	Slicing of computation without practical evaluation
Transcoding	Caching	Reactive	Classic	[138]	MEC, VNF	Transcoding and cache location in virtualized edge infrastructures	Latency factor not considered
Transcoding	Caching	Reactive / Proactive	Classic	[202]	MEC, VNF	Transcoding and cache location when content popularity is known (proactive) or not (reactive)	User experience factor not considered
Transcoding	Caching	Proactive	Classic	[118, 212]	MEC, VNF	Transcoding and cache location based on known content popularity	Lack of mobility
CDN Brokering	-	Reactive	Not applicable	[18–20]	L7	Proprietary solution for selection of CDN vendor at startup	Dynamic costs and experience excluded
CDN Brokering	-	Reactive	Classic	[98, 164, 200]	L3 DNS	Performance-driven solution based on DNS resolution	Prevention schemes not discussed
CDN Brokering	-	Not applicable	Not applicable	[48, 90, 195, 214]	L3 DNS	Design of CDN-ISP collaborative solutions	No practical evaluation
CDN Brokering	-	Proactive	ANN	[206]	L7, L3	Solution for proactive CDN selection employing ANN algorithm to forecast network metrics	Small real world field test
CDN Brokering	-	Reactive	Not applicable	[62, 73, 104]	L7	Cloud solution for cost-effective CDN switching	Fairness among users not discussed
Caching	CDN Brokering	Reactive / Proactive	Classic	[207]	L7, L3, MEC	Statistical solution for CDN selection (reactive) and content caching (proactive)	Popularity of content not considered
Caching	-	Not applicable	Not applicable	[53]	VNF, Orchestration	Design of virtual CDNs for media distribution	No evaluation or comparison
Caching	-	Proactive	Classic	[196]	MEC, SDR	Solution at edge exploiting radio network information	Lack of mobility
Caching	Fair QoE	Reactive	Classic	[144]	MEC, SDR	Solution at edge exploiting radio network information	Delay factor not evaluated
Caching	Fair QoE	Reactive	Classic	[94]	MEC, SDR	Solution at edge exploiting radio network information and content popularity	User experience factor not considered
Caching	-	Proactive	ANN	[59]	MEC	Solution for proactive caching employing ANN technologies to predict popularity	Computational workload and time not considered

4.2 Media Casting

For massive delivery of common data synchronously at once, the broadcast is still much more efficient than unicast communications widely employed by cellular networks. That is why 3^{rd} Generation Partnership Project (3GPP) introduced the Multimedia Broadcast/Multicast Service (MBMS) specification in Long-Term Evolution (LTE) Release 9. Later, it 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 [96]. Furthermore, Release 16 includes feedback for increased reliability [15].

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. It would need 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, keep the manifests unencrypted even for encrypted contents, allowing simple processing to parse them by intermediaries, such as CDNs or MEC systems, for efficient and intelligent media delivery.

This architecture brings three significant benefits by 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 the volume of users since everyone is consuming the same broadcast signal;
(2) Optimal fidelity, as the network can deliver the maximum resolution (bitrate representation) to all the audience;
(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. It means no bitrate or resolution changes to track time-varying network conditions and no freezes to refill the buffer.

This approach is possible thanks to virtualization and softwarization paradigms to RAN technologies, making vRAN and the containerization of some RAN network functions such as FeMBMS feasible [91].

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 the safety of autonomous driving [141]. In these applications, media flows are important as the vehicles get fitted with more camera-like sensors capturing the environment and exchanging the raw/compressed data or processed insights/summaries from onboard computer vision systems [205].

4.3 Media Transcoding and Transrating

As summarized in Table 3, HAS technologies, such as DASH or HLS, are widely employed and need the provision of several representations meaning different resolutions and bitrates [138]. Thus, VNF-based transcoders are being developed under international funding initiatives to empower different use cases, e.g., live 3D media streaming [70] or automotive [16]. 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 lower bitrate variants (transrating) at the edge. It would reduce overheads in the core to send all the possible media variants. To this end, the media transcoding at the edge is essential [202], stressing the fronthaul capacity and requiring Cloud-RANs (C-RANs) or MEC systems 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 to be performed at a resource-constrained MEC server. Then, it means a challenge for delay-sensitive services. Here, different works deal with the optimal position 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 the edge hosts. To overcome this challenge, a mechanism for optimal request forwarding which respects the resource limitations and minimizes serving latency is required [174]. In [212], 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 a transcoding process, other approaches consider different algorithms for reactive or proactive planning [202]. In this case, the dynamics significantly impact the reaction time and forecast range. These aspects are minimized using a segment duration in the HAS stream, which favors steady short-term conditions as changes come on a segment duration basis.

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

Finally, it is essential to underline that these solutions are often linked to caching strategies, as both transcoding and caching can be executed at the edge to better match the local conditions, patterns, and demand features. Therefore, joint transcoding processing and caching strategies are designed [118, 138, 174, 202, 212]. In [72], the authors only transcode the content on-the-fly if the content is not cached.

4.4 Content Caching

4.4.1 CDN brokering. Caching is the most employed network function to improve performance when accessing online content, particularly media streaming. A CDN is the most popular network solution to provide caching capabilities in this context. 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 so that users can quickly access it from nearby locations. The employment of CDN services by the CPs increased in the last few years as the number of CDN vendors increased. Furthermore, major CPs also moved to multi-CDN strategies to provide more reliable service while streaming their content. Thus, an improved service also generates more satisfaction among the customers. Nevertheless, the different CDNs employed can differ from one CP to another. Static selection of the CDN when a streaming session starts is the easiest and most widely employed solution among CPs. In 2012, this strategy was used by Netflix [20] and Hulu [18], with big similarities [19]. In both cases, they were using three different CDN vendors. They used to map the player device to a CDN depending on its location or the subscriber when the streaming session starts. Moreover, the CDN never changes during the streaming session, even when the performances decrease. Other solutions include client-side CDN selection [164] or Domain Name System (DNS)-based solutions [200]. The client has a privileged position to measure end-to-end QoS metrics (network bandwidth and latency) when choosing the CDN. However, this approach has the disadvantage of producing 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 decrease performance [98], affecting the user's satisfaction.

In recent years, other network caching solutions are also raising to empower delivery. The same Netflix changed its streaming strategies. It developed and deployed an in-house CDN, called Open Connect [48], to reduce the dependency on CDN vendors and streaming costs. Moreover, Open Connect is meant to be also run inside the ISP infrastructure, i.e., closer to the user, to guarantee better performances in terms of network bandwidth and latency [69]. Open Connect also helps Netflix and other CPs have in-house solutions to control better the resources enabled for the streaming

session and reduce the costs. Anyway, it requires a significant investment to have such a solution that could not be affordable by small CPs.

The Streaming Video Alliance (SVA) is a joint initiative that works on different aspects of media streaming and aims to standardize the employed protocols and technologies. Its membership includes some major world-wide content production and delivery agents. Among its activities, the SVA Open Caching Working Group [195] 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's Open Connect but with the advantage of being standardized.

Other collaborations between CDN and ISP are proposed in the literature. In [90], ISP provides the CDN provider with information concerning geographical user distribution and allows the CDN provider to allocate server resources inside the ISP network. The authors of [214] 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 for improvements. CDN Brokering [42] is proposed to make CDN utilization in a multi-CDN environment more effective. 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 [206] 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 [73], proposed by the European Broadcasting Union (EBU). Similar solutions are also provided by Citrix [62] and Haivision [104].

4.4.2 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 underperforming and cannot support the demanded traffic. Moreover, it features local edge caching to reduce network traffic. Recurrent content is downloaded and cached once for every player. In [53], a similar MEC cache is proposed for empowering the delivery.

With a deeper integration with RAN interfaces, in [196] and [144], the MEC cache is improved by exploiting RNI. The media segments and representations are selectively cached depending on the network state. In [94], 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 [59] 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 a higher probability of being requested by the users.

Table 5. Comparison of performance driven network function categories.

Network media function	Scalability	Video adaptation	Processing resources
Media Casting	High	No	Low
Transcoding	Low	Yes	High
CDN Brokering	Medium	Limited to HAS	Low
Edge Caching	High	Limited to HAS	Low

To sum up, the proposed performance-driven network functions for media streaming belong to four main categories, corresponding to specific actions on the content: Media Casting, Transcoding, CDN Brokering, and Edge Caching. Table 5 compares the four categories regarding scalability, video adaptation (resolution and bitrate), and required processing resources. Media Casting and Edge Caching are the best-performing solutions in terms of scalability. They are designed to receive the content from a remote server or CDN and then serve it to the UE(s). As the content is served from RAN or Edge Network, the traffic inside the Network Core is reduced (i.e., there is no redundant traffic). Another advantage is their low processing resource consumption, as they only have to forward packets and do not process the content. CDN Brokering also has low resource processing, like Casting and Edge caching, but its scalability is limited, as it does not reduce network traffic. CDN Brokering selects the remote CDN that the UE has to connect to at any time only. Therefore, the redundancy in Network Core traffic remains, and all the workload has to be absorbed by the remote CDN. Finally, when considering video adaptation (resolution and bitrate), Transcoding solutions are the best ones, as they allow to generate a personalized video representation in real-time. Each client can receive a representation according to its display capabilities and the network state. Nevertheless, this flexibility in video adaptation comes at the cost of using increased processing resources for the transcoding operation. Video adaptation is also available with CDN Brokering and Edge Caching. However, it is possible only when using HAS-based technologies and is limited to the pre-encoded video representations of the content at the CDN.

5 CHALLENGES OF VIRTUAL NETWORK FUNCTIONS FOR MEDIA STREAMING

VNF solutions play a significant role in successfully deploying 5G networks. It is backed by evidence, especially for supporting rich media applications such as multimedia streaming, as described in Section 4. However, VNF applications still require some challenges and open issues to be addressed, as shown in Figure 7. This section discusses and classifies these challenges around some key features studied concerning 5G networks and presents the open issues in the context of the 6G networks roadmap.

5.1 Self-Organizing Networks

Table 6. 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

Agile deployment and life cycle management of VNFs exploiting an 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 [79] represents a next step to achieve a fully virtualized and automated network. SON 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 [26]. SON capabilities were initially meant to be included as add-on features of LTE, as 3GPP Release 8 started defining LTE

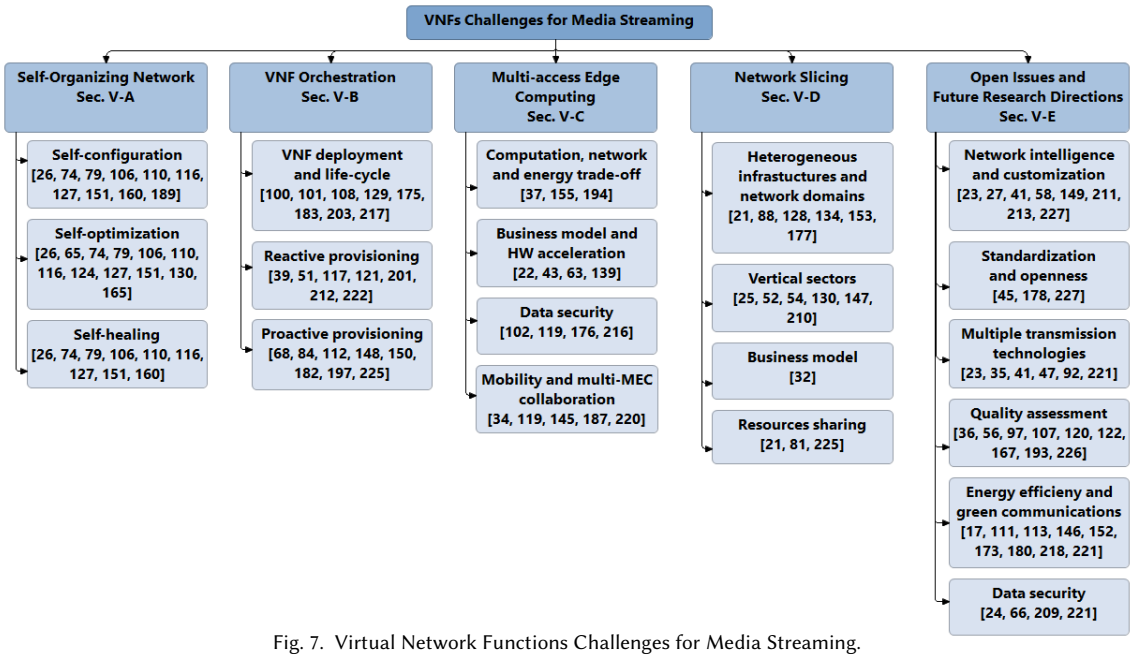

Fig. 7. Virtual Network Functions Challenges for Media Streaming.

and already set the basis for SON concepts and requirements [110]. However, SON is expected to enhance 5G network management by providing automation to cope with increasing network complexity [151].

