IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 30, 2020, accepted October 13, 2020, date of publication October 21, 2020, date of current version November 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3032726

A Survey of End-to-End Solutions for Reliable Low-Latency Communications in 5G Networks

DELIA RICO AND PEDRO MERINO
ITIS Software, Universidad de Málaga, Andalucía Tech, 29071 Málaga, Spain

Corresponding author: Delia Rico (deliarico@uma.es)

This work was supported in part by the 5Genesis Project (European Union Horizon 2020 research and innovation) under grant agreement No. 815178 and in part by the Ministry of Education of Spain through grant FPU17/04292 and grant agreement RTI2018-099777-B-I00 (RFOG Project).

ABSTRACT Very low latency and high reliability are two of the main requirements for new applications exploiting 5G networks. This is the case for the remote operation of robots or vehicles, the autonomous interaction of equipment in a factory, autonomous driving and tactile internet applications. Although the TCP/IP stack has been sufficient as the end-to-end solution for most of the history of the Internet, a number of surveys have appeared recently presenting many different methods for managing the end-to-end communication to meet the requirements of various technologies such as that of 5G networks. In this paper, we present a novel classification of the literature focused on new end-to-end solutions and the creation of services towards the support of low latency (1 ms) and high reliability (10^{-9} error rate) in current and future 5G networks. We specifically highlight how the proposals can be classified according to enabling technologies and the specific method used to achieve success in terms of the latency and reliability. The literature related to end-to-end solutions for reliable low-latency communications are organized according to three main topics: (i) end-to-end protocols that improve communication in terms of latency and reliability, (ii) functionality or technologies implemented on the network to support the current demands, and (iii) application programming interfaces that enhance the correct utilization of those protocols and additional technologies.

INDEX TERMS Context awareness, edge computing, high reliability, low latency, multi-connectivity, TCP/IP.

I. INTRODUCTION

Since their standardization in the 80s, IP, UDP and TCP have positioned themselves as the most important Internet and transport layer protocols [1]–[3]. Although the TCP/IP stack has been sufficient for most of the history of the Internet, recent tendencies in communications are creating a greater challenge with more stringent requirements. The ossification of the Internet stack is a well-known issue [4] that has been aggravated by the arrival of 5G networks. However, even though TCP/IP variants are expected to be the main end-to-end transport protocols for applications in 5G networks, these protocols will integrate and collaborate with other enabling technologies to comply with 5G critical requirements.

5G networks and their three categories of services, namely, eMBB (enhanced mobile broadband), MTC (machine-type communication) and URLLC (ultra-reliable low-latency Communication) [15], [16], present critical requirements in terms of reliability, latency, throughput and capacity, among others. Two of these requirements, reliability and latency, are especially important in communications for mission critical applications, where the three most representative use cases are remote surgery, factory automation and autonomous connected cars [6]. Remote surgery can occur during complex life-saving procedures in health emergencies [5], where networks should be able to support the communication needs since any noticeable error could lead to catastrophic outcomes. Factory automation is a high-reliability, low-latency and low-jitter use case [17] traditionally based on wired networks that is being directed into the wireless and cellular world for enhanced deployment flexibility, reduced cost of maintenance and higher long-term reliability through initiatives such as time-sensitive networking [18]–[20]. Finally, autonomous connected car communication [21] requires

The associate editor coordinating the review of this manuscript and approving it for publication was Cristian Zambelli.

TABLE 1. Use case definitions and required values for KPIs.

Use Case	Description	Max Latency (ms)	Max Error Rate	References
Remote Surgery	Real-time complex life-saving procedures which require reliable and low latency for video streaming and feedback to perform precise movements in the robotic arms.	1	10^{-9}	[5]
Factory Automation and Industrial Control	Precise operation in factory processes and power system automation services that require high reliability, low latency and jitter is not tolerated.	0.1-1	10^{-9}	[6] [5] [7] [8]
Intelligent Transportation Systems	Communication between cars that require high reliability and low latency to avoid misunderstanding control messages and to keep the car connected at all times.	5-10	10^{-5}	[6] [5] [7]
Internet of Drones (remote control)	Remote control of drones with a wide area of operation that demands cellular networks with enhanced reliability and latency.	5-50	10^{-3}	[9] [10] [11]
Tactile Internet	Real-time interactive systems where machines and humans interact in a low latency and high reliability scenario.	0.5-5	10^{-9}	[12] [7]
Cyber-physical Systems	Interacting networks of physical and computational components requiring ultra reliability and low latency.	0.5-2	10^{-5}	[13]
Networked Action Games	Gaming requires increasing confidence in transporting recent messages and processing them quickly.	10	10^{-4}	[14]
Virtual Reality	Audio-visual feedback to support human interaction with the environment, people or remote control.	5-10	10^{-4}	[14]
Health Monitoring	Real-time observation of the state of a person for monitoring and consultation.	100	10^{-4}	[14]
Smart Grids	Improving the efficiency of energy distribution with a prompt reaction reconfiguring in response to events.	8	10^{-5}	[10]

99.999% reliability to avoid misinterpreted control messages, low latency and seamless robust handover to keep the car connected at all times, and even information about other vehicles combined with edge computing solutions to increase the general performance.

Nevertheless, remote surgery, factory automation and autonomous connected cars are not the only existing use cases. The Internet of Drones [22]–[24], IEEE tactile internet (TI) [12], 3GPP cyber-physical systems (CPS) [13], networked action games [25], virtual reality/augmented reality, eHealth periodic monitoring, smart grids, etc., are just some examples of the wide variety of applications that are currently being developed with demanding requirements mainly in terms of low latency and high reliability. Table 1 offers a better understanding of the use cases and the KPI[1] target values. The first and second columns show some of the most important use cases for critical applications and their definitions, whereas columns 3 and 4 display the two KPIs under study, the maximum supported latency (in milliseconds) and the minimum reliability required (in terms of the maximum error rate tolerated). Finally, the last column presents references to scientific papers that justify these values.

The evolution towards new techniques for latency and reliability has been partially studied in other surveys, which are described in Section II. Most of them focus on low-layer protocols and technologies, while transport protocols and close technologies are not sufficiently analysed from a common latency-reliability perspective. In this paper, we present a comprehensive and updated survey on novel technologies and solutions to fill in the gaps of the previous papers and to identify research opportunities in the context of end-to-end solutions. The survey focuses mainly on technologies that are close to applications, instead of the lower layers, and considers the need to enhance communications as a whole and not just the protocols or concrete technologies. Furthermore, we consider contributions with the aim of enhancing reliability and latency jointly, instead of focusing only on the contributions of one KPI.

We distinguish 3 lines of research to improve communications: the enhancement of end-to-end protocols, the support of the network and the use of information from outside the scope (e.g., network state). The survey methodology relies on the identification of enabling technologies that fit into these categories (e.g., single-path, multipath or multicast protocols, edge computing, software-defined networking, network function virtualization and information-centric networking) and the study of APIs (application programming interfaces). We then evaluate the common methods and techniques used to enhance the performance of these enabling technologies and APIs. Finally, we select a number of 5G-related KPIs and other relevant parameters to determine the reliability and latency (such as low latency, high reliability, high throughput, partial reliability or heterogeneous network support). The parameters come from the 5G-PPP European initiative; however, they are aligned with other world-wide activities such as 5G Americas, 5G Forum, 5G Brasil and 5GMF [26].

In this context, we present more than 150 papers and organize recent contributions in a number of tables according to several classification criteria, like relevant parameters and the methods to reduce latency and/or to increase reliability. The output of this analysis is a new characterization of the current state-of-the-art and the identification of research topics where more effort is required to make the TCP/IP stack and other end-to-end technologies for managing services suitable for achieving reliable low-latency communications.

[1] Key performance indicators (KPIs) are measurements of specific network properties that help in monitoring, optimizing and characterizing services.

Compared with previous surveys, we provide a different view of the state-of-the-art protocols, technologies and APIs used to support enhanced reliability and latency services. In particular, we analyse each contribution simultaneously using a number of relevant parameters, some of which were not considered in previous works; and we evaluate common methods and techniques used to enhance the performance of the enabling technologies and APIs. Finally, we present a comprehensive evaluation to identify the current research efforts and future lines of study. It is worth noting that many contributions that were initially designed for 4G networks are included in the survey because they are still valid for 5G networks.

This paper is organized as follows. Section II introduces a comparative analysis of previous surveys. Section III explains the classification criteria used to select the contributions and the parameters evaluated. Then, Section IV analyses the contributions from the scientific literature and presents our characterization of the state-of-the-art, while Section V evaluates these contributions, identifying possible future lines of research. Finally, we conclude our paper in Section VI.

II. RELATED SURVEYS

The presented scenario of new use cases for 5G networks has led to a need for network evolution in both the lower and higher layers of the protocol stack. In this section, we collect previous research efforts made to gather contributions that improve the reliability and latency or study novel technologies aimed at achieving this network evolution.

The first point of study is the 5G-related surveys, which work on enhancements in latency or reliability over these novel cellular networks. We detect a large focus on lower-layer solutions and a lack of analysis of the reliability and latency conjointly when the focus is set on transport protocols and solutions. Some surveys of the evolution of protocols and techniques for 5G critical communications are those of Sutton *et al.* [27], Pocovi *et al.* [8], Zhang *et al.* [28] and Morgado *et al.* [29], which present a variety of enabling technologies to enhance communication in terms of latency or reliability but mostly focus on the lower layers. Furthermore, Mitra *et al.* [30], Agiwal *et al.* [31], Gupta *et al.* [32] and Olwal *et al.* [33] provide surveys aimed at the study of emergent technologies, paradigms and applications for 5G networks. However, even though they present some higher-layer contributions, the main focus is again mostly on lower-layer solutions, such as self-organizing networks (SON). Additionally, the surveys do not set their research efforts towards enhancing both reliability and latency but instead only the general performance. Similar studies are the ones of Jaber *et al.* [34], focused on 5G backhauling, and Chettri *et al.* [35], targeting 5G IoT systems. Finally, Nasrallah *et al.* [36] and Parvez *et al.* [37] introduced different methods and contributions towards enhanced performance but focused only on latency.

It is also interesting to highlight the research efforts made to survey concrete technologies, considered separately and not analysed from a common perspective. Habib *et al.* [38] and Li *et al.* [39] present studies of multipaths in different layers; Mao *et al.* [40] and Wan *et al.* [41] present surveys on mobile edge networks; Al-Anbagi *et al.* [42] carry out a survey on cross-layer approaches for delay and reliability-aware applications; Papastergiou *et al.* [4] present a fairly comprehensive overview of context-aware solutions; Taleb *et al.* [43] introduce a survey on mobile edge computing (MEC) and focuses on other fundamental key enabling technologies in 5G contributions such as software-defined networking (SDN) and network function virtualization (NFV); and Yürür *et al.* [44] present a survey on context awareness for mobile sensing.

Finally, other relevant surveys on the enhancement of reliability and latency, which study some of the enabling technologies presented in this paper, are the ones of Elbamby *et al.* [45], Briscoe *et al.* [46] and Antonakoglou *et al.* [47]. The approach of Elbamby *et al.* [45] is very theoretical, not presenting or analysing a wide variety of contributions. Briscoe *et al.* [46] study Internet enhancements but focus only on latency. Moreover, Antonakoglou *et al.* [47] focus their efforts on finding contributions to data compression and reduction, robust stability control, and multi-modal data streaming over the Internet.

Table 2 summarizes the technologies studied in each survey, the approaches taken to analyse them and whether the main focus of the survey was on the lower layers. Checkmarks indicate surveys with high treatment of the topic, while bullets highlight those surveys that mention the topic but not with a deep analysis or focus. We selected the columns according to the topics found in the surveys: the large focus on the lower layers; technologies such as transport protocols, multi-connectivity, edge computing, etc.; and approaches such as studying the contributions of low latency, high reliability, partial reliability, cellular networks, surveys that study 5G use cases, and those that analyse the contributions in detail or take into consideration the heterogeneous network paradigm.

In our evaluation, we detected several tendencies:

- Most of the research on the state-of-the-art for 5G network evolution is focused on lower-layer solutions.
- There is not enough analysis of both the reliability and latency conjointly when the focus is on the higher layers.
- For the higher layers, there is also a lack of research on transport protocols and network support solutions conjointly as a plausible solution to support the novel requirements.
- Heterogeneous network (HetNet) support, or the ability to work properly under these conditions, is a key point of study in most solutions due to the fact that different technologies with diverse characteristics coexist in the current networks.
- Partial reliability is often forgotten as a possible enabler for certain use cases.
- Technologies such as EDGE, SDN, and NFV and solutions such as multi-connectivity and context awareness

TABLE 2. Comparison with previous surveys.

	Focus	Technologies and methods under study								Approach & Analysis						
	Lower Layers	SON	Transport Protocols	Multi-connectivity	Context Awareness	EDGE	SDN	NFV	ICN/CDN	Low Latency	High Reliability	Partial Reliability	Cellular Networks	5G Use Cases	Numerous contributions	HetNets
Sutton et al. [27]	✓				✓	●	●			✓	✓		✓	✓	✓	
Pocovi et al [8]	✓									✓	✓		✓		●	
Zhang et al. [28]	✓						●	●		✓	✓		✓	✓	✓	
Morgado et al. [29]	✓					✓	✓	✓		✓	✓		✓	✓	✓	●
Mitra et al. [30]	●		✓			●				✓	●		✓	✓	✓	●
Agiwal et al. [31]	✓	✓			✓	●	✓		●	✓	●		✓	✓	✓	✓
Gupta et al. [32]	✓	✓			✓	●	✓	●		✓	●		✓	✓	✓	✓
Olwal et al. [33]	✓	●			✓	●	✓			✓	●		✓	●	✓	✓
Jaber et al. [34]	✓	●			●		●	●		✓	●		✓		✓	✓
Chettri et al. [35]	●		●			●				●			✓	●	✓	✓
Nasrallah et al [36]	●	●	●		✓	✓	✓	✓		✓			✓	✓	✓	✓
Parvez et al. [37]	●	●			✓	✓	✓	✓	●	✓			✓	✓	✓	●
Habib et al. [38]			✓	✓							✓	✓			✓	✓
Li et al. [39]	●		✓	✓	✓	●			✓	●	✓	●			✓	✓
Mao et al. [40]					✓	✓	●	●	●	✓	✓		✓		✓	
Wan et al. [41]					✓	✓	●	●	●	✓			✓		✓	✓
Al-Anbagi et al. [42]	●		●	●	✓					✓	✓			✓	✓	●
Papastergiou et al. [4]			✓	✓	✓	●				●	●				✓	
Taleb et al. [43]					✓	✓	✓	✓	●	●	✓		✓		✓	●
Yürür et al. [44]					✓										✓	●
Elbamby et al. [45]	●			●		✓				✓	✓		✓	✓	●	
Briscoe et al. [46]	●		✓	✓	●	✓	●		✓	✓	●	●			✓	
Antonakoglou et al. [47]	●		✓	●	●	●	●			✓	✓		✓	✓	✓	●
This Survey			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

are of high importance in enhancing the reliability and latency.

- Content delivery paradigms such as information-centric networking (ICN), aimed at redesigning the current Internet infrastructure, leaving behind the point-to point paradigm and embracing techniques such as catching, data replication and content distribution [48], are promising solutions that can enhance the latency of the content distribution.

From the analysis of these previous surveys, we determine a different approach to organize the literature related to end-to-end solutions for the enhancement of the reliability and latency. Section III presents this classification criteria proposed for the evaluation in more depth. Furthermore, the last row of Table 2 presents a direct comparison of the presented state-of-the-art with this survey. There, we can see how this paper analyses all end-to-end solutions and methods focused on higher layers that have been identified as relevant from all presented approaches.

III. CLASSIFICATION CRITERIA FOR PREVIOUS WORK

In this section, we present the classification criteria selected for this survey: enabling technologies, APIs, common methods and techniques used to enhance the performance of these approaches, and the parameters extracted from different relevant surveys that have helped in the characterization of the literature.

A. ENABLING TECHNOLOGIES FOR LOW LATENCY AND HIGH RELIABILITY

TCP, UDP and their variants are expected to be the main end-to-end transport protocols for applications in 5G networks. However, these protocols will be integrated and collaborate with other enabling technologies to reduce the latency and to increase the reliability. For instance, the use of MEC will move one of the final end points from the cloud to the EDGE; the use of NFV could even change the location of the end points if some reconfiguration is required; the use of information from the network to be used in the logic of the transport protocol implies new APIs for context awareness; and the use of cache mechanisms or some other ICN acceleration technique in ICN indicates some kind of offloading of work from the TCP/IP path to a different location. This strong interrelation motivates us to present all these enabling technologies or transport solutions in this section and the APIs in the following section.

Works such as that of Elbamby *et al.* [45] or that of Parvez *et al.* [37] have helped in the selection of the categories to be analysed. Elbamby *et al.* [45] study the importance of reliability and latency in virtual reality and present multi-connectivity, edge computing and multicasting as enabling solutions; whereas Parvez *et al.*, [37] study some of the increasingly important novel technologies for 5G, such as software-defined networking, network function virtualization and information-centric networking.

Based on these surveys mentioned and the stated above, the categories selected for this work as enabling technologies are shown jointly in Figure 1 and described in the following subsections.

1) END-TO-END PROTOCOLS

To improve communications, the first approach needed is to enhance the communication protocols themselves. Novel communication protocols have been classified into three main categories.

- **Single-path protocols:** A proper communication protocol is necessary in each case to exploit the full capabilities of a network [49]. It is equally important to focus on physical layer improvements as well as on protocols, since an inefficient protocol will limit the possibility of taking advantage of network capabilities. For this reason, this survey analyses enhancements in existing protocols (such as UDP, TCP and their variants) as well as novel protocols.
- **Multipath protocols:** Another approach is to take communication protocols further and improve their capabilities over multiple flows instead of single flows. As Qadir *et al.* [50] noted, the Internet's future is inherently multipath. Multihoming capabilities, path/interface/network diversity, data centre enhancements and wireless communications are leading networking into the use of multi-access connectivity. The benefits of multiple connectivity include better reliability, network offloading, improved availability, etc.
- **Multicast protocols:** Poularakis *et al.* [51] and Araniti *et al.* [52] reflect on the growth of mobile multicast applications and present multicasting as a key opportunity in future 5G networks. Sending the same copy of information to multiple receivers at a given moment of time can provide lower latencies, higher scalability and network offloading. In 5G applications such as intelligent transport systems, assisted driving, etc., this technology will play a key role. European Union's Horizon 2020 research and innovation programme projects such as 5G-Xcast [53] focus on enhancing this technology in terms of improving several KPIs such as the data rate, latency, reliability and power consumption.

2) NETWORK SUPPORT

The network technologies that support protocol or application operation selected to reduce latency and to increase reliability are as follows.

- **Edge computing (EDGE):** Edge, MEC and fog computing[2] are key enabling technologies for novel 5G requirements [40]. Edge computing consists of moving the cloud and some network functions closer to the user to provide services locally and consequently improve performance, such as a reduced latency. Intelligent transportation systems, virtual reality and network offloading are some examples of areas that can benefit from this technology.
- **Software-defined networking (SDN):** SDN is a novel approach that consists of creating a decoupled architecture that splits the control and data planes. SDN allows intelligent routing, flexibility, programmability and facilitates virtualization [56]. The increasing interest in SDN solutions by telecommunication service providers (e.g., Ericsson Cloud SDN [57] and Nokia Software-Defined Access Networks [58]) and the fact that some of the proven benefits of SDN are load balancing, signalling reduction and improvements in general parameters such as latency and reliability [59] make SDN a relevant technology for this survey.
- **Network function virtualization (NFV):** NFV [60] is a novel solution standardized by the ETSI in 2014 [61] aiming to virtualize network functionalities. NFV decouples software functionalities from physical equipment to offer better flexibility, scalability, latency, reliability, capacity, etc. NFV is a promising solution for 5G communications by itself and can be combined with other technologies, such as in the Huawei Cloud solution [62].
- **Information-centric networking (ICN):** ICN is an approach to redesign the current Internet infrastructure to leave behind the point-to-point paradigm and embrace data replication, content distribution, naming schemes and catching [48], [63]. Although it is not just a technology but a combination of techniques that can be used to evolve the current Internet architecture, ICN has a similar role to that of EDGE, SDN or NFV, providing network assistance and evolution to enhance KPIs such as latency in the case of content distribution applications. Thus, we found it appropriate to present this paradigm in this section.

Some of the papers that have helped in the study of the different enabling technologies are the following: Habib *et al.* [38] and Li *et al.* [39] present studies of the multipath in different layers; Mouradian *et al.* [64], Mao *et al.* [40] and Wan *et al.* [41] present surveys on mobile edge computing and fog computing; and Antonakoglou *et al.* [47] study the necessary infrastructure for tactile internet.

B. APPLICATION PROGRAMMING INTERFACE (API)

The concept of the API is present in almost all the enabling technologies. Application programming interfaces (APIs) are intermediaries that allow application layers to manage transport information and even the information of the lower layers in order to work flexibly according to the needs. Taking advantage of information outside of a protocol's or layer's scope of work and managing these functionalities could result

[2] MEC refers to the computation on the edge of the network standarized by the ETSI [54]. Edge computing is more flexible, since it does not necessarily use the technologies recommended by the standard. Fog computing refers to the computation carried out at computing nodes placed at any point of the architecture between the end devices and the cloud (fog layer) [55].

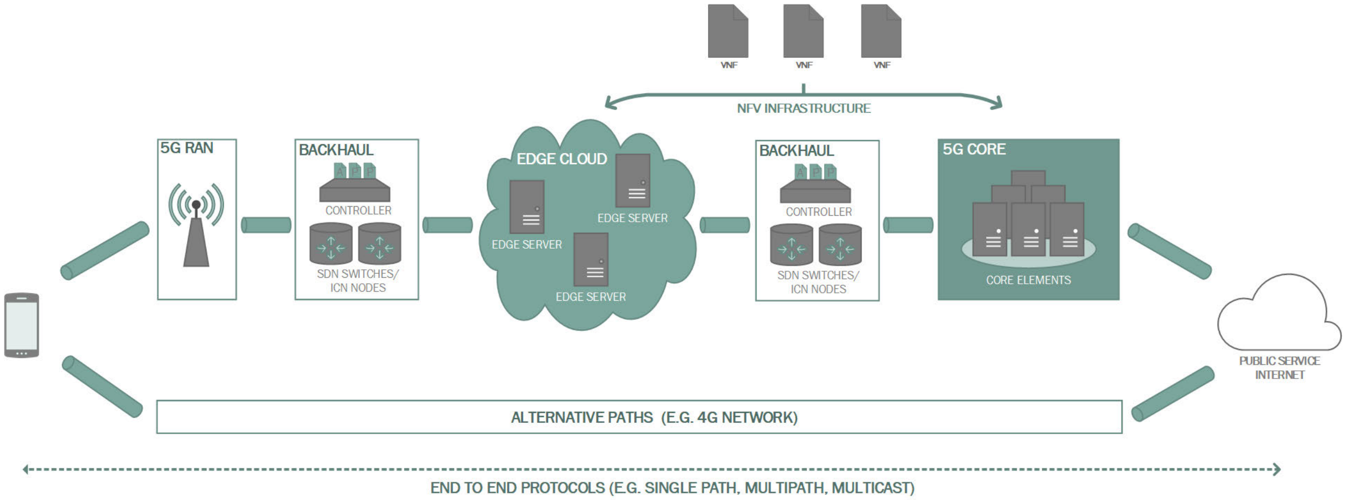

FIGURE 1. **Enabling technologies in 5G networks.**

in a better service, in terms of the general performance or concrete parameters such as latency and reliability, that would benefit both the services and the network, resulting in a better service with reduced overload. The traditional socket API is too low-level, simple and inflexible [65]–[67] and has been questioned for a long time. Due to the number of works on this subject, in this survey, we create a separate category for discussing papers on APIs to improve the reliability and latency.

C. COMMON METHODS AND TECHNIQUES

These enabling technologies and APIs have been presented in high-level summarizations; however, works based on these technologies can also be grouped regarding the common methods and techniques used to enhance their performance. In this section, we present different common solutions used by the enabling technologies in order to achieve the desired requirements in terms of latency or reliability. This grouping represents a novel taxonomy used to organize the enhancements of the papers considered in this survey. We expect most future papers to also be classified according to this taxonomy, summarized in Figure 2.

1) DATA PLANE MANAGEMENT

One of the most immediate way to deal with latency and reliability in end-to-end protocols such as TCP, UDP or SCTP over 5G networks is to modify the basic mechanisms for managing the data flow in these protocols. Such modifications include a) the large literature on **congestion control** mechanisms in protocols such as TCP for wireless networks, b) changes in **retransmission** algorithms for the early confirmation or deletion of unnecessary ACKs (possible in the lower radio level), or c) intelligent **traffic shaping** to prioritize certain types of traffic.

Another popular method in the revised literature is implementing a smart **scheduling** of packet delivery over one or multiple connections and/or interfaces in three different ways: a) partitioning packets in several chunks or tasks to be sent over a single connection (**scheduling packets**), b) using multiple connections over a single interface and selecting one or several connections to send the packet (**scheduling paths**), and c) similar techniques to b) but using a multi-homed device with several interfaces and conducting selection (and possible duplication) considering the interfaces (**scheduling interfaces**).

The last relevant method in data plane management is **caching**. Caching is based on storing frequently accessed data content and routing popular requests in order to reduce the retrieval delay. This technique is mainly applied in the ICN context and it results in a significant end-to-end latency improvement.

2) TRANSPORT PROTOCOL ENHANCEMENT

Some research efforts focus on the development of transport protocols to comply with novel requirements. A common technique is to start from a well-known tested single-path protocol and extend its capabilities to support multi-connectivity, in order to enhance the reliability, throughput and further KPIs. We refer to this category as **extension for multi-connectivity**. Another technique is based on a flexible stack that could select different protocols regarding parameters, requirements or the network state. This **protocol selection** is usually combined with context awareness information and could help reduce the latency and enhance the reliability.

3) CODING

In data transmissions, coding refers to sending information with some modifications in order to enhance communication. Most of the time, this coding is performed with redundancy in such a way that data can be received in different forms and decoded at the destination. The two main coding forms considered in this survey are **forward error correction (FEC)** and **network coding**. Forward error correction is an end-to-end technique used for the detection and correction of a

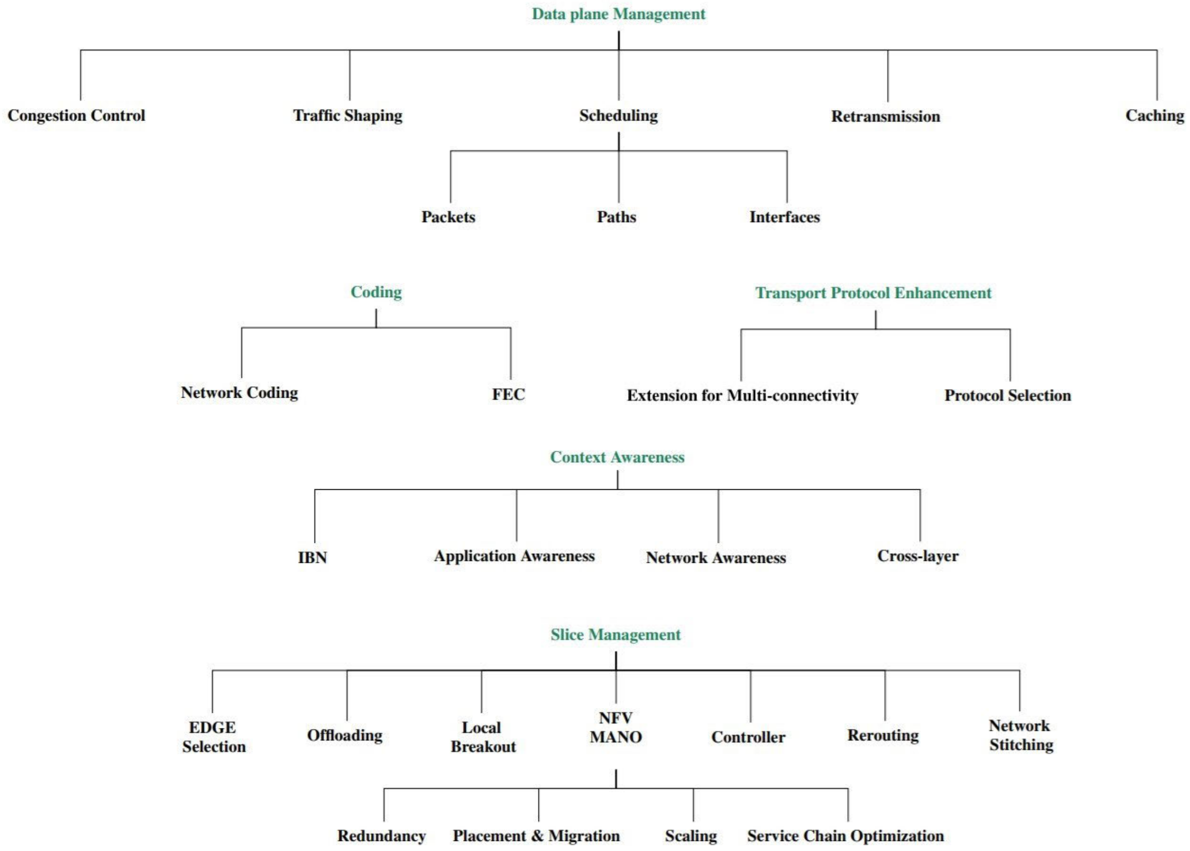

FIGURE 2. **The taxonomy of common methods and techniques for latency and reliability.**

limited number of errors over noisy communication channels without performing retransmissions. The proper use of this method could enhance the reliability, throughput or latency levels [68]. In addition, network coding allows intermediate nodes, such as routers, to send some data information coded in different packets. In most of its variants, if just a sufficient number of packets arrive at the destination, the original message can be decoded [69].