Specifically, in the media streaming context, SON should provide the necessary 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 [26]. 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 network performance is near optimal. They include real-time network monitoring and performance metrics processing to apply enhanced operational parameters proactively. 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 network faults, failures, or any operational range violations 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 the current setup, the appropriate countermeasure is applied to reestablish the desired network performance.

All these SON flavors need actionable data to process decision-making algorithms. Therefore, it is crucial to collect and exploit network data. Current networks are ready to probe and provide a considerable amount of data. However, specialized intelligence must be deployed within the network to infer valuable and helpful information from the collected data [116]. Such information helps take 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, ML

techniques will become prominent, even if selecting the suitable algorithm is not trivial and depends on the considered use case [127, 151]. Table 6 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 and included in their commercial hardware equipment. Some examples are HCL's SON [106], Nokia's EdenNet [160] and Ericsson's SON Optimization Manager [74].

In any case, SON systems need to have a more comprehensive view of the delivered traffic beyond the metrics from the network functions and including the service domain. Operated SON policies are usually steered by network statistics rather than application characteristics. In [143], the authors analyze network metrics of selected network topologies, such as bandwidth and latency, when the network nodes are generating media streaming traffic. It proposes an ML-powered algorithm to select the appropriate network topology that copes with the demanded network resources for the media traffic. Nevertheless, the communication dynamics of applications delivered on top of the network have an impact on network performance. Thus, the authors of [165] propose to design an application-driven SON to widen the view with both network performance and the user's QoE metrics. When considering media streaming applications, data are available from network functions in the path and playback devices. Thus, data exploitation inside a SON-enabled system needs further investigation, as the multi-domain data exploitation is still underexplored. In [131], the authors classify the network traffic into four classes, including one for interactive games and video telephony, and propose to take into account the QoS requirements of each class to enable a flexible mobility management (handover) scheme. In [189], the authors introduce a self-organizing Unmanned Aerial Vehicle (UAV)-based communication framework for media streaming. It aims at dynamically selecting the network configuration to achieve better network capabilities in terms of throughput, end-to-end transmission delay, packet re-transmission, and packet loss during video transmission. UAV-based communication is also considered in [65] to design an ML-based scheduling solution to react to the changeable networking conditions and make the best decisions to stream Live Ultra-High-Definition (UHD) Video Streaming from UAV to ground users. Finally, the authors of [124] propose a SON-enabled video transcoder to be deployed within the network to let the video propagate over the network by exploiting knowledge about the state of the network.

5.2 VNF Orchestration

The orchestration on top of SDN programs and coordinates different network functions to support specific applications and services further. The specialized network functions in the media context include servers, load balancers, caches, transcoders, transraters, and encryptors. These functions are containerized to make media functions ready to be deployed on top of cloud infrastructures and instantiated depending on resource policies. Thus a media system gains virtualization and flexible scaling of cloud resources available as NS at the network core and edge. This evolution is essential to reduce superfluous OPEX and lead to shorter time-to-market and lower CAPEX. In this context, DevOps and micro-services platforms gain relevance to accomplishing the Function-as-a-Service (FaaS) paradigm. Here, the ambition is to meet media-intensive use cases, where an on-demand VNF instantiation and deployment enables an elastic infrastructure that enforces the performance and restrains costs [203].

ETSI NFV MANO architecture brings several functionalities at different layers operating different systems. For example, MANO manages the allocation of virtualized resources of the NFVI layer using the VIM system, the VNFs live-cycle at the VNF layer driving the VNFM, and orchestration scaling or deleting NS instances steering the NFVO [100]. MANO is expected to support optimal media delivery in flexible networks toward handling target QoE and cost trade-offs. However, this has not been fully achieved yet. The main problems of orchestration, still open and under the focus of different research lines, are related to resource allocation and the instances placement.

First, deploying a VNF over a distributed platform requires allocating 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 [108]. 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 mainly includes: VNF Chain Composition (VNF-CC) and VNF Placement and Chaining (VNF-PC), 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. VNF-PC aims for the optimal placement and the required instances of VNFs needed to deploy Service Function Chains (SFCs) while optimizing the cost of resource provisioning. How the traffic flows between VNFs is also described through the definition of VNF Forwarding Graphs (VNF-FG). Thus, any network service can be considered composed of a set of VNFs and VNF-FGs. Each VNF executes a small function of the entire application or service [217]. It aims to find appropriate resources and locations to allocate the VNFs in NFVI. At this stage, resource selection and optimization must be accomplished concerning the specific constraints defined by SLA [183]. Finally, VNF-SCH determines how to schedule the processing operations of the deployed VNFs [175].

Specifically, when media VNFs come into play, the dynamic allocation and provisioning of VNF resources are essential to mitigate or even prevent QoS/QoE violations when bottlenecks at some network path arise. When the VNFs are already deployed and running, the required resources may vary during their life cycle, as they depend on the user demand of the running function provided by the VNFs. Allocated resources could be optimized to fit the user's variable demand. Increasing or decreasing the allocated resources means the VNFs also need to scale up/down dynamically. Then, an efficient orchestration and automation of the VNFs require supporting this dynamic allocation of resources. 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. Here, transcoder [129], and transrating [101] systems are instantiated with VNF technologies triggered by a changeable demand applying Machine Learning (ML) techniques to automate the monitoring and actuation.

Second, the placement of caches, transcoders, and transraters has been profoundly studied to solve this NP-hard problem. Here, [51] analyzes the QoE feedback to on-the-fly conclude an optimal placement of transcoding VNFs. Others such as [201, 212] go beyond and optimize at the same time caching, transcoding, and transrating in two timescales, short-term and long-term, to favor steadiness.

However, most of the proposed solutions include two main limitations, the lack of a cost function to add effectiveness to the equation and the inability to prevent issues firing actions to avoid SLA violations dynamically.

Thus, the related literature often employs an over-provisioning strategy, where the allocated resources for each VNF are larger than required. This approach is inefficient in terms of OPEX and energy consumption generated by the allocated resources which are not employed. It 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. Some approaches optimize the cache and transcoding, including simple cost models concurrently [39, 121]. They formulate a constrained optimization problem to minimize each user request's total caching, computing, and bandwidth utilization. In [222], the authors also include the hardware acceleration costs. Moreover, [117] meets the placement and chaining of VNFs for media cache, including VNF instantiation, migration, hosting, and routing costs.

Furthermore, the solutions widely involve a reactive provisioning and allocation approach, which means changing the allocated resources to react when traffic and/or demand change. In [68], the authors design a proactive algorithm for VNF placement and allocation of caching and protection VNFs. They also aim to minimize the OPEX by considering the bandwidth and host resource consumption trade-offs under diverse workload variations. Most of the literature

on generic VNF orchestration driven by traffic demand prediction employs ANN algorithms [182, 197]. However, the application of such algorithms in practical solutions is limited, being mostly theoretical. Among the most innovative solutions proposed, [84] describes a novel Follow The Regularized Leader (FTRL) online algorithm for VNF provisioning, which handles workload fluctuations. The solution in [148] employs an ANN algorithm to predict future resource requirements for each VNF contributing to a network service. The authors of [225] propose the POLAR algorithm, combining online learning and optimization to proactively provision resources with VNFs provisioning. In contrast, the VNFs chaining in a network service is ignored. In [112], proactive failure recovery is proposed when considering VNF deployed at distributed edge computing nodes. Finally, the authors of [150] 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.

5.3 Multi-access Edge Computing (MEC)

MEC represents a novel technological solution integrated into 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 the network edge. There are many aspects that 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 crucial application of MEC solutions, as ETSI considers it one of MEC core use cases [78]. MEC platforms can host edge services to empower media streaming applications traditionally 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, transcoding, and content caching.

More specifically, MEC resources are exploited by both the server and clients to offload computation tasks [37, 155]. Offloading server tasks reduces network traffic and latency, as the processing is performed close to UEs. Computation offloading at the MEC is necessary to enable use cases where the remote server has a very high delay and/or the client has not enough computing capabilities. 6G potential use cases include resource-expensive and delay-sensitive applications such as augmented and virtual reality [154]. In any case, it is important to note that MEC resources are shared between different service providers, but how the resources are distributed among different service providers is still undefined. The authors of [155] propose to allocate MEC resources proportionally to each service provider's demanded resources and payment. On the other side, if it is the UE that offloads its tasks to the MEC host, it reduces not only the device computation load but also the power consumption on the device, as computing-intensive tasks heavily impact the battery duration. In [37], 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 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 [194], optimization of the allocation of both computing and network resources is discussed while considering 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 [22]. 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 make the cost model more complex. Resource accounting and monitoring must also be determined 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 [139]. Some works suggest employing Field-Programmable Gate Array

(FPGA) approaches instead of GPU solutions due to their reduced price and power consumption [43, 63], but this possibility is again underexplored.

Regarding accessible information at MEC, the API to communicate with RNIS [77] has been recently standardized, and its development is ongoing [30, 199]. It means that services running at the MEC host cannot be further optimized. When RNIS implementations are available, edge services will embed more complex and precise algorithms (classic or ML models) to exploit RNI to improve their operations and overall system performance. However, improved capabilities due to RNI exploitation raise some security concerns regarding managing information at MEC hosts, an aspect that needs further investigation [102, 119]. 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 [176, 216]. Use cases like surveillance and CCAM may also include videos containing sensitive data. Thus, video and data stream anonymization [44] is an important matter to consider for further improving MEC-based solutions.

Mobility remains another primary concern and is becoming critical, as the explosion in availability and type of mobile devices (e.g., smartphones and tablets) involves an increasing number of UEs to be served. In the same way that 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 multi-MEC cooperation should be addressed to guarantee seamless session migration across MEC servers [119, 187]. Moreover, this seamless migration becomes critical as delays may raise security concerns in some use cases. An example is represented by CCAM use cases, where the video is streamed between two self-driving cars that make real-time decisions [34].

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. Balancing the cost of MEC-based caching and transcoding and provision of high user QoE is an essential direction for future research [119]. Moreover, finding suitable locations where MEC instances should be deployed becomes relevant, as it may affect the fulfillment 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 [145]. Finally, content caching mechanisms in the network have been studied both at the core and at the edge, but a convergent solution has not been identified yet. Caching solutions that integrate both core and edge caching could improve network performance in terms of energy consumption, network throughput, latency, and user QoE [220].

5.4 Network Slicing

Network slicing [21, 88] is introduced in 5G networks as a solution involving several virtual/logical networks (slices) on top of a shared physical network, where each virtual/logical network delivers the traffic generated by a specific service [153, 177]. To achieve it, it provides network and computing resources across different networks. It can be considered that a network slice is associated with a set of network resources and VNFs, which that slice can provide. In this context, NFV, together with SDN, plays an important role, especially in deploying and managing network slices [163, 224]. NFV 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. The introduction of network slicing is necessary since a best-effort network cannot guarantee that appropriate network resources are offered in each use case. With network slicing,

different sets of proprieties can be established, each one associated with specific network resources and supporting relevant use cases. For instance, use cases requiring high throughput, i.e., on-demand video streaming, are logically separated from use cases requiring low latency, i.e., real-time video streaming or mission-critical services, and are supported by different slices. Each network slice provides support for service in terms of 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 [128, 134].

Network slicing at the air interface refers to partitioning physical radio resources (physical layer or L1) into subsets of several physical resources. Each subset associated with a different network slice is mapped 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 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 [25] through the deployment of two different slices to split audio and video transmissions, as they have different requirements in terms of network throughput. In [210], 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 [54] to enable real-time communication between hospital staff and patients. In [52] and [147], applications of network slicing for Vehicle-to-Everything (V2X) services are investigated. Different use cases are considered in a vehicle, including those related to safety and traffic efficiency, autonomous or teleoperated driving, media & entertainment, and remote diagnostics. Each use case requires different latency, throughput, and communication reliability requirements. Consequently, different network slices with different configurations are required on the same physical network of resources. The authors of [130] 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 slicing operations relate to the exploitation of resources the network operator provides. However, the effects of changes in network operators' business models for operating network slicing are unknown [32]. 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 needs further investigation [225]. An end-to-end network slice implies that slice segments potentially stretch across different administrative domains. There are two requirements 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 control on the slice [21]. Finally, network slicing leverages algorithms to accommodate applications with diverse requirements over the same physical network. Thus, complex algorithms are necessary for deciding how to allocate efficiently, manage, and control the physical resources shared across diverse slices [204]. 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 [81].

5.5 Open Issues and Future Research Directions

The benefits of virtualization for media streaming communications will become 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 [115].

ML-powered network intelligence to manage NFV and VNFs is only partially achieved in 5G networks, but it will also be a key factor for the future 6G networks [227]. The concept of Intent-Based Network (IBN) [213] means employing ML solutions to transform business intents into a 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 [58, 149]. The advantages of an intelligent-native network are two-fold. First, the network can analyze the user's behavior in real-time and autonomously learn his needs to predict his future behavior. Then, the user's information can be employed for network customization to achieve a user-centric network [211]. Second, the network can meet 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 [41]. Moreover, the global COVID-19 pandemic is accelerating the digital transformation of multiple and heterogeneous verticals, such as the development of new services for smart cities and innovation in the eHealth, including telemedicine, medical and thermal imaging, and robotics for medicine practice [23, 27].

Openness is also essential to achieve flexible networks and services [227]. An open network platform and interfaces (O-RAN, NFV MANO, SDN, etc.) allow interconnection and interoperability of different vendors, which is essential for sharing physical infrastructure. Thus, agents of diverse vertical industries may deploy their private physical infrastructure and manage it through NFV MANO solutions and SDN controllers independent from public networks operated by mobile network operators [178]. The standardization process will continue in the following years to fulfill the remaining gaps and guarantee interoperability of heterogeneous implementations of open network solutions [45].

The cooperation of different physical networks will also attract attention. Multiple Radio Access Technology (multi-RAT) aims to employ different access networks to improve the overall connectivity [92]. Its application to improve media streaming is already being investigated [35, 47]. However, new transmission solutions based on space, UAV-based and underwater communications will be integrated with terrestrial ones [23, 41]. Flexibility to operate the network at any level (spectrum/band, physical and MAC, etc.), despite the different involved technologies, will be imperative [221].

In a media streaming context, performance assessment focuses on evaluating 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 involve quantifiable values to track and monitor a streaming session, i.e., video resolution and bitrate, or a related factor that can influence it, i.e., network bandwidth and latency, at any network device. Moreover, there exists lots of different metrics to describe the media system under different points of view, e.g., ISP, CP or user's view, including both objective (QoS) [56, 120] and subjective ones (QoE) [122, 193, 226]. In the context of the current 5G networks and devices, metrics collection is not so easy, as the heterogeneity in devices and virtualization solutions increases the complexity. In a virtualized and heterogeneous network environment, media and network-related metrics to describe the media system's performance should be further expanded to characterize this new context where the media streaming solution is deployed. Monitoring systems represent an essential part of a virtualized architecture. Monitoring the resources means defining metrics that apply to the virtualization environment and describing the state of the NFVI and the running VNFI. It allows to manage the physical and virtual resources more efficiently, including those allocated for encoding operations [36, 167] or content caching [97]. Furthermore, the network flexibility, guaranteed by NFV and SDN, raises the discussion on business costs monitoring [107]. The CP may be able to enforce its business policy to balance the QoS/QoE and business costs trade-off. Thus, the definition of business metrics is necessary to allow CP's business decisions when running each component of its media streaming solution.

Energy efficiency and green communications [113] are envisioned to enable more sustainable networking [173]. Energy efficiency concerns are also relevant for media streaming services [17, 146]. 6G network will be developed by taking into account self-sustainability devices and solutions [111]. Here, low-power wireless devices could harvest energy from the available high-power radio waves [221]. Thus, battery-free implementations will be an interesting topic to be further explored in different use cases. Clearly, IoT applications will benefit the most from battery-free implementations [218]. However, the high presence of video content traffic also suggests their use for media communications to support a self-sustainable 6G network [152, 180].

Finally, the growth of network and media traffic will have consequences on security. Critical media use cases, e.g., eHealth applications [24], and autonomous driving systems [66], need to be secured with security mechanisms that will complement the conventional cryptography-based ones. Increasing security will be assured with the design of cross-layer algorithms to protect the transferred information [209, 221].

6 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. Recently, the European Commission has funded numerous research projects to develop and implement VNFs for different research scenarios and vertical industries. Table 7 summarizes the most relevant actions. The project list includes initiatives targeting generic architectural design (i.e., CogNet [7], SELFNET [11] and SliceNet [11]), activities building testbed environments and pilot environments for use case definition and testing (FLAME [14], SoftFIRE [13] and 5GTango) [4], projects targeting specific application verticals and developing required functionalities (5G-Media [1], 5Growth [3], 5GCity [2]) and finally international software communities to provide open-source platforms (OpenAirInterface Software Alliance [10], Mosaic5G [8] and O-RAN Alliance [9]).

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 [55]. Among others, it proposed the smart integration of state-of-the-art technologies in NFV. One of the outcomes is presented in [157], where the SELFNET framework preserves the network's health 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 [57]. The project proposed an integrated network management, control and orchestration framework

Table 7. Major SDN/NFV related research activities.

Project	Period	Area	References
CogNet (Building an Intelligent System of Insights and Action for 5G Network Management)	2015-2018	Architecture	[7, 31, 38]
SELFNET (Framework for Self-Organized Network Management in Virtualized and Software Defined Networks)	2015-2018	Architecture	[11, 55]
SliceNet (End-to-End Cognitive Network Slicing and Slice Management Framework in Virtualised Multi-Domain, Multi-Tenant 5G Networks)	2017-2020	Architecture	[12, 181, 210]
SoftFIRE (Software Defined Networks and Network Function Virtualization Testbed within FIRE+)	2016-2018	TestBeds	[13, 132]
FLAME (Facility for Large-scale Adaptive Media Experimentation)	2017-2020	TestBeds	[14, 103]
5GTango (5G Development and validation platform for global industry-specific network services and Apps)	2017-2020	TestBeds	[4, 168, 192]
5G-Media (Programmable edge-to-cloud virtualization fabric for the 5G Media industry)	2017-2020	Application Verticals	[1, 29, 50]
5Growth (5G-enabled Growth in Vertical Industries)	2019-2021	Application Verticals	[3, 135]
5GCity (A Distributed Cloud and Radio Platform for 5G Neutral Hosts)	2017-2020	Application Verticals	[2, 64]
OpenAirInterface Software Alliance	2014-	Development Platforms	[10, 159]
Mosaic5G	2016 -	Development Platforms	[8, 158]
O-RAN Alliance	2018-	Development Platforms	[10, 219]

and applied the concept to various use cases. One of those cases related to multimedia health services is described in [210], where demanding QoS requirements (i.e., latency) must be fulfilled. The network intelligence topic is tackled by CogNet, a project focused on realizing the well-known control loop MAPE (Monitor, Analyze, Plan and Execute) with ML 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 [31], [38].

A second group of projects aimed at creating platforms and testbed environments where specific use cases, applications, algorithms, and interoperability solutions could be designed and validated. FLAME stands out in this area as a facility for large-scale experiments in the Adaptive Media field. Since 2017, FLAME has hosted different proposals [103] to offload video content proactively to the edge of the network on an SDN/NFV environment. FLAME tests include AR applications and smart video surveillance for aiding impaired citizens. SoftFIRE is another testbed environment to experiment with VNF services and applications in SDN/NFV. SoftFIRE aims to assess the maturity level of solutions in programmability, interoperability, and security and show how they can support the full potential of these properties in a real-world case [132]. Finally, 5GTango puts the focus on network flexible programmability [168] by providing software development kits (SDKs) [192]. 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 verticals through specific pilots: advanced manufacturing and immersive media [168].

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 project's primary outcomes was the 5GCity Orchestration Platform, which supported the NFV MANO model. In [64], the authors demonstrate that the virtualized platform could 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 [1] 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 [29]. The use cases include teleimmersive gaming, mobile journalism, and UHD content distribution [50]. 5Growth [3] supports diverse industry verticals by 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 [135].

Finally, OpenAirInterface Software Alliance [10], Mosaic5G [8], and O-RAN Alliance [9] 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 [10] 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 [159]. Mosaic5G [8] develops a set of 5G software solutions and has already hosted experiments targeting low latency MEC services, orchestration solutions, and programmable RANs [158]. O-RAN Alliance [9] is pushing the standardization and the development of the O-RAN. RAN industry is moving towards open, intelligent, virtualized, and fully interoperable RAN [219].

7 CONCLUSIONS

The popularity of media streaming services is constantly growing due to an increasing number of users, the diversity of rich media experiences, e.g., online gaming, VR/AR applications, and the utilization beyond entertainment services, i.e., in professional application domains. The latest smart mobile devices also have an essential role in the success of media streaming, as their processing and rendering capabilities support streaming content at very high resolutions, e.g., UHD or 4K. Consequently, media streaming traffic represents a substantial share of the total Internet traffic and, more importantly, 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 achieving the best trade-off with business costs and energy efficiency. 5G networking is bringing new possibilities to deploy intelligent network functions, which monitor the media streaming service through live and objective metrics and boost it in real-time. Under the 5G umbrella, NFV will have a prominent role in virtualizing 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 with the design and implementation of a stable VNF. VNF-based media streaming functions monitors and collects performance metrics to tune their configuration to enhance the processed media streaming sessions. Beyond that base features limited to session quality, advanced solutions adapt the VNF life cycle deployment and management, dealing with cost-performance trade-offs and the balanced setup when infrastructures with several VNFs come into play. Moreover, network traffic monitoring and analysis allow the creation of models to approximate the behavior of the network and predict future network events to take actions proactively. 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, transcoding, and content caching, can be employed at any network segment. Thanks to the NFV MANO architecture, the deployment of VNFs is interoperable and can be automatically operated and orchestrated across a virtualized infrastructure. However, the exploitation of media streaming VNFs is not limited to the Network Core, but they can also be run at MEC hosts. 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 main venues where the research will focus in the next few years are the achievement of dynamic VNF deployment and orchestration, complete MEC integration, and network slicing. Long-term research will also address the 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 capacity. Green communications and security will also be significant concerns, as future networks should reduce their environmental impact and guarantee the security of the processed information. In conclusion, VNFs represent an essential enabler in improving media streaming services. However, despite the research done under international initiatives pushing 5G and network virtualization, several research challenges still exist and provide opportunities for further research activities.