4) CONTEXT AWARENESS

Context awareness involves taking advantage of information outside of a protocol's or layer's scope to enhance its utilization and provide better service in terms of the general performance or concrete KPIs such as latency and reliability. The 5G network is expected to be completely context-aware [70], creating interest in this research field (for instance, through the creation of research groups such as the recently established Path Awareness Networking Research Group [71]).

Context information can be extracted from different sources such as higher layers or lower layers of the protocol stack. **Application awareness** means monitoring the application status of the information flow from the application layer in order to work in the lower layers to improve the latency, reliability, throughput or general performance. Likewise, **network awareness** means considering the network conditions or configurations to make decisions about parameters and the usage of certain higher layer protocols or applications. Furthermore, protocol stacks can benefit from the interaction between different layers. This exchange of information is not always about the network state or application requirements. A **cross-layer** approach would mean exposing information between layers and working conjointly to fulfil different requirements at execution time. Finally, as an alternative to the runtime monitoring of traffic or states, **intent-based networking (IBN)** is a concept first defined by Cisco [72] that consists on taking into account user preferences (intents) and applying a logical intelligence to map them or translate them into policies that can be applied in the current protocol, network or operating scope.

5) SLICE MANAGEMENT

In general, a network slice is a logical division of the network to isolate resources in order to maintain a certain level of quality (e.g., latency and reliability) for specific users and services. Since the 5G network slice is end-to-end, most of the common methods related to slice management are connected to the management of the enabling technologies in the category network support, such as EDGE, SDN and NFV.

The main objective of EDGE is to reduce the latency by placing itself closer to end users. In large networks, more than one EDGE is possible. A proper **EDGE selection** technique would reduce the distance between end users and would result in enhanced latencies. **Local breakout (LBO)** is a promising solution based on determining whether to send data packets through the central core network or through closer destinations (e.g., EDGE, local nodes, etc.) in order to reduce the excessive delay in the core network load. In addition, **offloading** is a network solution based on leveraging the processing or execution of tasks to the network. This solution usually partitions tasks and is highly coupled with EDGE since they allow easy deployment in the closer places of the network.

Another series of common methods and techniques in the literature are oriented to enhancing **network function virtualization management and orchestration** (in short, **NFV MANO**). Virtual network functions (VNFs) require management to enhance their utilization. This management and orchestration can be summarized in three main points. The first point is that NFV decouples software functionalities from physical equipment to offer better functionality. However, this software still needs a proper platform for its execution. Optimized selection of VNF **placement and migration** of services would mean offering a better service with enhanced KPIs. Moreover, the nature of NFV allows VNFs to be simultaneously deployed in different parts of a network. This **redundancy** provides better service in terms of the reliability and even the latency. Finally, resources are allocated to VNFs in terms of CPU cores and memory, among others. A proper dynamic allocation or **scaling** of these resources at runtime would enhance the overall communication and reduce potential overload. Finally, within NFV MANO, VNFs are placed as part of a **service chain**. Several methods, such as refactoring, pipelining, using parallelism, etc. resulting in enhanced communication with better latency, throughput or reliability, are studied to optimize this task.

Regarding the role of SDN in the network slice, the papers in this area focus on two problems. First, in SDN, proper **controller placement** would reduce the latency between SDN nodes. This reduced latency could impact end-to-end data when they need to be sent to the controller. Second, **rerouting** or dynamically changing routing tables (usually possible thanks to SDN switches) would provide sufficient network flexibility so as to adapt to network or application requirements.

Last, another relevant slice management method is **network stitching** or slice stitching. This is an operation that modifies the functionality of an existing slice by adding and merging the functions of another slice [73] in order to enhance the overall operations or concrete KPIs such as the reliability and latency; meanwhile, service chain optimization is based on the fact that in many network services, data pass through sequences of functions that are common to other services (e.g., firewalls, encryption, etc.).

D. PARAMETERS EVALUATED

Some parameters have been selected to help in the evaluation of the contributions. These parameters were extracted from the 5G key performance indicators [26] and other relevant parameters studied in the literature.

1) KEY PERFORMANCE INDICATORS (KPIs)

Key performance indicators (KPIs) are measurements of specific network properties that help in monitoring, optimizing and characterizing services. Some well-defined 5G-PPP KPIs have been taken from the 5Genesis[3] project [75]. These KPIs have been set as goals, and their different contributions have been evaluated as plausible enablers (in the tables of subsequent sections). The KPIs under study are the enhancement of latency (low latency), the increase in reliability (high reliability) and the improvement of the throughput (high throughput). The two first KPIs are essential in the communications under study, while the third one is due to the increase in the number of new applications such as UHD video transmission that continue to demand high throughput (apart from ultra-reliability and low latency), turning throughput into a desirable characteristic in most critical communications that share video content.

2) OTHER PARAMETERS

In addition to KPIs, there are a couple of qualities considered interesting and able to characterize the contributions.

- **Partial Reliability:** Sometimes latency is achieved by sacrificing reliability. This sacrifice does not necessarily make the communication unreliable, as some critical data transmissions will continue focusing on reliability, while other data can tolerate loss in favour of lower latency [76], [77]. Partial reliability does not strictly meet all critical communication requirements but can meet the demands of certain types of reliable low-latency communications, making it an interesting feature with which to characterize contributions.
- **Heterogeneous Networks:** An increasingly large number of different technologies with diverse characteristics coexist in current networks (e.g., WiFi, LTE, and 5G). In some occasions, protocols and other network solutions have to indistinctively use these technologies or even use them together through interface diversity. A protocol's ability to work properly, to work with fairness, to adapt to changes, etc., under these conditions of heterogeneity is a remarkable added value [78].

IV. ANALYSIS OF THE STATE-OF-THE-ART

In this section, the variety of solutions grouped by enabling technologies and APIs are presented in different subsections, plus several tables offering a better understanding of the contributions are presented. Table 3 presents the protocol comparison of single-path, multipath and multicast solutions. Table 4, Table 5, Table 6 and Table 7 show the EDGE, SDN, NFV and ICN network support solutions, respectively. Then, Table 8 presents the API solutions. These tables collect the

[3] European Union Framework Programme Horizon 2020, Grant Agreement N° 815178 (5Genesis) [74]

variety of contributions studied in this survey and present further information in terms of the main methods adopted and the parameters on which the contributions focus. However, in order to maintain clarity in this analysis of the state-of-the-art Section, the evaluation of the possible relevant concentrations in terms of methods, KPIs or enabling technologies and APIs is not made until Section V, which presents further tables and a tree diagram to contribute to the evaluation.

A. END-TO-END PROTOCOLS

Single-path, multipath and multicast protocol contributions are presented in Table 3. Single-path protocols greatly focus on low latency and high throughput, while reliability is often ignored or addressed only partially. In contrast, the main focus of multipath protocols is the reliability; however, they also have a large number of solutions to increase the throughput and even address latency several times. In the case of multicast protocols, the focus is on reliability, owing to the fact that most solutions take advantage of sending multiple copies of information. On some occasions, this redundancy also helps to improve throughput; however, latency is not considered extensively.

1) SINGLE-PATH PROTOCOLS

The transmission control protocol (TCP) [125] is one of the most important protocols of the Internet; hence, it has been common to conduct research on the enhancement of the TCP. For instance, Petlund [79] presents TCP and SCTP modifications to satisfy the requirements of interactive and thin-stream applications (low latency in small packet transmissions) such as games [126]; ER TCP Pert [80] is a solution that combines the delay-based TCP and early transmission to improve the performance in delivering real-time media by reducing the latency caused by retransmissions; TCP-ROME [81] is a transport-layer framework that allows establishing and coordinating multiple many-to-one TCP connections, increasing the reliability in streaming multimedia; and Massaro *et al.* [82], [127] implemented an algorithm based on TCP Vegas [128] to improve the coexistence of TCP and UDP data with high throughput and low latency in heterogeneous flows (multimedia applications).

Although the TCP does not meet the requirements of new technologies, it is not at all obsolete. The TCP is used all over the Internet, which encourages research on enhancements for 5G networks. First, some studies adapting the TCP to different cellular networks are those of Polese *et al.* [83], who study TCP enhancements though link-layer retransmissions to improve the TCP for 5G mmWave networks in terms of latency and throughput; and Petrov *et al.* [84], who present an advanced TCP version for 5G with the purpose of increasing the throughput rate and improving the reliability levels through enhancing the TCP friendliness, TCP recovery from time outs and the drop rate.

Another series of studies focused on improving the general performance of the TCP to make it suitable for all kinds of communications are the following. Google LLC [85] presents a congestion control algorithm (TCP bottleneck bandwidth and round-trip propagation time) that responds to the actual congestion rather than the packet loss and thus improves the throughput, latency and quality of experience. Gambhava *et al.* [86] present a discrete TCP (DTCP), an enhancement that differentiates slow start and congestion avoidance phases while tuning the data flow over a transport connection, resulting in an improvement in TCP performance in heterogeneous networks. Zhu *et al.* [87] present a TCP optimization using radio awareness which yields a significant gain in both latency and throughput setting parameters of the TCP layer and modifying the TCP congestion control mechanism according to the cross-layer information. Finally, Luo *et al.* [88] study an extension of TCP/IP, called explicit congestion notification (ECN), that helps realize low latency in the TCP. They present standardization efforts and propose an improved ECN as an enabler of ultra-low latency and high throughput.

Apart from the TCP, there is also research on additional communication protocols mainly focused on improving latency. First, there are some contributions regarding novel transport protocols: ASAP [89] is a transport protocol that reduces latency, eliminating unnecessary RTTs in the handshake and cutting the delay of small requests by up to two-thirds; the short-term reliable protocol for low-latency video transmission [90] relies on packet retransmission for only a limited amount of time to reduce latency and make it optimal for image/video communication; Cheng *et al.* [91] develop PrefCast, a preference-aware protocol used to satisfy user preferences for content objects achieving the required latency-critical VR game demands with reduced network usage; and Park *et al.* [92] present a simple protocol solution for video transmission based on the RUDP (reliable user datagram protocol) to provide low latency with short-term reliability.

Additionally, there are research efforts on congestion and rate control enhancements: SCReAM [93] is a window-based and byte-oriented congestion control protocol for RTP streams that achieves improvements in both video latency and throughput in real-time communications thanks to its adaptation ability, whereas Mittal *et al.* [94] propose a framework for rate control with the similar objective of improving the throughput and latency. Finally, there exists a work aiming to improve current architectures such as the Low Latency, Low Loss, Scalable Throughput (L4S) Internet Service Architecture [95]. The L4S Architecture is a solution that enables low latency, low loss and a scalable throughput in novel applications coexisting on shared network bottlenecks. It aims to break network ossification and calls for evolution, making it possible to run scalable transport protocols such as the DCTCP [129] and MDTCP [130] over the same access networks as those of non-scalable transport protocols such as TCP CUBIC/Reno.

2) MULTIPLE CONNECTIVITY AND MULTIPATH PROTOCOLS

Multiple connectivity technologies can be classified according to the layer in which they perform the aggregation of

TABLE 3. Comparison of protocol solutions.

		Main Methods[4]	KPI			Other Parameters	
			Low Latency	High Reliability	High Throughput	Partial Reliability	HetNet Support
Single-path	**TCP/SCTP modifications [79]**	DP Mgmt: Retransmission	✓			✓	
	ER TCP Pert [80]	DP Mgmt: Retransmission	✓			✓	
	TCP-ROME [81]	Scheduling: Paths		✓	✓		✓
	TCP VEGAS* [82]	DP Mgmt: Congestion Control	✓		✓	✓	✓
	TCP 5G mmWave Networks [83]	Ctxt Awa: Cross-layer DP Mgmt: Retransmission		✓	✓		
	Advanced 5G-TCP [84]	DP Mgmt: Congestion Control		✓	✓		✓
	TCP BBR [85]	DP Mgmt: Congestion Control	✓		✓	✓	
	Gambhava et al. [86]	DP Mgmt: Congestion Control		✓	✓		✓
	Zhu et al. [87]	Ctxt Awa: Cross-layer DP Mgmt: Congestion Control	✓		✓		✓
	Luo et al. [88]	DP Mgmt: Congestion Control	✓		✓		
	ASAP [89]	Ctxt Awa: Cross-layer	✓				
	STRP [90]	DP Mgmt: Retransmission	✓			✓	
	PrefCast [91]	Scheduling: Packets Ctxt Awa: IBN	✓				
	Park et al. [92]	DP Mgmt: Retransmission	✓			✓	
	SCReAM [93]	DP Mgmt: Congestion Control	✓		✓	✓	✓
	Mittal et al. [94]	Ctxt Awa: Cross-layer	✓		✓		
	L4S Architecture [95]	DP Mgmt: Congestion Control	✓		✓		✓
Multipath	**LISP-HA [96]**	Scheduling: Paths		✓	✓		✓
	Yap et al. [97]	Scheduling: Paths Ctxt Awa: Application		✓			✓
	Singh et al. [98]	Scheduling: Interfaces		✓	✓		✓
	Gonzalez-Muriel et al. [99]	Scheduling: Interfaces		✓	✓		✓
	MPTCP [100]	TP Enh.: Extension for MC		✓	✓		
	NC-MPTPC [101]	Coding: Network Coding Scheduling: Paths		✓	✓		✓
	FMTCP [102]	Coding: Network Coding Scheduling: Paths	✓	✓	✓		✓
	Hurtig et al. [103]	Scheduling: Paths	✓		✓		✓
	QoS-MPTPC [104]	Scheduling: Packets	✓			✓	
	ADMIT [105]	Coding: FEC	✓	✓	✓		✓
	PR-MPTCP+ [106]	Scheduling: Paths Ctxt Awa: Network	✓	⋆	✓	⋆	✓
	MPFlex [107]	Scheduling: Paths		✓	✓		
	Lee et al. [108]	Scheduling: Paths		✓	✓		
	Multipath PERT [109]	Scheduling: Paths	✓	✓	✓		✓
	Multipath QUIC [110]	TP Enh.: Extension for MC	✓	✓	✓		✓
	MPRTP [111]	TP Enh.: Extension for MC	✓	✓	✓		
	EMSTP [112]	TP Enh.: Extension for MC Coding: FEC	✓			✓	✓
	MPMTP [113]	TP Enh.: Extensions for MC Coding: FEC	✓		✓	✓	✓
	HMTP [114]	Coding: FEC		✓	✓		✓
	MPMTP-AR [115]	Scheduling: Paths		✓	✓		
	m^{2}CMT [116]	Scheduling: Paths		✓	✓		✓
	A-CMT [117]	Scheduling: Paths		✓	✓		✓
Multicast	**Zhu et al. [118]**	Scheduling: Paths		✓	✓		✓
	Tsimbalo et al. [119]	Coding: Network Coding		✓			✓
	Xiong et al. [120]	Scheduling: Paths		✓	✓		
	Chi et al. [121]	Coding: Network Coding DP Mgmt: Retransmission		✓			✓
	Roger et al. [122]	Ctxt Awa: Cross-layer	✓				
	ECast [123]	DP Mgmt: Retransmission		✓			
	Mahajan et al. [124]	Ctxt Awa: Cross-layer DP Mgmt: Congestion Control		✓			✓

independent flows in a multipath flow. In the following subsections, we provide a study of the state-of-the-art, focused on contributions of the IP and higher layers. When multi-connectivity is performed in the application layer, the application itself has to be aware of the paths and manage them [131]. The disadvantage found in this method is that every application that wants to benefit from multiple connectivity must be adapted and modified. Thus, one typical approach to manage lower-layer information in the application layer is to use application programming interfaces, a technique studied in the following Section IV-C due to its connection to context awareness.