ACKNOWLEDGMENTS

This work was fully supported by the 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 870610 for the TRACTION project. The support of the Science Foundation Ireland (SFI) Research Centres Programme grant 12/RC/2289_P2 (Insight SFI Research Centre for Data Analytics) and EJ/GV Grant IT1436-22 are also gratefully acknowledged.

REFERENCES

[1] [n.d.]. 5G-Media (Programmable Edge-to-Cloud Virtualization Fabric for the 5G Media Industry). http://www.5gmedia.eu/
[2] [n.d.]. 5GCity – A distributed cloud & radio platform for 5G Neutral Hosts. https://www.5gcity.eu/
[3] [n.d.]. 5Growth (5G-enabled Growth in Vertical Industries). https://5growth.eu/
[4] [n.d.]. 5GTango (5G Development and validation platform for global industry-specific network services and Apps). https://5gtango.eu/
[5] [n.d.]. *The affordability of ICT services 2020.* https://www.itu.int/en/ITU-D/Statistics/Documents/publications/prices2020/ITU_A4AI_Price_Briefing_2020.pdf
[6] [n.d.]. *Broadband Commission Agenda for Action for Faster and Better Recovery.* https://www.broadbandcommission.org/COVID19/Pages/default.aspx
[7] [n.d.]. CogNet (Building an Intelligent System of Insights and Action for 5G Network Management). http://www.cognet.5g-ppp.eu/
[8] [n.d.]. Mosaic5G. https://mosaic5g.io/
[9] [n.d.]. O-RAN Alliance. https://www.o-ran.org/
[10] [n.d.]. OpenAirInterface – 5G software alliance for democratising wireless innovation. https://openairinterface.org/
[11] [n.d.]. SELFNET (Framework for Self-Organized Network Management in Virtualized and Software Defined Networks). https://selfnet-5g.eu/
[12] [n.d.]. SliceNet (End-to-End Cognitive Network Slicing and Slice Management Framework in Virtualised Multi-Domain, Multi-Tenant 5G Networks). https://slicenet.eu/
[13] [n.d.]. SoftFIRE (Software Defined Networks and Network Function Virtualization Testbed within FIRE+). https://www.softfire.eu/
[14] 2017. FLAME (Facility for Large-scale Adaptive Media Experimentation). https://www.ict-flame.eu/
[15] 3GPP. 2020. Overall Description of LTE-Based 5G Broadcast; Version 16.0.0; Technical Report (TR) 36.976. *3rd Generation Partnership Project (3GPP)* (2020).
[16] 5GINFIRE. [n.d.]. *D2.2-5GinFIRE Experimental Infrastructure Architecture and 5G Automotive Use Case(update).* https://bscw.5g-ppp.eu/pub/bscw.cgi/d302168/D2-2-5GINFIRE_Experimental_Infrastructure_Architecture_and_5G_Automotive_Use_Case-v1-0.pdf
[17] Hatem Abou-Zeid and Hossam S Hassanein. 2013. Predictive green wireless access: Exploiting mobility and application information. *IEEE wireless communications* 20, 5 (2013), 92–99.

[18] Vijay Kumar Adhikari, Yang Guo, Fang Hao, Volker Hilt, and Zhi-Li Zhang. 2012. A tale of three CDNs: An active measurement study of Hulu and its CDNs. In *2012 Proceedings IEEE INFOCOM Workshops*. IEEE, 7–12.

[19] Vijay K Adhikari, Yang Guo, Fang Hao, Volker Hilt, Zhi-Li Zhang, Matteo Varvello, and Moritz Steiner. 2014. Measurement study of Netflix, Hulu, and a tale of three CDNs. *IEEE/ACM Transactions on Networking* 23, 6 (2014), 1984–1997.

[20] Vijay Kumar Adhikari, Yang Guo, Fang Hao, Matteo Varvello, Volker Hilt, Moritz Steiner, and Zhi-Li Zhang. 2012. Unreeling netflix: Understanding and improving multi-cdn movie delivery. In *2012 Proceedings IEEE INFOCOM*. IEEE, 1620–1628.

[21] Ibrahim Afolabi, Tarik Taleb, Konstantinos Samdanis, Adlen Ksentini, and Hannu Flinck. 2018. Network slicing and softwarization: A survey on principles, enabling technologies, and solutions. *IEEE Communications Surveys & Tutorials* 20, 3 (2018), 2429–2453.

[22] Ejaz Ahmed, Arif Ahmed, Ibrar Yaqoob, Junaid Shuja, Abdullah Gani, Muhammad Imran, and Muhammad Shoaib. 2017. Bringing computation closer toward the user network: Is edge computing the solution? *IEEE Communications Magazine* 55, 11 (2017), 138–144.

[23] Muhammad Waseem Akhtar, Syed Ali Hassan, Rizwan Ghaffar, Haejoon Jung, Sahil Garg, and M Shamim Hossain. 2020. The shift to 6G communications: vision and requirements. *Human-centric Computing and Information Sciences* 10, 1 (2020), 1–27.

[24] Yazan Al-Issa, Mohammad Ashraf Ottom, and Ahmed Tamrawi. 2019. eHealth cloud security challenges: a survey. *Journal of healthcare engineering* 2019 (2019).

[25] Pol Alemany, L Juan, Ana Pol, Anton Roman, Panagiotis Trakadas, Panagiotis Karkazis, Marios Touloupou, Evgenia Kapassa, Dimosthenis Kyriazis, Thomas Soenen, et al. 2019. Network slicing over a packet/optical network for vertical applications applied to multimedia real-time communications. In *2019 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*. IEEE, 1–2.

[26] Osianoh Glenn Aliu, Ali Imran, Muhammad Ali Imran, and Barry Evans. 2012. A survey of self organisation in future cellular networks. *IEEE Communications Surveys & Tutorials* 15, 1 (2012), 336–361.

[27] Zaheer Allam and David S Jones. 2021. Future (post-COVID) digital, smart and sustainable cities in the wake of 6G: Digital twins, immersive realities and new urban economies. *Land Use Policy* 101 (2021), 105201.

[28] Mohammed Alreshoodi and John Woods. 2013. Survey on QoE\QoS correlation models for multimedia services. *arXiv preprint arXiv:1306.0221* (2013).

[29] Federico Alvarez, David Breitgand, David Griffin, Pasquale Andriani, Stamatia Rizou, Nikolaos Zioulis, Francesca Moscatelli, Javier Serrano, Madeleine Keltsch, Panagiotis Trakadas, T. Khoa Phan, Avi Weit, Ugur Acar, Oscar Prieto, Francesco Iadanza, Gino Carrozzo, Harilaos Koumaras, Dimitrios Zarpalas, and David Jimenez. 2019. An Edge-to-Cloud Virtualized Multimedia Service Platform for 5G Networks. *IEEE Transactions on Broadcasting* 65, 2 (June 2019), 369–380. https://doi.org/10.1109/TBC.2019.2901400

[30] Sagar Arora, Pantelis A Frangoudis, and Adlen Ksentini. 2019. Exposing radio network information in a MEC-in-NFV environment: the RNISaaS concept. In *2019 IEEE Conference on Network Softwarization (NetSoft)*. IEEE, 306–310.

[31] Haytham Assem, Lei Xu, Teodora Sandra Buda, and Declan O'Sullivan. 2016. Machine learning as a service for enabling Internet of Things and People. *Personal and Ubiquitous Computing* 20, 6 (Nov. 2016), 899–914. https://doi.org/10.1007/s00779-016-0963-3

[32] Alcardo Alex Barakabitze, Arslan Ahmad, Rashid Mijumbi, and Andrew Hines. 2020. 5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges. *Computer Networks* 167 (2020), 106984.

[33] Alcardo Alex Barakabitze, Nabajeet Barman, Arslan Ahmad, Saman Zadtootaghaj, Lingfen Sun, Maria G Martini, and Luigi Atzori. 2019. QoE management of multimedia streaming services in future networks: a tutorial and survey. *IEEE Communications Surveys & Tutorials* 22, 1 (2019), 526–565.

[34] Hamid R Barzegar, Nabil El Ioini, Claus Pahl, et al. 2020. Service continuity for ccam platform in 5g-carmen. In *2020 International Wireless Communications and Mobile Computing (IWCMC)*. IEEE, 1764–1769.

[35] Pavlos Basaras, George Iosifidis, Stepan Kucera, and Holger Claussen. 2020. Multicast Optimization for Video Delivery in Multi-RAT Networks. *IEEE Transactions on Communications* 68, 8 (2020), 4973–4985.

[36] Sreejata Basu. 2020. *Cloud Video Encoding vs On-Premise: Pros, Cons and Beyond.* https://www.muvi.com/blogs/cloud-video-encoding-vs-on-premise.html

[37] Michael Till Beck, Sebastian Feld, Andreas Fichtner, Claudia Linnhoff-Popien, and Thomas Schimper. 2015. ME-VoLTE: Network functions for energy-efficient video transcoding at the mobile edge. In *2015 18th International Conference on Intelligence in Next Generation Networks*. IEEE, 38–44.

[38] Imen Grida Ben Yahia, Jaafar Bendriss, Alassane Samba, and Philippe Dooze. 2017. 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)*. IEEE, Paris, 252–259. https://doi.org/10.1109/ICIN.2017.7899421

[39] Ilias Benkacem, Tarik Taleb, Miloud Bagaa, and Hannu Flinck. 2018. Performance benchmark of transcoding as a virtual network function in CDN as a service slicing. In *2018 IEEE Wireless Communications and Networking Conference (WCNC)*. 1–6. https://doi.org/10.1109/WCNC.2018.8377402

[40] Divyashri Bhat, Amr Rizk, and Michael Zink. 2017. Not so QUIC: A performance study of DASH over QUIC. In *Proceedings of the 27th workshop on network and operating systems support for digital audio and video*. 13–18.

[41] Jagadeesha R Bhat and Salman A Alqahtani. 2021. 6G Ecosystem: Current Status and Future Perspective. *IEEE Access* 9 (2021), 43134–43167.

[42] Alexandros Biliris, Chuck Cranor, Fred Douglis, Michael Rabinovich, Sandeep Sibal, Oliver Spatscheck, and Walter Sturm. 2002. CDN brokering. *Computer Communications* 25, 4 (2002), 393–402.

[43] Saman Biookaghazadeh, Ming Zhao, and Fengbo Ren. 2018. Are fpgas suitable for edge computing?. In {*USENIX*} *Workshop on Hot Topics in Edge Computing (HotEdge 18)*.

[44] Pascal Birnstill, Daoyuan Ren, and Jürgen Beyerer. 2015. A user study on anonymization techniques for smart video surveillance. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 1–6.

[45] Leonardo Bonati, Michele Polese, Salvatore D'Oro, Stefano Basagni, and Tommaso Melodia. 2020. Open, Programmable, and Virtualized 5G Networks: State-of-the-Art and the Road Ahead. *Computer Networks* 182 (2020), 107516. https://doi.org/10.1016/j.comnet.2020.107516

[46] Michel S Bonfim, Kelvin L Dias, and Stenio FL Fernandes. 2019. Integrated NFV/SDN architectures: A systematic literature review. *ACM Computing Surveys (CSUR)* 51, 6 (2019), 1–39.

[47] Sem Borst, Aliye Özge Kaya, Doru Calin, and Harish Viswanathan. 2017. Dynamic path selection in 5G multi-RAT wireless networks. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 1–9.

[48] Timm Böttger, Felix Cuadrado, Gareth Tyson, Ignacio Castro, and Steve Uhlig. 2018. Open connect everywhere: A glimpse at the internet ecosystem through the lens of the netflix cdn. *ACM SIGCOMM Computer Communication Review* 48, 1 (2018), 28–34.

[49] Nassima Bouzakaria, Cyril Concolato, and Jean Le Feuvre. 2014. 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, 92–97.