- **IP layer**: Some recent contributions are the following. Locator/ID Separation Protocol - Hybrid Access [96] allows simultaneous usage of multiple access both upstream and downstream. It uses information about the packet loss and delay to improve the load balancing, bandwidth and resilience. Yap *et al.* [97] propose an algorithm to improve the scheduling of packets over multiple interfaces. Singh *et al.* [98] develop a framework for optimal traffic aggregation in multi-RAT (radio access technology) heterogeneous wireless networks. In addition, Gonzalez-Muriel *et al.* [99] present an implementation of LWIP, which consists of LTE-WLAN aggregation at the IP level, aiming to enhance the bandwidth and reliability. Their results show that the throughput increases without degrading the latency or increasing the packet loss.

[4] These methods and techniques are the ones introduced in Section III-C. The acronyms and abbreviations used for presenting those methods in the table can be found after Section VI.

- **Transport layer**: Going up to the transport layer, there is a large variety of protocols based on the idea of multiple connectivity. The most remarkable multi-connectivity protocol in the transport layer is the multipath TCP. The multipath TCP is a transport layer protocol extension of the TCP and standardized as experimental in January 2013 in the IETF RFC 6824 [100]. It tries to overcome some of the TCP limitations and to improve it with a higher quality of service, robustness, better performance, network decongestion, etc., using multiple paths. Based on these benefits, certain uses of the protocol such as the offloading of networks, mobility, the migration of virtual machines in a wide area, etc., have been foreseen. The MPTCP has been proven to perform better than the TCP when using paths with similar characteristics [132], [133], but it fails to outperform it in heterogeneous networks [134], [135].

  Due to the limitations of the multipath TCP, some papers have conducted research on its improvement. The NC-MPTCP [101] and fountain code-based multipath TCP (FMTCP) [102] utilize network coding to boost the overall goodput and outperform the MPTCP in the case of highly dissimilar subflow conditions. Hurtig *et al.* [103] present two novel scheduling techniques for the MPTCP (BLEST and STTF) that are shown to reduce latency when interfaces have asymmetric capacity and delay. The QoS-MPTPC [104], ADMIT [105] and PR-MPTCP+ [106] are extensions for interactive video, video streaming and real-time multimedia, respectively. Finally, MPFlex [107] is a flexible software architecture that enhances MPTCP scheduling and policies thanks to the use of multiplexing.

  One approach in cellular networks is to bring the MPTCP to 5G networks. The 3GPP 5G mobile core features ATSSS (access traffic steering, switching and splitting) and has already standardized the MPTCP as a foundational capability in 3GPP Release 16 [136]. Research labs such as Tessares [137] and CableLabs [138] are already working on the implementation of this 5G ATSSS functionality and bringing MPTCP contributions into 3GPP, respectively. For instance, Lee *et al.* [108] develop an offloading control scheme to make the MPTCP suitable for 5G NR and LTE networks, reducing the packet loss rate and enhancing the throughput in these upcoming networks.

  Nonetheless, the MPTCP is not the only protocol developed for multi-connectivity. There is a wide range of protocols regarding this matter. First, a set of protocols aiming to improve real-time communications or streaming (latency) can be found. They may be based on a TCP, such as the multipath PERT [109]. The multipath probabilistic early response TCP is a solution suitable for real-time data transfer that provides high throughput and efficient load balancing. However, the majority of these protocols are based on the UDP, such as the multipath QUIC [110], [139]. The MPQUIC is a protocol based on the QUIC that takes advantage of UDP features to provide lower latency and of multi-connectivity improvements to provide higher reliability and resilience. Multiple parallel paths for the RTP (MPRTP) [111] are also UDP-based and increase the reliability and throughput to enhance the user experience compared to the RTP. The energy-aware multipath streaming transport protocol (EMSTP) [112] aims to support high-quality streaming over heterogeneous networks working with UDP subflows as well as with Raptor codes. Furthermore, the multipath multimedia transport protocol (MPMTP) [113] works, similar to the EMSTP, using Raptor codes to support a seamless high-quality video streaming service over wireless networks, the difference being that it uses both TCP subflows and UDP subflows to manage the control information and data, respectively.

  There is also a set of protocols geared at improving the throughput, utilization or general performance instead of just the latency for real-time communications. The heterogeneous multipath transport protocol (HMTP) [114] is a protocol based on fountain codes, which recovers the original data if a sufficient number of packets are received regardless of their arrival order, solving the receive buffer blocking problem. The multipath message transport protocol based on the application-level relay (MPMTP-AR) [115] works in a multipath transport system based on the application-level relay (MPTS-AR) framework [140] to deliver reliable data service over multiple paths with high efficiency, throughput and resilience. Finally, concurrent multipath transfer for the SCTP (CMT-SCTP) [141] approaches such as m^{2}CMT [116] and A-CMT [117] exploit the multi-homing capability of the SCTP to improve its performance with multi-connectivity.

3) MULTICAST TECHNOLOGIES

Most of the work on multicasting focuses on reliability. For instance, Zhu *et al.* [118] present a new multicast protocol called the MCTCP. This protocol aims to outperform the state-of-the-art reliable multicast schemes by managing the multicast groups in a centralized manner and reactively scheduling flows to optimal links. The MCTCP achieves improvements in both reliability and throughput compared with the original and TCP-SMO schemes, (an alternative single-source multicast optimization scheme). In addition, the work of Tsimbalo *et al.* [119] considers a lossy multicast network in which reliability is provided by means of random linear network coding. Specifically, they utilize random linear network coding and verify that the mean square error in their tests can be as low as 9×10^{-5}.

Moreover, when the aim is to develop different network architectures through multicasting, the focus is also on reliability. Xiong *et al.* [120] present MTM, a novel reliable multicast for data centre networks. MTM improves the error resilience ability in the presence of various levels of packet loss and provides high application throughput. Chi *et al.* [121] propose enhancing multicast transmissions by means of D2D-communication-based retransmission. They propose an efficient reliable multicast scheme for 5G networks that utilizes D2D communication and network coding to achieve 100 percent reliability. However, with the expansion of critical communications, we can find some recent works also aiming to improve the latency, such as that of Roger *et al.* [122]. They address the challenges imposed by 5G V2X (vehicle-to-everything) services in terms of latency and reliability, which generally cannot be guaranteed using the current MBMS (multimedia broadcast multicast services) architecture, and propose a low-latency multicast scheme to decrease the end-to-end communication latency, ensuring, at the same time, the correct operation of high-demand services.

Another approach is to combine multicast technologies with other enabling technologies to enhance their capabilities. For instance, Zhang *et al.* [123] present an OpenFlow-enabled elastic loss recovery solution, called ECast, for reliable multicasting that uses elastic area multicast to enhance the retransmission of multicast recovery packets, whereas Mahajan *et al.* [124] design and implement a platform named ATHENA that enables multicast in SDN-based data centres, providing high reliability and, at the same time, congestion control mechanisms to ensure fairness.

B. NETWORK SUPPORT

The presented protocols must be integrated and collaborate with other enabling technologies to reduce latency and to increase reliability. A network should be able to provide different mechanisms flexibly to achieve the desired operation. The four technologies selected as the main ones to provide this network assistance in current and future networks are edge computing, software-defined networking, network function virtualization and information-centric networking.

1) EDGE COMPUTING (EDGE)

As shown in Table 4, latency reduction is the goal in every edge computing contribution. The proximity of EDGE to devices reduces the end-to-end distance between the end sides of communication, resulting in an enhanced latency. Nevertheless, reliability and throughput improvements are not usually considered in this technology.

Several studies focus on bringing edge computing to current and future cellular networks. Garcia *et al.* [142] introduce the idea of fog and edge computing to LTE networks. They introduce two new elements: the fog gateway and the GTP gateway [143]. These new components allow the processing of specific services on the edge, preventing all traffic from reaching the core and resulting in an improvement of up to 78% in terms of latency reduction. Zhang *et al.* [144] present a mobility-aware edge computing framework for emerging 5G applications such as IoT for intelligent transportation, intelligent healthcare, etc. The solution speeds up the application response (latency), improves the user experience, reduces congestion, increases the speed of data, etc., and exposes critical challenges for EDGE that still need to be addressed, such as further improvement of the efficiency and security. In addition, Piran *et al.* [145] propose a context-aware streaming over 5G HetNets (CASH) video streaming framework that allocates the resources in an intelligent manner based on Edge-UE communication and the actual requirements of the content and network characteristics, outperforming the existing works in terms of the peak data rate, latency, user experience and spectral efficiency.

An increasingly important solution in edge computing is the ‘‘distributed SGW with local breakout (SGW-LBO)’’ approach. It stems from the desire of operators to have greater control of the traffic that needs to be steered [159], and the idea behind it is to control the redirection of data planes. Some examples of contributions improving communications by means of this method are the following: Lee *et al.* [146] propose a local breakout of mobile access network traffic in base stations by MEC to reduce end-to-end latency, whereas Cattaneo *et al.* [147] combine MEC and NFV to deploy CPU-intensive applications and enhance the latency of the immersive video use case.

EDGE is usually combined with novel technologies such as SDN or NFV to take its performance to the next level. Such is the case of Huang *et al.* [148], who implement an SDN-based MEC framework solution for LTE/LTE-A. The solution is compliant with the ETSI and 3GPP architectures and enables latency reduction and traffic offloading. Heinonen *et al.* [149] present a prototype of a 5G network slice that selects the mobility anchor during an attach procedure from the closest network edge (and re-evaluates it in each handover). The selection of the optimal network edge node results in a decrease in the end-to-end latency. Schiller *et al.* [150] develop an NFV/EDGE/SDN platform that uses VNFs to flexibly manage EDGE applications and improve the user QoE (e.g., latency and throughput). Yang *et al.* [151] propose a solution to take advantage of the low latency benefit of edge computing without wasting resources during stable/low-workload periods of the fixed-location traditional solution. They adopt network function virtualization in edge computing to create a dynamic resource allocation framework, offering flexibility in hosting MEC services in any virtualized network node, which consequently reduces the cost by up to 33% compared to existing solutions. Finally, Cziva *et al.* [152] combine edge computing with virtualization, deploying VNFs (virtual network functions) in different scenarios. Their results show that using edge servers can deliver up to a 70% improvement in user-to-VNF latency.

Moreover, EDGE is also combined with different methods to exploit context awareness. Nunna *et al.* [153] propose combining novel communication architectures of 5G with mobile edge computing to provide ultra-low latency data transmissions. This MEC integration at the edge of 5G networks provides a robust real-time context-aware collaboration platform. Dutta *et al.* [154] combine EDGE, QoE awareness and the NFV technology to create an edge-assistive transcoding and adaptative streaming that ensures reduced latency and

TABLE 4. EDGE proposals.

	Main Methods	KPI			Other Parameters	
		Low Latency	High Reliability	High Throughput	Partial Reliability	HetNet Support
Garcia et al. [142], [143]	Ctxt Awa: Cross-layer Slice Mgmt: Local Breakout	✓				
Zhang et al. [144]	Slice Mgmt: Offloading	✓				✓
CASH [145]	Ctxt Awa: Application Ctxt Awa: Network	✓		✓		✓
Lee et al. [146]	Slice Mgmt: Local Breakout	✓				
Cattaneo et al. [147]	Slice Mgmt: Local Breakout	✓				✓
Huang et al. [148]	Ctxt Awa: Network Ctxt Awa: Cross-layer	✓				
Heinonen et al. [149]	Slice Mgmt: EDGE Selection	✓				✓
Schiller et al. [150]	DP Mgmt: Caching	✓		✓		✓
Yang et al. [151]	VNF MANO: Scaling	✓				✓
Cziva et al. [152]	VNF MANO: Plcmt & Migr.	✓				✓
Nunna et al. [153]	Ctxt Awa: Network	✓				
Dutta et al. [154]	Slice Mgmt: Offloading	✓				
Taleb et al. [155]	Slice Mgmt: EDGE Selection VNF MANO: Plcmt & Migr.	✓				
Edgent [156]	Slice Mgmt: Offloading	✓				
Maier et al. [157]	Slice Mgmt: Offloading	✓	✓	✓		✓
Liu et al. [158]	Slice Mgmt: Offloading	✓	✓			✓

better quality of experience. Finally, Taleb *et al.* [155] proposes an approach to enhance users' experience by bringing MEC to smart cities. They aim to ensure ultra-short latency through a smart architecture that allows applications/services to follow the mobility of users.

Another interesting and recent tendency in EDGE is edge intelligence, where big data analytics and edge computing are combined to provide near-real-time analysis of data. Some works on edge intelligence include those of Li *et al.* [156] and Maier *et al.* [157]. Li *et al.* [156] propose Edgent, a collaborative and on-demand deep neural network (DNN) co-inference framework with device-edge synergy. Their prototype implementation and evaluations demonstrate the effectiveness of Edgent in enabling on-demand low-latency edge intelligence. Maier *et al.* [157] propose the utilization of edge intelligence to achieve a low-latency FiWi-enhanced mobile network. Their solution makes use of machine learning in the context of FiWi-enhanced heterogeneous networks to decouple haptic feedback from the impact of propagation delays and, ultimately, enable an ultra-low latency tactile Internet.

It can be seen that most of the works on EDGE focus mainly on latency. In fact, as highlighted by Liu *et al.* [158], mobile edge computing research lacks a focus on reliability, the complementary aspect of the critical communications under study. Thus, they propose a framework and algorithms to strike a good balance between latency and reliability, offloading tasks from a single UE to multiple edge nodes.

2) SOFTWARE-DEFINED NETWORKING (SDN)

The SDN proposals are introduced in Table 5. They focus mostly on a decrease in latency, similar to EDGE solutions, although in some papers, this technology also considers reliability and throughput enhancements.