[50] David Breitgand, Avi Weit, Stamatia Rizou, David Griffin, Ugur Acar, Gino Carrozzo, Nikolaos Zioulis, Pasquale Andriani, and Francesco Iadanza. 2018. Towards Serverless NFV for 5G Media Applications. In *Proceedings of the 11th ACM International Systems and Storage Conference (SYSTOR '18)*. Association for Computing Machinery, Haifa, Israel, 118. https://doi.org/10.1145/3211890.3211916

[51] Utku Bulkan, Muddesar Iqbal, and Tasos Dagiuklas. 2018. Load-Balancing for Edge QoE-Based VNF Placement for OTT Video Streaming. In *2018 IEEE Globecom Workshops (GC Wkshps)*. 1–6. https://doi.org/10.1109/GLOCOMW.2018.8644214

[52] Claudia Campolo, Antonella Molinaro, Antonio Iera, and Francesco Menichella. 2017. 5G network slicing for vehicle-to-everything services. *IEEE Wireless Communications* 24, 6 (2017), 38–45.

[53] Gino Carrozzo, Francesca Moscatelli, Gabriel Solsona, Oscar Prieto Gordo, Madeleine Keltsch, and Martin Schmalohr. 2018. Virtual CDNs over 5G networks: scenarios and requirements for ultra-high definition media distribution. In *2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 1–5.

[54] Alberto Huertas Celdrán, Manuel Gil Pérez, Félix J García Clemente, Fabrizio Ippoliti, and Gregorio Martínez Pérez. 2019. Dynamic network slicing management of multimedia scenarios for future remote healthcare. *Multimedia Tools and Applications* 78, 17 (2019), 24707–24737.

[55] Alberto Huertas Celdran, Manuel Gil Perez, Felix J. Garcia Clemente, and Gregorio Martinez Perez. 2017. Enabling Highly Dynamic Mobile Scenarios with Software Defined Networking. *IEEE Communications Magazine* 55, 4 (April 2017), 108–113. https://doi.org/10.1109/MCOM.2017.1600117CM

[56] Dan Chalmers and Morris Sloman. 1999. A survey of quality of service in mobile computing environments. *IEEE Communications surveys* 2, 2 (1999), 2–10.

[57] Chia-Yu Chang, Navid Nikaein, Osama Arouk, Kostas Katsalis, Adlen Ksentini, Thierry Turletti, and Konstantinos Samdanis. 2018. Slice Orchestration for Multi-Service Disaggregated Ultra-Dense RANs. *IEEE Communications Magazine* 56, 8 (Aug. 2018), 70–77. https://doi.org/10.1109/MCOM.2018.1701044

[58] Dawei Chen, Yin-Chen Liu, BaekGyu Kim, Jiang Xie, Choong Seon Hong, and Zhu Han. 2020. Edge computing resources reservation in vehicular networks: A meta-learning approach. *IEEE Transactions on Vehicular Technology* 69, 5 (2020), 5634–5646.

[59] Yu Chen, Yong Liu, Jingya Zhao, and Qinghua Zhu. 2020. Mobile edge cache strategy based on neural collaborative filtering. *IEEE Access* 8 (2020), 18475–18482.

[60] Pete Chown. 2002. *Advanced encryption standard (AES) ciphersuites for transport layer security (TLS)*. Technical Report. RFC 3268, June.

[61] Cisco. 2020. *Cisco Annual Internet Report (2018–2023) White Paper*. https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[62] Citrix. [n.d.]. *Intelligent Traffic Management*. https://www.citrix.com/products/citrix-intelligent-traffic-management/

[63] Ian Colbert, Jake Daly, Ken Kreutz-Delgado, and Srinjoy Das. 2021. A Competitive Edge: Can FPGAs Beat GPUs at DCNN Inference Acceleration in Resource-Limited Edge Computing Applications? *arXiv preprint arXiv:2102.00294* (2021).

[64] Carlos Colman-Meixner, Hamzeh Khalili, Konstantinos Antoniou, Muhammad Shuaib Siddiqui, Apostolos Papageorgiou, Antonino Albanese, Paolo Cruschelli, Gino Carrozzo, Luca Vignaroli, Alexandre Ulisses, Pedro Santos, Jordi Colom, Ioannis Neokosmidis, David Pujals, Rita Spada, Antonio Garcia, Sergi Figerola, Reza Nejabati, and Dimitra Simeonidou. 2019. Deploying a Novel 5G-Enabled Architecture on City Infrastructure for Ultra-High Definition and Immersive Media Production and Broadcasting. *IEEE Transactions on Broadcasting* 65, 2 (June 2019), 392–403. https://doi.org/10.1109/TBC.2019.2901387

[65] Ioan-Sorin Comşa, Gabriel-Miro Muntean, and Ramona Trestian. 2020. An innovative machine-learning-based scheduling solution for improving live UHD video streaming quality in highly dynamic network environments. *IEEE Transactions on Broadcasting* 67, 1 (2020), 212–224.

[66] Jin Cui, Lin Shen Liew, Giedre Sabaliauskaite, and Fengjun Zhou. 2019. A review on safety failures, security attacks, and available countermeasures for autonomous vehicles. *Ad Hoc Networks* 90 (2019), 101823.

[67] Quentin De Coninck and Olivier Bonaventure. 2020. Multipath Extensions for QUIC (MP-QUIC). *draft-deconinck-quic-multipath-06* (2020).

[68] Mouhamad Dieye, Shohreh Ahvar, Jagruti Sahoo, Ehsan Ahvar, Roch Glitho, Halima Elbiaze, and Noel Crespi. 2018. CPVNF: Cost-efficient proactive VNF placement and chaining for value-added services in content delivery networks. *IEEE Transactions on Network and Service Management* 15, 2 (2018), 774–786.

[69] Trinh Viet Doan, Vaibhav Bajpai, and Sam Crawford. 2020. A Longitudinal View of Netflix: Content Delivery over IPv6 and Content Cache Deployments. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 1073–1082.

[70] Alexandros Doumanoglou, Nikolaos Zioulis, David Griffin, Javier Serrano, Truong Khoa Phan, David Jiménez, Dimitrios Zarpalas, Federico Alvarez, Miguel Rio, and Petros Daras. 2018. A system architecture for live immersive 3D-media transcoding over 5G networks. In *2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 11–15.

[71] Kerem Durak, Mehmet N Akcay, Yigit K Erinc, Boran Pekel, and Ali C Begen. 2020. Evaluating the Performance of Apple's Low-Latency HLS. In *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 1–6.

[72] Sunny Dutta, Tarik Taleb, Pantelis A Frangoudis, and Adlen Ksentini. 2016. On-the-fly QoE-aware transcoding in the mobile edge. In *2016 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 1–6.

[73] EBU. [n.d.]. *Eurovision FLOW*. https://tech.ebu.ch/docs/groups/flow/Eurovision%20Flow%20Brochure.pdf

[74] Ericsson. [n.d.]. *SON Optimization Manager*. https://www.ericsson.com/en/portfolio/digital-services/automated-network-operations/network-management/son-optimization-manager

[75] ETSI. [n.d.]. *Open Source MANO (OSM)*. https://osm.etsi.org/

[76] ETSI. 2014. *ETSI GS NFV-MAN 001: Network Functions Virtualisation (NFV); Management and Orchestration*. https://www.etsi.org/deliver/etsi_gs/nfv-man/001_099/001/01.01.01_60/gs_nfv-man001v010101p.pdf

[77] ETSI. 2017. *ETSI GS MEC 012: Mobile Edge Computing (MEC); Radio Network Information API*. https://www.etsi.org/deliver/etsi_gs/MEC/001_099/012/01.01.01_60/gs_MEC012v010101p.pdf

[78] ETSI. 2018. *ETSI GS MEC 002: Multi-access Edge Computing (MEC): Phase 2: Use Cases and Requirements*. https://www.etsi.org/deliver/etsi_gs/MEC/001_099/002/02.01.01_60/gs_MEC002v020101p.pdf

[79] ETSI. 2020. *ETSI TS 128 313: 5G;Self-Organizing Networks (SON) for 5G networks*. https://www.etsi.org/deliver/etsi_ts/128300_128399/128313/16.00.00_60/ts_128313v160000p.pdf

[80] Kristian Evensen, Tomas Kupka, Haakon Riiser, Pengpeng Ni, Ragnhild Eg, Carsten Griwodz, and Pål Halvorsen. 2014. Adaptive media streaming to mobile devices: challenges, enhancements, and recommendations. *Advances in Multimedia* 2014 (2014).

[81] Zubair Md Fadlullah, Fengxiao Tang, Bomin Mao, Nei Kato, Osamu Akashi, Takeru Inoue, and Kimihiro Mizutani. 2017. State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems. *IEEE Communications Surveys & Tutorials* 19, 4 (2017), 2432–2455.

[82] Ching-Ling Fan, Wen-Chih Lo, Yu-Tung Pai, and Cheng-Hsin Hsu. 2019. A survey on 360 video streaming: Acquisition, transmission, and display. *ACM Computing Surveys (CSUR)* 52, 4 (2019), 1–36.

[83] Thomas Favale, Francesca Soro, Martino Trevisan, Idilio Drago, and Marco Mellia. 2020. Campus traffic and e-Learning during COVID-19 pandemic. *Computer Networks* 176 (2020), 107290.

[84] Xincai Fei, Fangming Liu, Hong Xu, and Hai Jin. 2018. Adaptive VNF scaling and flow routing with proactive demand prediction. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 486–494.

[85] Xincai Fei, Fangming Liu, Qixia Zhang, Hai Jin, and Hongxin Hu. 2020. Paving the way for NFV acceleration: A taxonomy, survey and future directions. *ACM Computing Surveys (CSUR)* 53, 4 (2020), 1–42.

[86] Anja Feldmann, Oliver Gasser, Franziska Lichtblau, Enric Pujol, Ingmar Poese, Christoph Dietzel, Daniel Wagner, Matthias Wichtlhuber, Juan Tapiador, Narseo Vallina-Rodriguez, Oliver Hohlfeld, and Georgios Smaragdakis. 2020. The Lockdown Effect: Implications of the COVID-19 Pandemic on Internet Traffic. *Proceedings of the ACM Internet Measurement Conference* (2020), 1–18. https://doi.org/10.1145/3419394.3423658

[87] Alan Ford, Costin Raiciu, Mark Handley, Sebastien Barre, Janardhan Iyengar, et al. 2011. Architectural guidelines for multipath TCP development. *IETF, Informational RFC* 6182 (2011), 2070–1721.

[88] Xenofon Foukas, Georgios Patounas, Ahmed Elmokashfi, and Mahesh K Marina. 2017. Network slicing in 5G: Survey and challenges. *IEEE Communications Magazine* 55, 5 (2017), 94–100.

[89] Linux Foundation. [n.d.]. *Open Network Automation Platform (ONAP)*. https://www.onap.org/

[90] Benjamin Frank, Ingmar Poese, Yin Lin, Georgios Smaragdakis, Anja Feldmann, Bruce Maggs, Jannis Rake, Steve Uhlig, and Rick Weber. 2013. Pushing CDN-ISP collaboration to the limit. *ACM SIGCOMM Computer Communication Review* 43, 3 (2013), 34–44.

[91] Alvaro Gabilondo, Javier Morgade, Roberto Viola, Pablo Angueira, and Jon Montalbán. 2020. Realising a vRAN based FeMBMS Management and Orchestration Framework. In *2020 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 1–7.

[92] Olga Galinina, Alexander Pyattaev, Sergey Andreev, Mischa Dohler, and Yevgeni Koucheryavy. 2015. 5G multi-RAT LTE-WiFi ultra-dense small cells: Performance dynamics, architecture, and trends. *IEEE Journal on Selected Areas in Communications* 33, 6 (2015), 1224–1240.