Software-defined networking is a solution for current and future networks. Pagé *et al.* [160] propose a modification to 4G architecture by integrating SDN to achieve low latency. The idea is that the SGW would be replaced by SDN switches with routing algorithms and intelligence to improve the network performance. Costa-Requena *et al.* [161] deploy a modular SDN-based user plane in real testbeds for 5G. This platform allows optimized transport for low latency, throughput and reliability, with EDGE taking advantage of slicing and the flexibility of SDN. Additionally, J. Wang *et al.* [162] design an SDN framework for a smart factory based on an industrial Internet of things (IIoT) system. Their method is based on computing mode selection (CMS) and the execution of sequences based on task priority, achieving real-time performance and high reliability.

A large part of SDN research for new applications focuses on improving rerouting and controller placement to lower the latency of multimedia applications. Lakiotakis *et al.* [163] create a collaboration between the network and network music performance (NMP) applications to reduce the delay by up to 59% over the traditional solutions. SDN is used to increase the performance during link congestion and perform rerouting or send orders to applications to modify the audio processing configuration. Awobuluyi *et al.* [56] present a holistic SDN control plane approach to multimedia transmission. A QoE and context-aware application make decisions regarding rerouting, load balancing and adapting flows in an SDN network to achieve the required ultra-reliable low-latency video streaming. Garg *et al.* [164] present an SDN framework combined with edge computing and QoS awareness to enhance the routing capabilities and mobility management of autonomous vehicles (AVs). The performance assessment reports an overall improvement in terms of the end-to-end delay. Furthermore, Wang *et al.* [165] study the placement of SDN controllers to shorten the latency between controllers and switches in wide-area networks. They present the concepts of network partitioning and a clustering-based network partitioning algorithm, which result in a reduced maximum latency between controllers and their associated switches.

TABLE 5. SDN proposals.

	Main Methods	KPI			Other Parameters	
		Low Latency	High Reliability	High Throughput	Partial Reliability	HetNet Support
Pagé et al. [160]	DP Mgmt: Traffic Shaping	✓				
Costa-Requena et al. [161]	DP Mgmt: Traffic Shaping	✓	✓	✓		
Wang et al. [162]	Slice Mgmt: EDGE Selection Slice Mgmt: Offloading	✓	✓			
Lakiotakis et al. [163]	Ctxt Awa: Network Slice Mgmt: Rerouting	✓				
Awobuluyi et al. [56]	Ctxt Awa: Network Slice Mgmt: Rerouting	✓	✓	✓		
Garg et al. [164]	Slice Mgmt: EDGE Selection Slice Mgmt: Offloading	✓	✓			
Han et al. [59]	Slice Mgmt: Controller Placement	✓				
G. Wang et al. [165]	Slice Mgmt: Controller Placement	✓				
Yap et al. [166]	Slice Mgmt: Network Stitching Scheduling: Paths	✓	✓	✓		✓
RLMD [167]	Slice Mgmt: Controller Placement	✓	✓			

Nonetheless, while most solutions focus on latency, a combination of multiple connectivity and SDN could also be optimal, providing enhancements in both reliability and latency. Yap *et al.* [166] present a distribution across multiple interfaces using SDN, enhancing the multipath with dynamic selection intelligence, and Hu *et al.* [167] develop a reliable and load-balance-aware multi-controller deployment (RLMD) strategy to address the controller placement selection and explore the reliable deployments of the controllers. Their simulations show a better performance in improving the reliability of the control plane and balancing the distribution of the controller loads.

3) NETWORK FUNCTION VIRTUALIZATION (NFV)

Table 6 shows the similar tendencies of NFV and SDN. The contributions studied in this survey focus in both cases on reducing latency, while reliability and throughput are also considered but at a lower level.

Some NFV contributions focus on improving the current and future networks. Raza *et al.* [168], [169] present a vIMS (virtualized IP multimedia subsystem) design that refactors network function modules and results in significant improvements in both latency and reliability (compared with the existing 3GPP IMS implementation). Qu *et al.* [170] develop a series of algorithms to reduce the delay in network service chains for NFV-enabled data centre networks. The algorithms presented can reduce up to 18.5% the average end-to-end delay and increase the reliability from 7.4% to 14.8%. Furthermore, Mekikis *et al.* [171] work on an NFV-enabled experimental platform for 5G tactile Internet support in industrial environments and demonstrate that their setup can achieve the end-to-end communication latency required for this kind of application.

There are also proposals to enhance the NFV operation itself. Ding *et al.* [172] present an enhancement of the existing redundancy method for NFV architectures. The proposed CERA algorithm achieves a better estimation for the services, resulting in higher reliability and higher cost efficiency. Nascimento *et al.* [173] propose an acceleration mechanism for NFV platforms, aiming to improve their performance and scalability. Their results show an enhancement in both latency and cost efficiency, and the goal of higher throughput is presented as future research. Cho *et al.* [174] address the VNF migration problem for low network latency among VNFs and develop a novel VNF migration algorithm (VNF real-time migration) to minimize the network latency in rapidly changing network environments (an up to 70.90% network latency reduction ratio). Sun *et al.* [175] present an NFV framework that enables network function parallelism to improve NFV performance. This network function parallelism describes and orchestrates NF chaining intents to achieve significant latency reduction for real-world service chains. Finally, Fan *et al.* [176] present GREP (guaranteeing reliability with enhanced protection), a novel algorithm developed to guarantee high reliability in NFV while minimizing resource consumption. Their evaluation shows that GREP performs reliable service function chain (SFC) mapping in NFV networks, minimizing the amount of resources allocated to SFC requests while meeting clients' SLA requirements.

EDGE and SDN, the previous two technologies presented, can be combined with NFV, as shown by Huang *et al.* [148], Yang *et al.* [151], Costa-Requena *et al.* [161], etc. Other relevant contributions combining NFV with other technologies include the works of Bekkouche *et al.* [177], Valsamas *et al.* [178] and Yao *et al.* [179]. Bekkouche *et al.* [177] propose an extended framework for the management and orchestration of unmanned aerial vehicles (UAVs). The framework combines NFV and MEC with the functionalities of a UAV traffic management system to satisfy the end-to-end latency requirement without fully compromising the reliability. Finally, Valsamas *et al.* [178] and Yao *et al.* [179] present network slicing platforms that support different interconnected services and achieve improved latency and reliability, respectively.

4) INFORMATION-CENTRIC NETWORKING (ICN)

ICN is an approach used to leave behind the point-to-point paradigm and evolve the Internet infrastructure. Data become independent from the location, application or storage to enable desirable features that can enhance the informa-

TABLE 6. NFV proposals.

	Main Methods	KPI			Other Parameters	
		Low Latency	High Reliability	High Throughput	Partial Reliability	HetNet Support
Raza et al. [168], [169]	NFV MANO: SC Optimization	✓	✓			
Qu et al. [170]	NFV MANO: Plcmt & Migr.	✓	✓	✓		
Mekikis et al. [171]	NFV MANO: Scaling	✓	✓	✓		✓
Ding et al. [172]	NFV MANO: Redundancy		✓			
Nascimento et al. [173]	NFV MANO: SC Optimization	✓				
Cho et al. [174]	NFV MANO: Plcmt & Migr.	✓				✓
Sun et al. [175]	NFV MANO: SC Optimization	✓		✓		
Fan et al. [176]	NFV MANO: Redundancy NFV MANO: SC Optimization		✓			
Bekkouche et al. [177]	Slice Mgmt: EDGE Selection Slice Mgmt: Offloading NFV MANO: Plcmt & Migr.	✓			✓	
Valsamas et al. [178]	Slice Mgmt: Network Stitching	✓				✓
Yao et al. [179]	Slice Mgmt: Network Stitching		✓			✓

TABLE 7. ICN proposals.

	Main Methods	KPI			Other Parameters	
		Low Latency	High Reliability	High Throughput	Partial Reliability	HetNet Support
Liang et al. [180]	NFV MANO: Plcmt & Migr. DP Mgmt: Caching	✓				✓
Carofiglio et al. [181]–[183]	DP Mgmt: Caching	✓				
Zhang et al. [184]	DP Mgmt: Caching	✓				
Sardara et al. [185]	Ctxt Awa: Cross-layer	✓		✓	✓	✓
Dannewitz et al [186]	DP Mgmt: Caching	✓				✓
Wang et al. [187]	Ctxt Awa: Cross-layer	✓	✓	✓		
Vakilina et al. [188]	DP Mgmt: Congestion control	✓	✓	✓		✓

tion distribution [48]. The intention of ICN contributions in the scientific literature is generally to enhance the latency KPI, although reliability and throughput are occasionally addressed as well (see Table 7).

Some relevant ICN contributions in novel cellular 5G networks are as follows. Liang *et al.* [180] presents an ICN over 5G networks approach based on improving the end-to-end network performance by integrating ICN techniques with wireless network virtualization. They develop some key components for the architecture to enhance resource allocation and catching and ultimately minimize traffic and latency. Carofiglio *et al.* [181]–[183] develop LAC and later FOCAL, an approach combining novel caching and forwarding strategies to preferentially route popular content requests through the optimal path. Their evaluation shows that FOCAL reduces the end-user-experienced latency. Zhang *et al.* [184] present a ICN-based caching approach that considers both the mobility of users and the popularity of videos to reduce the retrieval delay caused by frequent handoffs in 5G networks. Another work along the same line is that of Sardara *et al.* [185], who present a transport layer solution and socket API for ICN, providing a better throughput rate as well as latency in these novel networks.

ICN solutions usually focus their efforts on enhancing latency, as in the work of Dannewitz et al [186], which presents an architecture that achieves low latencies through efficient caching and a scalable name resolution service. Nonetheless, although reliability is not the main focus of the enhancement in the latest ICN research efforts, some pertinent contributions in that direction are those of Wang *et al.* [187] and Vakilina *et al.* [188]. Wang *et al.* [187] propose a reliable hop-by-hop transport mechanism for ICN that guarantees the content reliability in packets and forwards all the received packets downstream so that the end-to-end latency can be remarkably decreased. Meanwhile, Vakilina *et al.* [188] develop a distributed algorithm at the backhaul and an SDN-based centralized algorithm aimed at minimizing congestion and enhancing latency levels without sacrificing reliability.

C. APPLICATION PROGRAMMING INTERFACES

Application programming interfaces are key enablers that provide additional control and flexibility for protocol stacks to enhance the performance. As shown in Table 8, the majority of API contributions presented help to reduce the latency and increase the throughput and reliability at the same time.

The traditional socket API has been questioned for a long time. In fact, works such as Sockets++ [203], Florissi *et al.* [204], Abbasi *et al.* [205] and Reuther *et al.* [206] present some of the first relevant enhancements of the socket API. However, they are too far from the 5G requirements studied in this survey. Newer contributions that try to fit new tendencies such as multiple connectivity, as enablers of the demands on latency and reliability, are Jones *et al.* [65] and Trammell *et al.* [66]. Jones *et al.* [65] propose raising the datagram API to be able to implement some missing features in the current socket datagram API: establishing connectivity, control over the QoS (reliability and congestion control) and support for multiple interfaces; while Trammell *et al.* [66] propose a new API solution based on message carriers and policies to make it an independent

TABLE 8. API proposals[5].

	Main Methods	KPI			Other Parameters	
		Low Latency	High Reliability	High Throughput	Partial Reliability	HetNet Support
Jones et al. [65]	TP Enh.: Protocol Selection Ctxt Awa: IBN	⋆	⋆	⋆	⋆	
Trammell et al. [66]	TP Enh.: Protocol Selection	⋆	⋆	⋆	⋆	⋆
Scharf et al. [189]	Ctxt Awa: Cross-layer		✓	✓		
Hesmans et al. [190]	Scheduling: Paths		✓	✓		✓
Hesmans et al. [191]	Scheduling: Paths		✓	✓		✓
NEAT [192], [193]	Ctxt Awa: IBN Ctxt Awa: Network TP Enh.: Protocol Selection	⋆	⋆	⋆	⋆	⋆
TAPS [194]	TP Enh.: Protocol Selection	⋆	⋆	⋆	⋆	⋆
Nielsen et al. [195]	Scheduling: Paths	✓	✓			✓
Multi-sockets [196]	Ctxt Awa: Cross-layer TP Enh.: Protocol Selection	✓		✓	✓	✓
Msocket [197]	TP Enh.: Protocol Selection	⋆	⋆	⋆	⋆	⋆
Socket Intents [198]	Ctxt Awa: IBN Scheduling: Paths	✓		✓		
Chronos [199]	Ctxt Awa: Cross-layer	✓		✓		
Belay et al. [200], [201]	Ctxt Awa: Cross-layer	✓		✓		
Siddiqui et al. [202]	Ctxt Awa: IBN	⋆	⋆	⋆	⋆	⋆

platform and transport protocol (and to support multipaths if necessary).

Because the MPTCP is the most extended multipath protocol, there are API solutions developed specifically to enhance its utilization, reliability and throughput levels. For instance, Scharf *et al.* [189] present a simple extension of the TCP interface for MPTCP-aware applications; Hesmans *et al.* [190] propose raising the MPTCP path manager to provide control over multipath TCP decisions and path management to applications, resulting in energy savings and enhancements in backup mode, streaming and flow selection; and Hesmans *et al.* [191] propose an enhanced socket API for the multipath TCP that enables application programmers to control the MPTCP and enhance the operation of the underlying stack.

Another approach also based on APIs is the NEAT (new, evolutive API and transport-layer architecture) solution [192], [193]. NEAT uses application requirements such as reliability or latency levels not to choose parameters or interfaces but to directly select different transport protocols. It aims to break the current ossification of the Internet transport architecture, enabling the incremental flexible deployment of new transport services and features. One of its latest improvements is to supply applications with detailed network information, creating a more complex and complete API. The development of NEAT was continued with the contribution TAPS (an architecture for transport services) [194], exposing transport protocol features to applications for a flexible network communication. Nielsen *et al.* [195] also take advantage of both network monitoring measurements and user objectives to improve interface diversity (selecting optimal interfaces through weighted KPIs, such as latency or reliability). Higgins *et al.* [196] present a similar approach with Multi-sockets, a solution that uses the knowledge of application needs to select interfaces efficiently, enhancing latency and throughput. Finally, Msocket [197] is an extension of the Berkeley socket API for supporting multiple stacks. When there are multiple distinct TCP/IP stacks available, it allows the application to specify which one to use for the communication. This exposure allows different behaviours according to the network requirements, permissions, QoS demands and levels of protection.

Other API solutions relevant to this survey but not directly related to multi-connectivity or multiple stacks are works such as Schmidt *et al.*'s [198]. They developed Socket Intents, a socket solution for the purpose of managing user and application information (e.g., small and sensitive packet delay, background traffic, etc.) to select network parameters, resulting in the enhancement of KPIs such as latency or throughput. Along the same line, Kapoor *et al.* [199] propose Chronos, a framework that can deliver predictable, low latency in data centre applications. This framework uses a combination of techniques for that purpose, one of which is a user-level networking API that supports efficient load balancing, a kernel bypass, etc., to reduce the latency in data centre networks. The contribution of Belay *et al.* [200], [201] presents a data plane operating system that provides high performance. The data plane architecture works with a native API to optimize both the latency and bandwidth, managing and dedicating hardware threads and networking queues to data plane instances. The solution results in significant improvements in both the end-to-end latency and throughput. Finally, Siddiqui *et al.* [202] present a requirement-based API as an abstraction layer to make applications independent of network mechanisms. The aim is to reduce the coupling between applications and underlying protocols and to evolve into a future Internet architecture flexible enough to adapt to an application's requirements.

V. EVALUATION AND CHALLENGES

A. EVALUATION

In this evaluation, we present a table, some graphs and a diagram to characterize the contributions and discuss open research topics. Specifically, we follow three classification

[5] When ⋆ is used, it means that the parameter can be achieved at the cost of another marked parameter.

TABLE 9. Proposals combining multiple enabling technologies.

References	Multipath Protocols	Multicast Protocols	EDGE	SDN	NFV	ICN
[151] [161] [150] [207] [171] [149]			✓	✓	✓	
[146] [147] [152] [154] [177]			✓		✓	
[124] [118] [120] [123] [167]		✓		✓		
[142] [148] [162] [164]			✓	✓		
[166] [56]	✓			✓		
[170]	✓			✓	✓	
[173]				✓	✓	
[158]	✓		✓			
[183]		✓	✓			✓
[184]			✓			✓
[188]			✓	✓		✓
[185]		✓				✓
[180]					✓	✓

criteria to report some conclusions: first, we analyse the combination of different enabling technologies and evaluate API contributions; then, we study the KPI concentrations and coverage according to each approach; and finally, we reflect on the different methods used in the literature.

1) COMBINATION OF ENABLING TECHNOLOGIES AND API EVALUATION

First, due to the overlap experienced while presenting different enabling technologies, Table 9 was generated to achieve a better understanding. It identifies the relationships between the enabling technologies and shows the contributions that combine multiple enabling technologies. This table does not include the category ‘‘single-path protocols’’ since it would not contribute to the evaluation and would add a large number of unnecessary entries in the table (every communication needs a protocol; thus, when it is not a multipath or multicast, it is usually single path).

Regarding the table content, note the strong link between EDGE, SDN and NFV. There are 5 contributions on EDGE and SDN, 5 articles on EDGE and NFV, 2 papers incorporating SDN and NFV, and 6 additional contributions combining all three. Likewise, the association between multicast protocols and SDN and works very well with 5 contributions combining both technologies. Finally, it is worth highlighting the strong coupling between ICN and other technologies with 5 contributions (of the 7 contributions of ICN studied in this paper) combined with others technologies. One of the most relevant combinations is ICN and EDGE due to the capabilities of the latter in providing a platform for caching support.

Furthermore, API contributions are commonly combined with the different enabling technologies due to their ubiquitous nature. There is in fact an especially strong coupling between API and multipath protocols owing to the fact that API provides better control of the different paths and protocol decisions. Some examples of this are contributions such as NEAT [192], [193], TAPS [194] and Schmidt *et al.* [198], among others. However, API is not only combined with multipath protocols, and many other contributions include API as a means of achieving the requirements without being the main topic. This is the case of Nunna *et al.* [153], Taleb *et al.* [155], Mahajan *et al.* [124] and Sardara *et al.* [185], among others.

Overall, API has proven itself to be a valuable approach to enhance the latency and reliability. A great variety of papers on the topic assess the increased reliability and the decreased latency without ignoring the support of other KPIs and parameters as we evaluate in next subsection V-A2. Because of everything that has been stated above, we consider API to be a promising topic that will be discussed in Section V-B.

2) KPI COVERAGE

Figure 3 shows some graphs summarizing the enabling qualities found in each technology and in APIs based on the contributions analysed in this survey. Each technology is presented in a radar chart or spider chart that sets the line closer to the edges (which represent each KPI) proportional to the percentage of contributions that focus on that KPI. Therefore, in the literature under study, we can see generally a high level of treatment of the latency KPI, followed by a medium-high level of treatment of the reliability. Throughput and HetNet support are parameters frequently addressed while partial reliability is not commonly a main point of study in contributions aimed at enhancing the latency or reliability. As enablers, APIs, multipath protocols and single-path protocols are the approaches most frequently used to address the different KPIs. API addresses all parameters with at least a medium-high level of coverage. In multipath protocols, there is a high level of work aimed at improving reliability, throughput and addressing HetNet support, with a medium level on latency; meanwhile, in single-path protocols, latency and throughput are the main points, but HetNet support and partial reliability are also addressed several times. Each remaining technology has a clearly differentiated main topic: in EDGE, SDN and NFV, it is the latency; in ICN, it is also the latency with additional high importance given to HetNet support; and in multicast protocols, it is the reliability with also a relevant importance of HetNet support. This does not mean that the solutions related to these topics cannot help to meet the requirements of other parameters, such as we can see in the increasing importance of high reliability on SDN or NFV; nevertheless, it means that the latest research efforts have been mostly aimed in these directions.

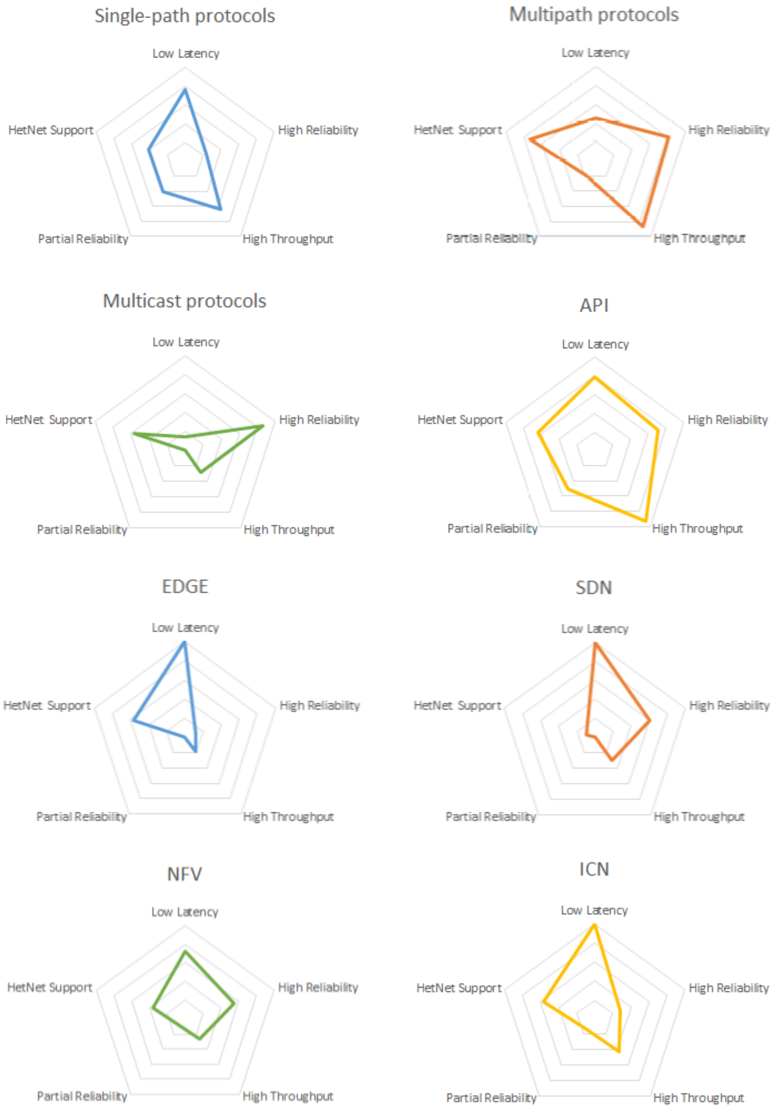

FIGURE 3. Mapping of API and the enabling technologies used to classify parameters based on their contributions.

All things considered, in our study of the state-of-the-art, we can see that each enabling technology has its strengths and weaknesses, which can be combined to achieve the desired requirements, as discussed in Section V-B. However, to guarantee the KPI levels presented in the Introduction (Section I), all end-to-end levels have to contribute. Currently, the state-of-the-art focuses on partial solutions that do not allow a complete end-to-end validation, and thus it is not possible to perform an evaluation with concrete numerical values. Ongoing research efforts along this line include projects such as 5Genesis, one of the main objectives of which is the creation of a 5G full-stack environment [74], [208].

3) ON THE COMMON METHODS AND TECHNIQUES

Focusing on the underlying methods or techniques used to achieve low latency and/or high reliability, Figure 4 gives a clear picture of their use in existing literature. One thing learned from this picture is that they reuse and enhance existing mechanisms to improve 5G, but there is some small novelty in the new mechanisms for 5G. Many of the methods are classic versions of basic mechanisms coming from fixed networks (congestion control or cross-layer) or techniques used to optimize end-to-end communications in previous 4G mobile communication networks (coding, context awareness, EDGE enhancements and multi-connectivity). Only a small number are specifically focused on 5G, like NFV or SDN methods, but they are also widely used in cloud (not wireless) environments. This first observation confirms that the core methods for end-to-end solutions in different network technologies have some continuity and there are improvements and adaptations to new networks, but 5G does not mean a hard break when considering the latency and reliability. More revolutionary techniques are probably in the radio access part,

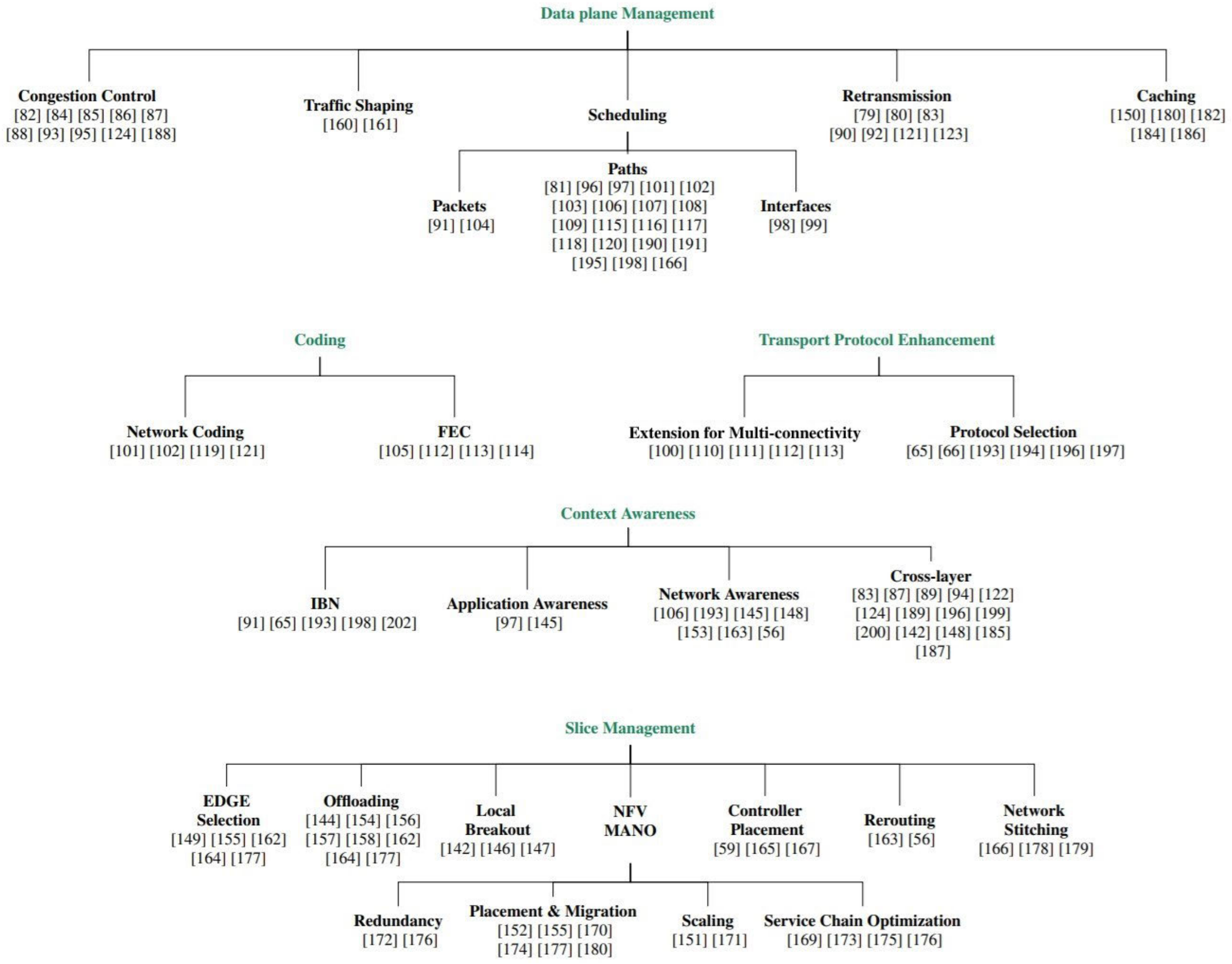

FIGURE 4. Classification of the literature according to the proposed taxonomy.

but they are not part of the survey because we limit this survey to end-to-end solutions.

The figure also reflects the most popular topics: scheduling paths, cross-layer, congestion control and EDGE support (for both offloading and selection). However, there are some methods with only a few related papers that we see as having more potential, probably combined with some of the most popular methods, such as automatic scaling in the NFV domain and IBN. We discuss some of them from our perspective in the promising research topics discussion (Section V-B).

B. PROMISING RESEARCH TOPICS

We have identified a number of topics connected with possible future challenges and research work.

1) EDGE COMPUTING

First, we highlight the importance of edge computing in latency constrained communications. It cannot be forgotten that there is an inevitable latency that comes with distance. Protocols, technologies and physical layer development are necessary to enhance the latency; however, EDGE is an essential starting point of any attempt to reduce the latency. Applications and services must deploy their instances in edge clouds (closer to the user than the core networks) to be able to provide a reasonable low-latency communication. This deployment has to be done in an efficient manner; therefore, EDGE has to rely on further technological innovations.

2) APIs

One of the key enhancers for EDGE is the use of context information through APIs. Communications with stringent requirements such as low latency and high reliability require a flexible network and protocols that can be adapted. Knowledge about network, application information and the use of the IBN can help a protocol to choose the parameters, characteristics and options that will optimize the communication. These APIs could manage this information and be used to determine the EDGE placement and different configurations. In some use cases, such as the vehicular use case presented in the Introduction (Section I), information about the position of the other cars, their speed and additional parameters (standardized or not) are critical to maintain a safe environment.

3) MULTI-CONNECTIVITY

One technology that can be combined with the API to enhance its capabilities is multi-connectivity. Multi-connectivity is one of the most important solutions for improving reliability, taking advantage of multiple paths when possible and combining their potentials. Current multi-connectivity solutions rely mostly on schedulers that decide how to use these different paths. However, using context information, an application or a transport layer could be able to determine which scheduling algorithm to select or, ultimately, how to use the different paths offered to enhance the communication and provide sufficient reliability, latency, throughput, etc.