[93] Liljana Gavrilovska, Valentin Rakovic, and Daniel Denkovski. 2020. From Cloud RAN to Open RAN. *Wireless Personal Communications* (2020), 1–17.

[94] Chang Ge, Ning Wang, Severin Skillman, Gerry Foster, and Yue Cao. 2016. QoE-driven DASH video caching and adaptation at 5G mobile edge. In *Proceedings of the 3rd ACM Conference on Information-Centric Networking*. 237–242.

[95] Francesco Giannone, Pantelis A Frangoudis, Adlen Ksentini, and Luca Valcarenghi. 2020. Orchestrating heterogeneous MEC-based applications for connected vehicles. *Computer Networks* 180 (2020), 107402.

[96] Jordi Joan Gimenez, Jose Luis Carcel, Manuel Fuentes, Eduardo Garro, Simon Elliott, David Vargas, Christian Menzel, and David Gomez-Barquero. 2019. 5G new radio for terrestrial broadcast: A forward-looking approach for NR-MBMS. *IEEE Transactions on Broadcasting* 65, 2 (2019), 356–368.

[97] Kostas Giotis, Yiannos Kryftis, and Vasilis Maglaris. 2015. Policy-based orchestration of NFV services in software-defined networks. In *Proceedings of the 2015 1st ieee conference on network softwarization (netsoft)*. IEEE, 1–5.

[98] Utkarsh Goel, Mike P Wittie, and Moritz Steiner. 2015. Faster web through client-assisted CDN server selection. In *2015 24th International conference on computer communication and networks (ICCCN)*. IEEE, 1–10.

[99] Ismael Gomez-Miguelez, Andres Garcia-Saavedra, Paul D Sutton, Pablo Serrano, Cristina Cano, and Doug J Leith. 2016. srsLTE: An open-source platform for LTE evolution and experimentation. In *Proceedings of the Tenth ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation, and Characterization*. 25–32.

[100] Andres J. Gonzalez, Gianfranco Nencioni, Andrzej Kamisiński, Bjarne E. Helvik, and Poul E. Heegaard. 2018. Dependability of the NFV Orchestrator: State of the Art and Research Challenges. *IEEE Communications Surveys Tutorials* 20, 4 (2018), 3307–3329. https://doi.org/10.1109/COMST.2018.2830648

[101] Yashuang Guo, F. Richard Yu, Jianping An, Kai Yang, Chuqiao Yu, and Victor C. M. Leung. 2020. Adaptive Bitrate Streaming in Wireless Networks With Transcoding at Network Edge Using Deep Reinforcement Learning. *IEEE Transactions on Vehicular Technology* 69, 4 (2020), 3879–3892. https://doi.org/10.1109/TVT.2020.2968498

[102] Gürkan Gür, Pawani Porambage, and Madhusanka Liyanage. 2020. Convergence of ICN and MEC for 5G: Opportunities and Challenges. *IEEE Communications Standards Magazine* 4, 4 (2020), 64–71.

[103] Kay Haensge, Dirk Trossen, Sebastian Robitzsch, Michael Boniface, and Stephen Phillips. 2019. Cloud-Native 5G Service Delivery Platform. In *2019 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*. 1–7. https://doi.org/10.1109/NFV-SDN47374.2019.9040042

[104] Haivision. [n.d.]. *Haivision Lightflow MultiCDN*. https://www.haivision.com/products/haivision-lightflow-multicdn/

[105] Bo Han, Vijay Gopalakrishnan, Lusheng Ji, and Seungjoon Lee. 2015. Network function virtualization: Challenges and opportunities for innovations. *IEEE Communications Magazine* 53, 2 (2015), 90–97.

[106] HCL. [n.d.]. *SON*. https://www.hcltech.com/ERX/Telecom-and-5G/SON

[107] Enrique Hernandez-Valencia, Steven Izzo, and Beth Polonsky. 2015. How will NFV/SDN transform service provider opex? *IEEE Network* 29, 3 (2015), 60–67.

[108] Juliver Gil Herrera and Juan Felipe Botero. 2016. Resource allocation in NFV: A comprehensive survey. *IEEE Transactions on Network and Service Management* 13, 3 (2016), 518–532.

[109] Christer Holmberg, Stefan Hakansson, and G Eriksson. 2015. Web real-time communication use cases and requirements. *Request for Comments (RFC)* 7478 (2015).

[110] Honglin Hu, Jian Zhang, Xiaoying Zheng, Yang Yang, and Ping Wu. 2010. Self-configuration and self-optimization for LTE networks. *IEEE Communications Magazine* 48, 2 (2010), 94–100.

[111] Jie Hu, Qing Wang, and Kun Yang. 2020. Energy self-sustainability in full-spectrum 6G. *IEEE Wireless Communications* 28, 1 (2020), 104–111.

[112] Huawei Huang and Song Guo. 2019. Proactive failure recovery for NFV in distributed edge computing. *IEEE Communications Magazine* 57, 5 (2019), 131–137.

[113] Tongyi Huang, Wu Yang, Jun Wu, Jin Ma, Xiaofei Zhang, and Daoyin Zhang. 2019. A survey on green 6G network: Architecture and technologies. *IEEE Access* 7 (2019), 175758–175768.

[114] K Hughes and D Singer. 2017. Information technology–Multimedia application format (MPEG-A)–Part 19: Common media application format (CMAF) for segmented media. *ISO/IEC* 19 (2017), 23000.

[115] Chih-Lin I, Sławomir Kuklinskí, and Tao Chen. 2020. A Perspective of O-RAN Integration with MEC, SON, and Network Slicing in the 5G Era. *IEEE Network* 34, 6 (2020), 3–4. https://doi.org/10.1109/MNET.2020.9277891

[116] Ali Imran, Ahmed Zoha, and Adnan Abu-Dayya. 2014. Challenges in 5G: how to empower SON with big data for enabling 5G. *IEEE network* 28, 6 (2014), 27–33.

[117] Narjes Tahghigh Jahromi, Somayeh Kianpisheh, and Roch H Glitho. 2018. 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, 19–24.

[118] Qingmin Jia, Renchao Xie, Haijun Lu, Wenbin Zheng, and Hong Luo. 2019. Joint Optimization Scheme for Caching, Transcoding and Bandwidth in 5G Networks with Mobile Edge Computing. In *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*. IEEE, 999–1004.

[119] Xiantao Jiang, F Richard Yu, Tian Song, and Victor CM Leung. 2021. A Survey on Multi-Access Edge Computing Applied to Video Streaming: Some Research Issues and Challenges. *IEEE Communications Surveys & Tutorials* (2021).

[120] Jingwen Jin and Klara Nahrstedt. 2004. QoS specification languages for distributed multimedia applications: A survey and taxonomy. *IEEE multimedia* 11, 3 (2004), 74–87.

[121] Yichao Jin, Yonggang Wen, and Cedric Westphal. 2015. Optimal Transcoding and Caching for Adaptive Streaming in Media Cloud: an Analytical Approach. *IEEE Transactions on Circuits and Systems for Video Technology* 25, 12 (2015), 1914–1925. https://doi.org/10.1109/TCSVT.2015.2402892

[122] Parikshit Juluri, Venkatesh Tamarapalli, and Deep Medhi. 2015. Measurement of quality of experience of video-on-demand services: A survey. *IEEE Communications Surveys & Tutorials* 18, 1 (2015), 401–418.

[123] Madeleine Keltsch, Sebastian Prokesch, Oscar Prieto Gordo, Javier Serrano, Truong Khoa Phan, and Igor Fritzsch. 2018. Remote production and mobile contribution over 5G networks: scenarios, requirements and approaches for broadcast quality media streaming. In *2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 1–7.

[124] Javad I Khan, Seung Su Yang, Qiong Gu, Darsan Patel, Patrick Mail, Oleg Komogortsev, Wansik Oh, and Zhong Guo. 2001. Resource adaptive netcentric systems: A case study with sonet-A self-organizing network embedded transcoder. In *Proceedings of the ninth ACM international conference on Multimedia*. 617–620.

[125] Hyun Jong Kim and Seong Gon Choi. 2010. 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, 1377–1382.

[126] Daniel L King, Paul H Delfabbro, Joel Billieux, and Marc N Potenza. 2020. Problematic online gaming and the COVID-19 pandemic. *Journal of Behavioral Addictions* 9, 2 (2020), 184–186.

[127] Paulo Valente Klaine, Muhammad Ali Imran, Oluwakayode Onireti, and Richard Demo Souza. 2017. A survey of machine learning techniques applied to self-organizing cellular networks. *IEEE Communications Surveys & Tutorials* 19, 4 (2017), 2392–2431.

[128] Zbigniew Kotulski, Tomasz Nowak, Mariusz Sepczuk, Marcin Tunia, Rafal Artych, Krzysztof Bocianiak, Tomasz Osko, and Jean-Philippe Wary. 2017. On end-to-end approach for slice isolation in 5G networks. Fundamental challenges. In *2017 Federated conference on computer science and information systems (FedCSIS)*. IEEE, 783–792.

[129] Harilaos Koumaras, Christos Sakkas, Michail Alexandros Kourtis, Christos Xilouris, Vaios Koumaras, and Georgios Gardikis. 2016. Enabling agile video transcoding over SDN/NFV-enabled networks. In *2016 International Conference on Telecommunications and Multimedia (TEMU)*. 1–5. https://doi.org/10.1109/TEMU.2016.7551916

[130] Fabian Kurtz, Caner Bektas, Nils Dorsch, and Christian Wietfeld. 2018. Network slicing for critical communications in shared 5G infrastructures-an empirical evaluation. In *2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft)*. IEEE, 393–399.

[131] Yeunwoong Kyung and Tae-Kook Kim. 2020. QoS-Aware Flexible Handover Management in Software-Defined Mobile Networks. *Applied Sciences* 10, 12 (2020), 4264.

[132] David Lake, Gerry Foster, Serdar Vural, Yogaratnam Rahulan, Bong-Hwan Oh, Ning Wang, and Rahim Tafazolli. 2017. Virtualising and orchestrating a 5G evolved packet core network. In *2017 IEEE Conference on Network Softwarization (NetSoft)*. 1–5. https://doi.org/10.1109/NETSOFT.2017.8004215

[133] Adam Langley, Alistair Riddoch, Alyssa Wilk, Antonio Vicente, Charles Krasic, Dan Zhang, Fan Yang, Fedor Kouranov, Ian Swett, Janardhan Iyengar, et al. 2017. The quic transport protocol: Design and internet-scale deployment. In *Proceedings of the conference of the ACM special interest group on data communication*. 183–196.

[134] Qian Li, Geng Wu, Apostolos Papathanassiou, and Udayan Mukherjee. 2016. An end-to-end network slicing framework for 5G wireless communication systems. *arXiv preprint arXiv:1608.00572* (2016).

[135] Xi Li, Andres Garcia-Saavedra, Xavier Costa-Perez, Carlos J Bernardos, Carlos Guimarães, Kiril Antevski, Josep Mangues-Bafalluy, Jorge Baranda, Engin Zeydan, Daniel Corujo, et al. 2021. 5Growth: An End-to-End Service Platform for Automated Deployment and Management of Vertical Services over 5G Networks. *IEEE Communications Magazine* 59, 3 (2021), 84–90.

[136] Yue Li, Pantelis A Frangoudis, Yassine Hadjadj-Aoul, and Philippe Bertin. 2016. A mobile edge computing-based architecture for improved adaptive HTTP video delivery. In *2016 IEEE Conference on Standards for Communications and Networking (CSCN)*. IEEE, 1–6.

[137] Chengchao Liang, F Richard Yu, and Xi Zhang. 2015. Information-centric network function virtualization over 5G mobile wireless networks. *IEEE network* 29, 3 (2015), 68–74.