4) 5G END-TO-END SELF-ORGANIZING NETWORKS

Combining all these ideas, we highlight the 5G end-to-end self-organizing network open research topic. We are familiar with the idea of self-organizing networks, where networks configure themselves (mostly configuring their antennas) to be able to adapt to the user's conditions. In fact, this solution was presented in Section II and discarded as a main point of study of this survey due to its closer relation to lower-layer development. However, the suggested solution is to take this idea a step further to develop a 5G end-to-end network that is able to organize itself. This means that the application would be able to self-organize and select where to be deployed (EDGE or core), what NFVI (network functions virtualization infrastructure) or presence points to use, or what configuration to select. The network, on its behalf, would be able to self-organize itself and configure characteristics about NFV technologies (or slicing) and the SDN configuration (and paths). This 5G end-to-end self-organizing network would rely on a massive amount of context information that would be extracted with the help of APIs.

In general, this solution means a large development in several areas that must work jointly to succeed. First, there has to be an orchestrator to reorganize the services, make decisions about them and deploy them where necessary (a management and orchestrator entity if we refer to NFV); then, there is also a need to develop flexible novel protocols that change end-to-end connections to different endpoints without affecting user experience; and finally, API solutions have to be designed to collect information from lower and higher layers and offer them to the correspondent entity that needs them. All of this assumes a sufficient lower-layer infrastructure support.

Some research is being conducted in a direction similar to this suggested solution, i.e., the aforementioned work of 5Genesis [74]. The 5Genesis project is developing a 5G end-to-end network that implements EDGE, SDN and NFV. Moreover, API solutions are being developed to expose context-aware information and manage different configurations such as multi-connectivity. All things considered, the current 5Genesis solution is a first step towards a 5G end-to-end self-organizing network; however, there is still much research and development to be done using the presented technologies and additional techniques such as, for instance, artificial intelligence and machine learning. In fact, we still see a lack of papers using machine learning and artificial intelligent methods to reduce latency or increase reliability in a closed loop way. These approaches are much more complex than the SON methods used in the RAN segment in 3G and 4G and imply a hard modelling of the network to represent the reference behaviour KPIs. Some interesting contributions that are leading the research in this direction are Balevi *et al.* [209], Morocho-Cayamcela *et al.* [210] or the aforementioned work of Eurecom [150].

VI. CONCLUSION

In this paper, we presented a comprehensive survey of end-to-end solutions for 5G reliable low-latency communications. The main topic is the need to enhance the Internet and higher layers with our research efforts focused on end-to-end protocols, network support and APIs.

The solutions studied were selected based on technologies with plausible future perspectives such as novel protocols, multipath protocols, multicast protocols, EDGE, SDN, NFV, ICN and APIs; and they were characterized by the enhancement of latency, reliability or some other relevant parameters such as the throughput, HetNet support and partial reliability. In addition, we extracted the common methods used by the contributions in order to analyse current trends.

We identified some lines of research regarding these enabling technologies and additional aspects and focused on the idea of a 5G end-to-end self-organizing network combining edge computing, APIs, multi-connectivity, NFV, etc. Projects such as 5Genesis [74] aim to contribute in that direction, by considering most of the presented technologies to create a 5G end-to-end network.

LIST OF ACRONYMS

3GPP 3rd Generation Partnership Project

4G 4th Generation

5G 5th Generation

5GMF Fifth Generation Mobile Communications Promotion Forum

5G-PPP 5G Infrastructure Public Private Partnership

ACK Acknowledgement

A-CMT Adaptive Concurrent Multipath Transfer

ADMIT QuAlity-Driven MultIpath TCP

API Application Programming Interface

ASAP Accelerated Secure Association Protocol

ATSSS Access Traffic Steering, Switching and Splitting

AV Autonomous Vehicle

BBR Bottleneck Bandwidth and Round-trip propagation time

BLER Block Error Rate

BLEST Block Estimation Scheduler

CASH Context-Aware Streaming over 5G HetNets

CDN Content Delivery Network
CMS Computing Mode Selection
CMT-SCTP Concurrent Multipath Transfer for SCTP
CPS Cyber-physical System
Ctx Awa Context Awareness
D2D Device to Device
DCTCP Data Centre TCP
DNN Deep Neural Network
DP Data plane
DTCP Discrete TCP
ECN Explicit Congestion Notification
EDGE Edge Computing
eMBB Enhanced Mobile Broadband
EMSTP Energy-aware Multipath Streaming Transport Protocol
eNodeB Evolved Node B
ER Early Retransmission
ETSI European Telecommunications Standards Institute
FEC Forward Error Correction
FiWi Fibre-Wireless
FMTCP Fountain code-based Multipath TCP
FOCAL Forwarding and Caching with Latency Awareness
GREP Guaranteeing Reliability with Enhanced Protection
GTP General Packet Radio Service Tunnelling Protocol Gateway
HetNet Heterogeneous Network
HMTP Heterogeneous Multipath Transport Protocol
IBN Intent-Based Networking
ICN Information-Centric Networking
IEEE Institute of Electrical and Electronics Engineers
IETF Internet Engineering Task Force
IIoT Industrial Internet of Things
IMS IP Multimedia Subsystem
IoT Internet of Things
IP Internet Protocol
IPsec Internet Protocol security
KPI Key Performance Indicator
L4S Low Latency, Low Loss, Scalable Throughput
LAC Latency-Aware Caching
LBO Local Breakout
LISP-HA Locator/ID Separation Protocol - Hybrid Access
LL-MEC Low-Latency Multi-access Edge Computing
LLC Limited Liability Company
LTE Long-Term Evolution
LTE-A LTE Advanced
LWIP LTE-WLAN Integration using IPsec Tunnel
m^{2}CMT Modified mobile CMT
MBMS Multimedia Broadcast Multicast Service
MC Multi-connectivity
MDTCP Multipath Transport Protocol for Telco Cloud Data Centres
MEC Mobile Edge Computing
Mgmt Management
Migr. Migration
mmWave Millimetre Wave
MPFlex Multipath Flexible Software Architecture
MPMTP Multipath Multimedia Transport Protocol
MPMTP-AR Message Transport Protocol based on Application-level Relay
MPRTP Multiple parallel paths for RTP
MPTCP Multipath Transmission Control Protocol
MPTS-AR Multipath Transport System based on Application-Level Relay
ms Millisecond
MTC Machine-Type Communication
MTM Multiple Trees Multicast
NC-MPTPC Network Coding MPTCP
NEAT New Evolutive API and Transport-layer architecture
NF Network Function
NFV MANO NFV Management and Orchestration
NFV Network Function Virtualization
NFVI Network Function Virtualization Infrastructure
NMP Network Music Performance
NR New Radio
PERT Probabilistic Early Response
Plcmt Placement
PR-MPTCP+ Context-aware QoE-oriented MPTCP Partial Reliability extension
QoE Quality of Experience
QoS Quality of Service
QUIC Quick UDP Internet Connection
RAN Radio Access Network
RAT Radio Access Technology
RFC Request for Comments
RG Research Group
RLMD Reliable and Load balance-aware Multi-controller Deployment
ROME Real-time On-line Multimedia Environment
RTP Real-time Transport Protocol
RTT Round-Trip Time
RUDP Reliable UDP
SC Service Chain
SCReAM Self-Clocked Rate Adaptation for Multimedia
SCTP Stream Control Transmission Protocol
SDN Software-Defined Networking
SFC Service Function Chain
SGW Serving Gateway
SLA Service Level Agreement
SMO Single-source Multicast Optimization
SON Self-Organizing Network
STRP Short-Term Reliable Protocol
STTF Shortest Transfer Time First Scheduler
TAPS An Architecture for Transport Services
TCP Transmission Control Protocol
TI Tactile Internet
TP Enh. Transport Protocol Enhancement
UAV Unmanned Aerial Vehicle
UDP User Datagram Protocol
UE User Equipment
UHD Ultra Hi-Definition
URLLC Ultra-Reliable Low-Latency Communications

V2X Vehicle to Everything
vIMS virtualized IMS
VNF Virtual Network Function
VR Virtual Reality
WG Working Group
WLAN Wireless Local Area Network

REFERENCES

[1] S. Mccreary and K. Claffy, "Trends in wide area IP traffic patterns—A view from ames internet exchange," in *Proc. 13th ITC Specialist Seminar Internet Traffic Meas. Modeling*, Monterey, CA, USA, Jan. 2000.

[2] M. Zhang, M. Dusi, W. John, and C. Chen, "Analysis of UDP traffic usage on Internet backbone links," in *Proc. 9th Annu. Int. Symp. Appl. Internet*, Jul. 2009, pp. 280–281.

[3] D. Lee, B. E. Carpenter, and N. Brownlee, "Media streaming observations: Trends in UDP to TCP ratio," *Int. J. Adv. Syst. Meas.*, vol. 3, nos. 3–4, pp. 1–16, 2010.

[4] G. Papastergiou, G. Fairhurst, D. Ros, A. Brunstrom, K.-J. Grinnemo, P. Hurtig, N. Khademi, M. Tuxen, M. Welzl, D. Damjanovic, and S. Mangiante, "De-ossifying the Internet transport layer: A survey and future perspectives," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 619–639, 1st Quart., 2017.

[5] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, and B. Vucetic, "Ultra-reliable low latency cellular networks: Use cases, challenges and approaches," *IEEE Commun. Mag.*, vol. 56, no. 12, pp. 119–125, Dec. 2018.

[6] O. N. C. Yilmaz, "Ultra-reliable and low-latency (URLLC) 5G communication," in *Proc. EuCNC*, Jun. 2016, pp. 1–2. [Online]. Available: http://kom.aau.dk/~nup/2016-06-27_Yilmaz-5G%20Ultra-reliable-Low-latency_final.pdf

[7] *Technical Specification Group Services and System Aspects; Service Requirements for the 5G System; Stage 1 (Release 16)*, document TS 22.261, Version 16.0.0, 3GPP, Jun. 2017. [Online]. Available: http://www.3gpp.org/ftp//Specs/archive/22_series/22.261/22261-g00.zip

[8] G. Pocovi, H. Shariatmadari, G. Berardinelli, K. Pedersen, J. Steiner, and Z. Li, "Achieving ultra-reliable low-latency communications: Challenges and envisioned system enhancements," *IEEE Netw.*, vol. 32, no. 2, pp. 8–15, Mar. 2018.

[9] Z. Yuan, J. Jin, L. Sun, K.-W. Chin, and G.-M. Muntean, "Ultra-reliable IoT communications with UAVs: A swarm use case," *IEEE Commun. Mag.*, vol. 56, no. 12, pp. 90–96, Dec. 2018.

[10] *Technical Specification Group Services and System Aspects; Feasibility Study on New Services and Markets Technology Enablers for Critical Communications; Stage 1 (Release 14)*, document TR 22.862, 3GPP, Version 14.1.0, Sep. 2016. [Online]. Available: http://www.3gpp.org/ftp/Specs/archive/22_series/22.862/22862-e10.zip

[11] 3GPP, "Technical specification group radio access network; study on enhanced LTE support for aerial vehicles (release 15), version 15.0.0," 3rd Gener. Partnership Project (3GPP), Tech. Rep. (TR) 36.777, Dec. 2017. [Online]. Available: http://www.3gpp.org/ftp//Specs/archive/36_series/36.777/36777-f00.zip

[12] O. Holland, E. Steinbach, R. V. Prasad, Q. Liu, Z. Dawy, A. Aijaz, N. Pappas, K. Chandra, V. S. Rao, S. Oteafy, M. Eid, M. Luden, A. Bhardwaj, X. Liu, J. Sachs, and J. Araújo, "The IEEE 1918.1 'tactile Internet'-standards working group and its standards," *Proc. IEEE*, vol. 107, no. 2, pp. 256–279, Feb. 2019.

[13] *Technical Specification Group Services and System Aspects; Service Requirements for Cyber-Physical Control Applications in Vertical Domains; Stage 1 (Release 17)*, document TS 22.104, Version 17.2.0, 3GPP, Dec. 2019. [Online]. Available: http://www.3gpp.org/ftp//Specs/archive/22_series/22.104/22104-h20.zip

[14] 3GPP, "Technical specification group services and system aspects; service requirements for the 5G system; stage 1 (release 16), version 17.1.0," 3rd Gener. Partnership Project (3GPP), Tech. Specification (TS) 22.261, Dec. 2019. [Online]. Available: http://www.3gpp.org/ftp//Specs/archive/22_series/22.261/22261-h10.zip

[15] *Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s)*, document ITU-R M.2410-0, Nov. 2017. [Online]. Available: https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2410-2017-PDF-E.pdf

[16] *Study on New Radio Access Technology Physical Layer Aspects*, document TR 38.802, Version 14.2.0, 3GPP, Sep. 2017.

[17] Ericsson Inc., "5G systems: Enabling the transformation of industry and society," Ericsson, Stockholm, Sweden, White Paper UEN 284 23-3251 rev B, Jan. 2017. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/white-papers/5g-systems–enabling-the-transformation-of-industry-and-society

[18] Cisco Systems Inc., "White paper: Time-sensitive networking: A technical introduction," Cisco Public, San Jose, CA, USA, Tech. Rep. C11-738950-00, 2017. [Online]. Available: https://www.cisco.com/c/dam/en/us/solutions/collateral/industry-solutions/white-paper-c11-738950.pdf

[19] C. Cruces, R. Torrego, A. Arriola, and I. Val, "Deterministic hybrid architecture with time sensitive network and wireless capabilities," in *Proc. IEEE 23rd Int. Conf. Emerg. Technol. Factory Automat. (ETFA)*, vol. 1, Sep. 2018, pp. 1119–1122.

[20] N. Finn, "Introduction to time-sensitive networking," *IEEE Commun. Standards Mag.*, vol. 2, no. 2, pp. 22–28, Jun. 2018.

[21] Z. MacHardy, A. Khan, K. Obana, and S. Iwashina, "V2X access technologies: Regulation, research, and remaining challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1858–1877, 3rd Quart., 2018.

[22] A. Fotouhi, H. Qiang, M. Ding, M. Hassan, L. G. Giordano, A. Garcia-Rodriguez, and J. Yuan, "Survey on UAV cellular communications: Practical aspects, standardization advancements, regulation, and security challenges," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3417–3442, 4th Quart., 2019.

[23] M. Gharibi, R. Boutaba, and S. L. Waslander, "Internet of drones," *IEEE Access*, vol. 4, pp. 1148–1162, 2016.

[24] R. J. Hall, "An Internet of drones," *IEEE Internet Comput.*, vol. 20, no. 3, pp. 68–73, May 2016.

[25] J. Brun, F. Safaei, and P. Boustead, "Managing latency and fairness in networked games," *Commun. ACM*, vol. 49, no. 11, pp. 46–51, Nov. 2006.