[138] Chunyu Liu, Heli Zhang, Hong Ji, and Xi Li. 2021. MEC-assisted flexible transcoding strategy for adaptive bitrate video streaming in small cell networks. *China Communications* 18, 2 (2021), 200–214.

[139] Shaoshan Liu, Liangkai Liu, Jie Tang, Bo Yu, Yifan Wang, and Weisong Shi. 2019. Edge computing for autonomous driving: Opportunities and challenges. *Proc. IEEE* 107, 8 (2019), 1697–1716.

[140] Andra Lutu, Diego Perino, Marcelo Bagnulo, Enrique Frias-Martinez, and Javad Khangosstar. 2020. A Characterization of the COVID-19 Pandemic Impact on a Mobile Network Operator Traffic. In *Proceedings of the ACM Internet Measurement Conference*. 19–33.

[141] Huisheng Ma, Shufang Li, Erqing Zhang, Zhengnan Lv, Jing Hu, and Xinlei Wei. 2020. Cooperative Autonomous Driving Oriented MEC-Aided 5G-V2X: Prototype System Design, Field Tests and AI-Based Optimization Tools. *IEEE Access* 8 (2020), 54288–54302.

[142] Rohan Mahy, Philip Matthews, and Jonathan Rosenberg. 2010. *Traversal using relays around nat (turn): Relay extensions to session traversal utilities for nat (stun)*. Technical Report. RFC 5766 (Proposed Standard), Internet Engineering Task Force.

[143] Angel Martin, Jon Egaña, Julián Flórez, Jon Montalban, Igor G Olaizola, Marco Quartulli, Roberto Viola, and Mikel Zorrilla. 2018. Network resource allocation system for QoE-aware delivery of media services in 5G networks. *IEEE Transactions on Broadcasting* 64, 2 (2018), 561–574.

[144] Angel Martin, Roberto Viola, Mikel Zorrilla, Julián Flórez, Pablo Angueira, and Jon Montalbán. 2019. MEC for Fair, Reliable and Efficient Media Streaming in Mobile Networks. *IEEE Transactions on Broadcasting* 66, 2 (2019), 264–278.

[145] Jorge Martín-Pérez, Luca Cominardi, Carlos J Bernardos, Antonio de la Oliva, and Arturo Azcorra. 2019. Modeling mobile edge computing deployments for low latency multimedia services. *IEEE Transactions on Broadcasting* 65, 2 (2019), 464–474.

[146] Abbas Mehrabi, Matti Siekkinen, and Antti Ylä-Jääski. 2019. Energy-aware QoE and backhaul traffic optimization in green edge adaptive mobile video streaming. *IEEE Transactions on Green Communications and Networking* 3, 3 (2019), 828–839.

[147] Jie Mei, Xianbin Wang, and Kan Zheng. 2019. Intelligent network slicing for V2X services toward 5G. *IEEE Network* 33, 6 (2019), 196–204.

[148] Rashid Mijumbi, Sidhant Hasija, Steven Davy, Alan Davy, Brendan Jennings, and Raouf Boutaba. 2016. A connectionist approach to dynamic resource management for virtualised network functions. In *2016 12th International Conference on Network and Service Management (CNSM)*. IEEE, 1–9.

[149] Jean-Baptiste Monteil, Jernej Hribar, Pieter Barnard, Yong Li, and Luiz A DaSilva. 2020. Resource Reservation within Sliced 5G Networks: A Cost-Reduction Strategy for Service Providers. In *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 1–6.

[150] FJ Moreno-Muro, M Garrich, C San-Nicolás-Martínez, M Hernández-Bastida, P Pavón-Mariño, A Bravalheri, AS Muqaddas, N Uniyal, R Nejabati, D Simeonidou, et al. 2019. Joint VNF and multi-layer resource allocation with an open-source optimization-as-a-service integration. In *45th European Conference on Optical Communication (ECOC 2019)*. IET, 1–4.

[151] Jessica Moysen and Lorenza Giupponi. 2018. From 4G to 5G: Self-organized network management meets machine learning. *Computer Communications* 129 (2018), 248–268.

[152] Saman Naderiparizi, Mehrdad Hessar, Vamsi Talla, Shyamnath Gollakota, and Joshua R Smith. 2018. Towards battery-free {HD} video streaming. In *15th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 18)*. 233–247.

[153] Akihiro Nakao, Ping Du, Yoshiaki Kiriha, Fabrizio Granelli, Anteneh Atumo Gebremariam, Tarik Taleb, and Miloud Bagaa. 2017. End-to-end network slicing for 5G mobile networks. *Journal of Information Processing* 25 (2017), 153–163.

[154] Sabuzima Nayak and Ripon Patgiri. 2022. 6G Communication: A Vision on the Potential Applications. In *Edge Analytics*. Springer, 203–218.

[155] Anselme Ndikumana, Saeed Ullah, Tuan LeAnh, Nguyen H Tran, and Choong Seon Hong. 2017. Collaborative cache allocation and computation offloading in mobile edge computing. In *2017 19th Asia-Pacific Network Operations and Management Symposium (APNOMS)*. IEEE, 366–369.

[156] Van-Giang Nguyen, Anna Brunstrom, Karl-Johan Grinnemo, and Javid Taheri. 2017. SDN/NFV-based mobile packet core network architectures: A survey. *IEEE Communications Surveys & Tutorials* 19, 3 (2017), 1567–1602.

[157] James Nightingale, Qi Wang, Jose M Alcaraz Calero, Enrique Chirivella-Perez, Marian Ulbricht, Jesus A Alonso-Lopez, Ricardo Preto, Tiago Batista, Tiago Teixeira, Maria Joao Barros, et al. 2016. QoE-Driven, Energy-Aware Video Adaptation in 5G Networks: The SELFNET Self-Optimisation Use Case. *IJDSN* 12, 1 (2016), 7829305–1.

[158] Navid Nikaein, Chia-Yu Chang, and Konstantinos Alexandris. 2018. Mosaic5G: agile and flexible service platforms for 5G research. *ACM SIGCOMM Computer Communication Review* 48, 3 (Sept. 2018), 29–34. https://doi.org/10.1145/3276799.3276803

[159] Navid Nikaein, Mahesh K. Marina, Saravana Manickam, Alex Dawson, Raymond Knopp, and Christian Bonnet. 2014. OpenAirInterface: A Flexible Platform for 5G Research. *ACM SIGCOMM Computer Communication Review* 44, 5 (Oct. 2014), 33–38. https://doi.org/10.1145/2677046.2677053

[160] Nokia. [n.d.]. *EdenNet*. https://www.nokia.com/networks/portfolio/self-organizing-networks

[161] Andres F Ocampo, Thomas Dreibholz, Mah-Rukh Fida, Ahmed Elmokashfi, and Haakon Bryhni. 2020. Integrating Cloud-RAN with Packet Core as VNF Using Open Source MANO and OpenAirInterface. In *Proceedings of the 45th IEEE Conference on Local Computer Networks (LCN), Sydney, New South Wales/Australia (November 2020)*.

[162] Lyndon Ong, John Yoakum, et al. 2002. *An introduction to the stream control transmission protocol (SCTP)*. Technical Report. RFC 3286 (Informational), May.

[163] Jose Ordonez-Lucena, Pablo Ameigeiras, Diego Lopez, Juan J Ramos-Munoz, Javier Lorca, and Jesus Folgueira. 2017. Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges. *IEEE Communications Magazine* 55, 5 (2017), 80–87.

[164] John S Otto, Mario A Sánchez, John P Rula, Ted Stein, and Fabián E Bustamante. 2012. namehelp: Intelligent client-side DNS resolution. In *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication*. 287–288.

[165] Ye Ouyang, Zhongyuan Li, Le Su, Wenyuan Lu, and Zhenyi Lin. 2018. Application behaviors driven self-organizing network (SON) for 4G LTE networks. *IEEE Transactions on Network Science and Engineering* 7, 1 (2018), 3–14.

[166] Roger Pantos and William May. 2017. HTTP Live Streaming. *rfc 8216, August* (2017).

[167] Alain Pellen. 2020. *Cost Comparison: On-Premises Vs Cloud Computing*. https://www.harmonicinc.com/insights/blog/on-prem-vs-cloud

[168] Manuel Peuster, Stefan Schneider, Mengxuan Zhao, George Xilouris, Panagiotis Trakadas, Felipe Vicens, Wouter Tavernier, Thomas Soenen, Ricard Vilalta, George Andreou, Dimosthenis Kyriazis, and Holger Karl. 2019. Introducing Automated Verification and Validation for Virtualized Network Functions and Services. *IEEE Communications Magazine* 57, 5 (May 2019), 96–102. https://doi.org/10.1109/MCOM.2019.1800873

[169] Lucian Popa, Ali Ghodsi, and Ion Stoica. 2010. HTTP as the Narrow Waist of the Future Internet. In *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*. 1–6.

[170] Jon Postel et al. 1980. User datagram protocol. *STD 6, RFC 768, August* (1980).

[171] Jon Postel et al. 1981. Transmission control protocol. *STD 7, RFC 793, September* (1981).

[172] Dan Rayburn. 2020. *CDN/Media Pricing See's Big Drop for Largest Customers: Pricing Down to $0.0006*. https://www.streamingmediablog.com/2020/05/q1-cdn-pricing.html

[173] Daniela Renga and Michela Meo. 2018. From self-sustainable green mobile networks to enhanced interaction with the smart grid. In *2018 30th International Teletraffic Congress (ITC 30)*, Vol. 1. IEEE, 129–134.

[174] Sepehr Rezvani, Saeedeh Parsaeefard, Nader Mokari, Mohammad R Javan, and Halim Yanikomeroglu. 2019. Cooperative multi-bitrate video caching and transcoding in multicarrier NOMA-assisted heterogeneous virtualized MEC networks. *IEEE Access* 7 (2019), 93511–93536.

[175] Jordi Ferrer Riera, Eduard Escalona, Josep Batalle, Eduard Grasa, and Joan A Garcia-Espin. 2014. Virtual network function scheduling: Concept and challenges. In *2014 international conference on smart communications in network technologies (SaCoNeT)*. IEEE, 1–5.

[176] Rodrigo Roman, Javier Lopez, and Masahiro Mambo. 2018. Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges. *Future Generation Computer Systems* 78 (2018), 680–698.

[177] Peter Rost, Christian Mannweiler, Diomidis S Michalopoulos, Cinzia Sartori, Vincenzo Sciancalepore, Nishanth Sastry, Oliver Holland, Shreya Tayade, Bin Han, Dario Bega, et al. 2017. Network slicing to enable scalability and flexibility in 5G mobile networks. *IEEE Communications*

magazine 55, 5 (2017), 72–79.

[178] Ahmad Rostami. 2019. Private 5G networks for vertical industries: Deployment and operation models. In *2019 IEEE 2nd 5G World Forum (5GWF)*. IEEE, 433–439.

[179] Dario Sabella, Vadim Sukhomlinov, Linh Trang, Sami Kekki, Pietro Paglierani, Ralf Rossbach, Xinhui Li, Yonggang Fang, Dan Druta, Fabio Giust, et al. 2019. Developing software for multi-access edge computing. *ETSI white paper* 20 (2019), 1–38.

[180] Ali Saffari, Mehrdad Hessar, Saman Naderiparizi, and Joshua R Smith. 2019. Battery-free wireless video streaming camera system. In *2019 IEEE International Conference on RFID (RFID)*. IEEE, 1–8.

[181] Pablo Salva-Garcia, Jose M. Alcaraz-Calero, Qi Wang, Miguel Arevalillo-Herraez, and Jorge Bernal Bernabe. 2020. Scalable Virtual Network Video-Optimizer for Adaptive Real-Time Video Transmission in 5G Networks. *IEEE Transactions on Network and Service Management* 17, 2 (June 2020), 1068–1081. https://doi.org/10.1109/TNSM.2020.2978975

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

[183] Frederico Schardong, Ingrid Nunes, and Alberto Schaeffer-Filho. 2020. NFV resource allocation: A systematic review and taxonomy of VNF forwarding graph embedding. *Computer Networks* (2020), 107726.

[184] Henning Schulzrinne, Stephen Casner, Ron Frederick, Van Jacobson, et al. 1996. RTP: A transport protocol for real-time applications. *rfc 1889, January* (1996).

[185] Henning Schulzrinne, Anup Rao, and Robert Lanphier. 1998. Real time streaming protocol (RTSP). *rfc 2326, April* (1998).

[186] Michael Seufert, Sebastian Egger, Martin Slanina, Thomas Zinner, Tobias Hoßfeld, and Phuoc Tran-Gia. 2014. A survey on quality of experience of HTTP adaptive streaming. *IEEE Communications Surveys & Tutorials* 17, 1 (2014), 469–492.

[187] Sonia Shahzadi, Muddesar Iqbal, Tasos Dagiuklas, and Zia Ul Qayyum. 2017. Multi-access edge computing: open issues, challenges and future perspectives. *Journal of Cloud Computing* 6, 1 (2017), 1–13.

[188] Maria Sharabayko, Maxim Sharabayko, Jean Dube, Joonwoong Kim, and Jeongseok Kim. 2021. The SRT Protocol. *draft-sharabayko-srt-00* (2021).

[189] Chetna Singhal and BN Chandana. 2021. Aerial-SON: UAV-based Self-Organizing Network for Video Streaming in Dense Urban Scenario. In *2021 International Conference on COMmunication Systems & NETworkS (COMSNETS)*. IEEE, 7–12.

[190] Lea Skorin-Kapov, Martín Varela, Tobias Hoßfeld, and Kuan-Ta Chen. 2018. A survey of emerging concepts and challenges for QoE management of multimedia services. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 2s (2018), 1–29.

[191] Iraj Sodagar. 2011. The mpeg-dash standard for multimedia streaming over the internet. *IEEE multimedia* 18, 4 (2011), 62–67.

[192] Thomas Soenen, Wouter Tavernier, Manuel Peuster, Felipe Vicens, George Xilouris, Stavros Kolometsos, Michail-Alexandros Kourtis, and Didier Colle. 2019. Empowering network service developers: enhanced nfv devops and programmable mano. *IEEE Communications Magazine* (May 2019). https://doi.org/10.1109/MCOM.2019.1800810

[193] Guan-Ming Su, Xiao Su, Yan Bai, Mea Wang, Athanasios V Vasilakos, and Haohong Wang. 2016. QoE in video streaming over wireless networks: perspectives and research challenges. *Wireless networks* 22, 5 (2016), 1571–1593.

[194] Yang Sun, Tingting Wei, Huixin Li, Yanhua Zhang, and Wenjun Wu. 2020. Energy-efficient multimedia task assignment and computing offloading for mobile edge computing networks. *IEEE Access* 8 (2020), 36702–36713.

[195] SVA. [n.d.]. *Open Caching*. https://www.streamingvideoalliance.org/working-group/open-caching/

[196] Yiming Tan, Ce Han, Ming Luo, Xiang Zhou, and Xing Zhang. 2018. Radio network-aware edge caching for video delivery in MEC-enabled cellular networks. In *2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*. IEEE, 179–184.

[197] David Alexander Tedjopurnomo, Zhifeng Bao, Baihua Zheng, Farhana Choudhury, and AK Qin. 2020. A survey on modern deep neural network for traffic prediction: Trends, methods and challenges. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[198] Michael Thornburgh. 2014. Adobe's RTMFP Profile for Flash Communication. *Internet Engineering Task Force (IETF)* (2014).

[199] Lechosław Tomaszewski, Sławomir Kukliński, and Robert Kołakowski. 2020. A new approach to 5G and MEC integration. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 15–24.

[200] Ruben Torres, Alessandro Finamore, Jin Ryong Kim, Marco Mellia, Maurizio M Munafo, and Sanjay Rao. 2011. Dissecting video server selection strategies in the youtube cdn. In *2011 31st International Conference on Distributed Computing Systems*. IEEE, 248–257.

[201] Tuyen X. Tran, Parul Pandey, Abolfazl Hajisami, and Dario Pompili. 2016. Collaborative Multi-bitrate Video Caching and Processing in Mobile-Edge Computing Networks. arXiv:1612.01436 [cs.NI]

[202] Tuyen X Tran and Dario Pompili. 2018. Adaptive bitrate video caching and processing in mobile-edge computing networks. *IEEE Transactions on Mobile Computing* 18, 9 (2018), 1965–1978.

[203] Refik Fatih Ustok, Ugur Acar, Selcuk Keskin, David Breitgand, Avi Weit, Petros Drakoulis, Alexandros Doumanoglou, Nikolaos Zioulis, Dimitrios Zarpalas, Petros Daras, Francesco Iadanza, Francesca Moscatelli, and Giacomo Bernini. 2020. Service Development Kit for Media-Type Virtualized Network Services in 5G Networks. *IEEE Communications Magazine* 58, 7 (2020), 51–57. https://doi.org/10.1109/MCOM.001.1900613

[204] Spyridon Vassilaras, Lazaros Gkatzikis, Nikolaos Liakopoulos, Ioannis N Stiakogiannakis, Meiyu Qi, Lei Shi, Liu Liu, Merouane Debbah, and Georgios S Paschos. 2017. The algorithmic aspects of network slicing. *IEEE Communications Magazine* 55, 8 (2017), 112–119.

[205] Gorka Velez, Ángel Martín, Giancarlo Pastor, and Edward Mutafungwa. 2020. 5G Beyond 3GPP Release 15 for Connected Automated Mobility in Cross-Border Contexts. *Sensors* 20, 22 (2020), 6622.

[206] Roberto Viola, Angel Martin, Javier Morgade, Stefano Masneri, Mikel Zorrilla, Pablo Angueira, and Jon Montalbán. 2020. Predictive CDN Selection for Video Delivery Based on LSTM Network Performance Forecasts and Cost-Effective Trade-Offs. *IEEE Transactions on Broadcasting* (2020).

[207] Roberto Viola, Angel Martin, Mikel Zorrilla, and Jon Montalbán. 2018. MEC proxy for efficient cache and reliable multi-CDN video distribution. In *2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 1–7.

[208] W3C. 2020. *QUIC API for Peer-to-peer Connections*. https://w3c.github.io/webrtc-quic/

[209] Minghao Wang, Tianqing Zhu, Tao Zhang, Jun Zhang, Shui Yu, and Wanlei Zhou. 2020. Security and privacy in 6G networks: New areas and new challenges. *Digital Communications and Networks* 6, 3 (2020), 281–291.

[210] Qi Wang, Jose Alcaraz-Calero, Ruben Ricart-Sanchez, Maria Barros Weiss, Anastasius Gavras, Navid Nikaein, Xenofon Vasilakos, Bernini Giacomo, Giardina Pietro, Mark Roddy, Michael Healy, Paul Walsh, Thuy Truong, Zdravko Bozakov, Konstantinos Koutsopoulos, Pedro Neves, Cristian Patachia-Sultanoiu, Marius Iordache, Elena Oproiu, Imen Grida Ben Yahia, Ciriaco Angelo, Cosimo Zotti, Giuseppe Celozzi, Donal Morris, Ricardo Figueiredo, Dean Lorenz, Salvatore Spadaro, George Agapiou, Ana Aleixo, and Cipriano Lomba. 2019. Enable Advanced QoS-Aware Network Slicing in 5G Networks for Slice-Based Media Use Cases. *IEEE Transactions on Broadcasting* 65, 2 (June 2019), 444–453. https://doi.org/10.1109/TBC.2019.2901402

[211] Xinwei Wang, Jiandong Li, Lingxia Wang, Chungang Yang, and Zhu Han. 2019. Intelligent user-centric network selection: A model-driven reinforcement learning framework. *IEEE Access* 7 (2019), 21645–21661.

[212] Yanting Wang, Yan Zhang, Min Sheng, and Kun Guo. 2019. On the Interaction of Video Caching and Retrieving in Multi-Server Mobile-Edge Computing Systems. *IEEE Wireless Communications Letters* 8, 5 (2019), 1444–1447. https://doi.org/10.1109/LWC.2019.2921759

[213] Yiming Wei, Mugen Peng, and Yaqiong Liu. 2020. Intent-based networks for 6G: Insights and challenges. *Digital Communications and Networks* 6, 3 (2020), 270–280.

[214] Matthias Wichtlhuber, Robert Reinecke, and David Hausheer. 2015. An SDN-based CDN/ISP collaboration architecture for managing high-volume flows. *IEEE Transactions on Network and Service Management* 12, 1 (2015), 48–60.

[215] Dan Wing, Philip Matthews, Rohan Mahy, and Jonathan Rosenberg. 2008. Session traversal utilities for NAT (STUN). *RFC5389, October* (2008).

[216] Yinhao Xiao, Yizhen Jia, Chunchi Liu, Xiuzhen Cheng, Jiguo Yu, and Weifeng Lv. 2019. Edge computing security: State of the art and challenges. *Proc. IEEE* 107, 8 (2019), 1608–1631.

[217] Yanghao Xie, Zhixiang Liu, Sheng Wang, and Yuxiu Wang. 2016. Service function chaining resource allocation: A survey. *arXiv preprint arXiv:1608.00095* (2016).

[218] Chenren Xu, Lei Yang, and Pengyu Zhang. 2018. Practical backscatter communication systems for battery-free Internet of Things: A tutorial and survey of recent research. *IEEE Signal Processing Magazine* 35, 5 (2018), 16–27.

[219] Mao Yang, Yong Li, Depeng Jin, Li Su, Shaowu Ma, and Lieguang Zeng. 2013. OpenRAN: A Software-Defined Ran Architecture via Virtualization. *SIGCOMM Comput. Commun. Rev.* 43, 4 (Aug. 2013), 549–550. https://doi.org/10.1145/2534169.2491732

[220] Jingjing Yao, Tao Han, and Nirwan Ansari. 2019. On mobile edge caching. *IEEE Communications Surveys & Tutorials* 21, 3 (2019), 2525–2553.

[221] Ahmet Yazar, Seda Doğan Tusha, and Huseyin Arslan. 2020. 6G VISION: AN ULTRA-FLEXIBLE PERSPECTIVE. *ITU Journal on Future and Evolving Technologies* 1, 1 (2020).

[222] Jongwon Yoon and Suman Banerjee. 2020. Hardware-Assisted, Low-Cost Video Transcoding Solution in Wireless Networks. *IEEE Transactions on Mobile Computing* 19, 3 (2020), 581–597. https://doi.org/10.1109/TMC.2019.2898834

[223] Chuanji Zhang, Harshvardhan P Joshi, George F Riley, and Steven A Wright. 2019. Towards a virtual network function research agenda: A systematic literature review of vnf design considerations. *Journal of Network and Computer Applications* 146 (2019), 102417.

[224] Shunliang Zhang. 2019. An overview of network slicing for 5G. *IEEE Wireless Communications* 26, 3 (2019), 111–117.

[225] Xiaoxi Zhang, Chuan Wu, Zongpeng Li, and Francis CM Lau. 2017. Proactive VNF provisioning with multi-timescale cloud resources: Fusing online learning and online optimization. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 1–9.

[226] Tiesong Zhao, Qian Liu, and Chang Wen Chen. 2016. QoE in video transmission: A user experience-driven strategy. *IEEE Communications Surveys & Tutorials* 19, 1 (2016), 285–302.

[227] Yiqing Zhou, Ling Liu, Lu Wang, Ning Hui, Xinyu Cui, Jie Wu, Yan Peng, Yanli Qi, and Chengwen Xing. 2020. Service aware 6G: an intelligent and open network based on convergence of communication, computing and caching. *Digital Communications and Networks* (2020).