[26] 5G-PPP. (2018). *6th Global 5G Event: 5G Technology Changing Pardigms of a New Society*. [Online]. Available: https://5g-ppp.eu/6th-global-5g-event-5g-technology-changing-pardigms-of-a-new-society/

[27] G. J. Sutton, J. Zeng, R. P. Liu, W. Ni, D. N. Nguyen, B. A. Jayawickrama, X. Huang, M. Abolhasan, Z. Zhang, E. Dutkiewicz, and T. Lv, "Enabling technologies for ultra-reliable and low latency communications: From PHY and MAC layer perspectives," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2488–2524, 3rd Quart., 2019.

[28] S. Zhang, X. Xu, Y. Wu, and L. Lu, "5G: Towards energy-efficient, low-latency and high-reliable communications networks," in *Proc. IEEE Int. Conf. Commun. Syst. (ICCS)*, Nov. 2014, pp. 197–201.

[29] A. Morgado, K. M. S. Huq, S. Mumtaz, and J. Rodriguez, "A survey of 5G technologies: Regulatory, standardization and industrial perspectives," *Digit. Commun. Netw.*, vol. 4, no. 2, pp. 87–97, Apr. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352864817302584

[30] R. N. Mitra and D. P. Agrawal, "5G mobile technology: A survey," *ICT Express*, vol. 1, no. 3, pp. 132–137, Dec. 2015.

[31] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart., 2016.

[32] A. Gupta and R. K. Jha, "A survey of 5G network: Architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206–1232, 2015.

[33] T. O. Olwal, K. Djouani, and A. M. Kurien, "A survey of resource management toward 5G radio access networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1656–1686, 3rd Quart., 2016.

[34] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, "5G backhaul challenges and emerging research directions: A survey," *IEEE Access*, vol. 4, pp. 1743–1766, 2016.

[35] L. Chettri and R. Bera, "A comprehensive survey on Internet of Things (IoT) toward 5G wireless systems," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 16–32, Jan. 2020.

[36] A. Nasrallah, A. S. Thyagaturu, Z. Alharbi, C. Wang, X. Shao, M. Reisslein, and H. ElBakoury, "Ultra-low latency (ULL) networks: The IEEE TSN and IETF DetNet standards and related 5G ULL research," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 88–145, 1st Quart., 2019.

[37] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3098–3130, 4th Quart., 2018.

[38] S. Habib, J. Qadir, A. Ali, D. Habib, M. Li, and A. Sathiaseelan, "The past, present, and future of transport-layer multipath," *J. Netw. Comput. Appl.*, vol. 75, pp. 236–258, Nov. 2016.

[39] M. Li, A. Lukyanenko, Z. Ou, A. Ylä-Jääski, S. Tarkoma, M. Coudron, and S. Secci, "Multipath transmission for the Internet: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2887–2925, 4th Quart., 2016.

[40] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 4th Quart., 2017.

[41] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, and W. Wang, "A survey on mobile edge networks: Convergence of computing, caching and communications," *IEEE Access*, vol. 5, pp. 6757–6779, 2017.

[42] I. Al-Anbagi, M. Erol-Kantarci, and H. T. Mouftah, "A survey on cross-layer quality-of-service approaches in WSNs for delay and reliability-aware applications," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 525–552, 1st Quart., 2016.

[43] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1657–1681, 3rd Quart., 2017.

[44] Ö. Yürür, C. H. Liu, Z. Sheng, V. C. M. Leung, W. Moreno, and K. K. Leung, "Context-awareness for mobile sensing: A survey and future directions," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 68–93, 1st Quart., 2016.

[45] M. S. Elbamby, C. Perfecto, M. Bennis, and K. Doppler, "Toward low-latency and ultra-reliable virtual reality," *IEEE Netw.*, vol. 32, no. 2, pp. 78–84, Mar. 2018.

[46] B. Briscoe, A. Brunstrom, A. Petlund, D. Hayes, D. Ros, I.-J. Tsang, S. Gjessing, G. Fairhurst, C. Griwodz, and M. Welzl, "Reducing Internet latency: A survey of techniques and their merits," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 2149–2196, 3rd Quart., 2016.

[47] K. Antonakoglou, X. Xu, E. Steinbach, T. Mahmoodi, and M. Dohler, "Towards haptic communications over the 5G tactile Internet," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3034–3059, 4th Quart., 2018.

[48] D. Kutscher, S. Eum, K. Pentikousis, I. Psaras, D. Corujo, D. Saucez, T. Schmidt, and M. Waehlisch, *Information-Centric Networking (ICN) Research Challenges*, document RFC 7927, IRTF, 2016, pp. 1–38.

[49] M. Handley, "Why the Internet only just works," *BT Technol. J.*, vol. 24, no. 3, pp. 119–129, Jul. 2006.

[50] J. Qadir, A. Ali, K.-L.-A. Yau, A. Sathiaseelan, and J. Crowcroft, "Exploiting the power of multiplicity: A holistic survey of network-layer multipath," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2176–2213, 4th Quart., 2015.

[51] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Exploiting caching and multicast for 5G wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2995–3007, Apr. 2016.

[52] G. Araniti, M. Condoluci, P. Scopelliti, A. Molinaro, and A. Iera, "Multicasting over emerging 5G networks: Challenges and perspectives," *IEEE Netw.*, vol. 31, no. 2, pp. 80–89, Mar. 2017.

[53] 5G-XCast. (2018). *5G-XCast Project*. [Online]. Available: http://5g-xcast.eu/

[54] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing—A key technology towards 5G," Eur. Telecommun. Standards Inst. (ETSI), Sophia Antipolis, France, ETSI White Paper 11, Sep. 2015. [Online]. Available: https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp11_mec_a_key_technology_towards_5g.pdf

[55] K. Dolui and S. K. Datta, "Comparison of edge computing implementations: Fog computing, cloudlet and mobile edge computing," in *Proc. Global Internet Things Summit (GIoTS)*, Jun. 2017, pp. 1–6.

[56] O. Awobuluyi, J. Nightingale, Q. Wang, and J. M. Alcaraz-Calero, "Video quality in 5G networks: Context-aware QoE management in the SDN control plane," in *Proc. IEEE Int. Conf. Comput. Inf. Technol.; Ubiquitous Comput. Commun.; Dependable, Autonomic Secure Comput.; Pervas. Intell. Comput.*, Oct. 2015, pp. 1657–1662.

[57] Ericsson. (2020). *Cloud SDN*. [Online]. Available: https://www.ericsson.com/en/portfolio/digital-services/cloud-infrastructure/cloud-sdn

[58] Nokia. (2020). *Software-Defined Access Networks*. [Online]. Available: https://www.nokia.com/networks/solutions/software-defined-access-networks/

[59] L. Han, Z. Li, W. Liu, K. Dai, and W. Qu, "Minimum control latency of SDN controller placement," in *Proc. IEEE Trustcom/BigDataSE/ISPA*, Aug. 2016, pp. 2175–2180.

[60] R. Guerzoni, "Network functions virtualisation: An introduction, benefits, enablers, challenges and call for action, introductory white paper," in *Proc. SDN OpenFlow World Congr.*, vol. 1, 2012, pp. 5–7.

[61] ETSI. (2014). *Network Functions Virtualisation*. [Online]. Available: http://www.etsi.org/technologies-clusters/technologies/nfv

[62] Huawei. (2020). *Cloudedge*. [Online]. Available: https://carrier.huawei.com/en/solutions/cloud-enabled-digital-operations/cloudedge

[63] A. Al-Dulaimi, X. Wang, and C. I, *Network Softwarization View of 5G Networks*. Piscataway, NJ, USA: Wiley, 2018, pp. 499–518. [Online]. Available: https://ieeexplore.ieee.org/document/8496391

[64] C. Mouradian, D. Naboulsi, S. Yangui, R. H. Glitho, M. J. Morrow, and P. A. Polakos, "A comprehensive survey on fog computing: State-of-the-art and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 416–464, 1st Quart., 2018.

[65] T. Jones, G. Fairhurst, and C. Perkins, "Raising the datagram API to support transport protocol evolution," in *Proc. IFIP Netw. Conf. (IFIP Networking) Workshops*, Jun. 2017, pp. 1–6.

[66] B. Trammell, C. Perkins, and M. Kuhlewind, "Post sockets: Towards an evolvable network transport interface," in *Proc. IFIP Netw. Conf. (IFIP Networking) Workshops*, Jun. 2017, pp. 1–6.

[67] E. Atxutegi, "Moving toward the intra-protocol de-ossification of TCP in mobile networks: Start-up and mobility," Ph.D. dissertation, Univ. Basque Country, Vizcaya, Spain, Jan. 2018.

[68] A. Bensky, "Wireless personal area networks," in *Short-Range Wireless Communication*, A. Bensky, Ed., 3rd ed. London, U.K.: Newnes, 2019, ch. 12, pp. 317–360. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128154052000129

[69] F. R. Kschischang, "An introduction to network coding," in *Network Coding*, M. Médard and A. Sprintson, Eds. Boston, MA, USA: Academic, 2012, ch. 1, pp. 1–37. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780123809186000019

[70] B. Bangerter, S. Talwar, R. Arefi, and K. Stewart, "Networks and devices for the 5G era," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 90–96, Feb. 2014.

[71] IETF. *Path Aware Networking RG (Panrg)*. Accessed: Jan. 23, 2020. [Online]. Available: https://datatracker.ietf.org/rg/panrg/about/

[72] *Cisco Intent-Based Networking (IBN)*. Accessed: Jul. 31, 2020. [Online]. Available: https://www.cisco.com/c/en/us/solutions/intent-based-networking.html?dtid=osscdc000283

[73] A. Galis and K. Makhijani, "Network slicing landscape: A holistic architectural approach, orchestration and management with applicability in mobile and fixed networks and clouds," in *Proc. IEEE Conf. Netw. Softw. (NetSoft)*, Jun. 2018.

[74] H. Koumaras, D. Tsolkas, G. Gardikis, P. M. Gomez, V. Frascolla, D. Triantafyllopoulou, M. Emmelmann, V. Koumaras, M. L. G. Osma, D. Munaretto, E. Atxutegi, J. S. D. Puga, O. Alay, A. Brunstrom, and A. M. C. Bosneag, "5GENESIS: The genesis of a flexible 5G facility," in *Proc. IEEE 23rd Int. Workshop Comput. Aided Modeling Design Commun. Links Netw. (CAMAD)*, Sep. 2018, pp. 1–6.

[75] 5GENESIS. (2018). *5GENESIS Objectives*. [Online]. Available: https://5genesis.eu/objectives/

[76] R. Marasli, P. D. Amer, and P. T. Conrad, "Retransmission-based partially reliable transport service: An analytic model," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Mar. 1996, pp. 621–629.

[77] R. Stewart, M. Ramalho, Q. Xie, M. Tuexen, and P. Conrad, *Stream Control Transmission Protocol (SCTP) Partial Reliability Extension*, document RFC 3758, RFC Editor, Internet Requests for Comments, May 2004. [Online]. Available: http://www.rfc-editor.org/rfc/rfc3758.txt. http://www.rfc-editor.org/rfc/rfc3758.txt

[78] A. Khandekar, N. Bhushan, J. Tingfang, and V. Vanghi, "LTE-advanced: Heterogeneous networks," in *Proc. Eur. Wireless Conf.*, Apr. 2010, pp. 978–982.

[79] A. Petlund, "Improving latency for interactive, thin-stream applications over reliable transport," *ACM SIGMultimedia Records*, vol. 2, no. 1, pp. 17–18, Mar. 2010.

[80] Z. Yin, H. Alnuweiri, A. L. N. Reddy, H. Celebi, and K. Qaraqe, "Improving the performance of delay based protocol in delivering real time media via early retransmission," in *Proc. 18th Int. Conf. Telecommun.*, May 2011, pp. 511–516.

[81] J.-W. Park, R. P. Karrer, and J. Kim, "TCP-Rome: A transport-layer parallel streaming protocol for real-time online multimedia environments," *J. Commun. Netw.*, vol. 13, no. 3, pp. 277–285, Jun. 2011.

[82] M. Massaro, C. E. Palazzi, and A. Bujari, "Exploiting TCP vegas' algorithm to improve real-time multimedia applications," in *Proc. 12th Annu. IEEE Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2015, pp. 316–321.

[83] M. Polese, R. Jana, and M. Zorzi, "TCP in 5G mmWave networks: Link level retransmissions and MP-TCP," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2017, pp. 343–348.

[84] I. Petrov and T. Janevski, "Advanced 5G-TCP: Transport protocol for 5G mobile networks," in *Proc. 14th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2017, pp. 103–107.

[85] Google LLC. (Jul. 2017). *TCP BBR Congestion Control Comes to GCP—Your Internet Just Got Faster, Google Cloud Platform Blog*. [Online]. Available: https://cloudplatform.googleblog.com/2017/07/TCP-BBR-congestion-control-comes-to-GCP-your-Internet-just-got-faster.html

[86] B. Gambhava and C. Bhensdadia, "Discrete TCP: Differentiating slow start and congestion avoidance," *Int. J. Intell. Eng. Syst.*, vol. 11, no. 5, pp. 206–214, Oct. 2018.

[87] X. Zhu, R. Zheng, D. Yang, H. Liu, and J. Hou, "Radio-aware TCP optimization in mobile network," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–5.

[88] J. Luo, J. Jin, and F. Shan, "Standardization of low-latency TCP with explicit congestion notification: A survey," *IEEE Internet Comput.*, vol. 21, no. 1, pp. 48–55, Jan. 2017.

[89] W. Zhou, Q. Li, M. Caesar, and P. B. Godfrey, "ASAP: A low-latency transport layer," in *Proc. 7th Conf. Emerg. Netw. Exp. Technol. (CoNEXT)*, New York, NY, USA, 2011, pp. 20:1–20:12.

[90] C. Park and H. Kim, "Short-term reliable protocol for low latency video transmission," in *Proc. Int. Conf. Comput. Sci. Comput. Intell.*, Mar. 2014, pp. 311–312.

[91] H.-H. Cheng and K. C.-J. Lin, "Source selection and content dissemination for preference-aware traffic offloading," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 11, pp. 3160–3174, Nov. 2015.

[92] C. Park and H. Kim, "Low latency video transmission device," in *Information Science and Applications*. Berlin, Germany: Springer, 2015, pp. 217–222.

[93] I. Johansson and Z. Sarker, *Self-Clocked Rate Adaptation for Multimedia*, document RFC 8298, RFC Editor, Internet Requests for Comments, Dec. 2017.

[94] R. Mittal, V. T. Lam, N. Dukkipati, E. Blem, H. Wassel, M. Ghobadi, A. Vahdat, Y. Wang, D. Wetherall, and D. Zats, "TIMELY: RTT-based congestion control for the datacenter," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 4, pp. 537–550, 2015.

[95] B. Briscoe, K. Schepper, M. Bagnulo, and G. White. (Mar. 2020). *Low Latency, Low Loss, Scalable Throughput (L4S) Internet Service: Architecture*. [Online]. Available: https://tools.ietf.org/html/draft-ietf-tsvwg-l4s-arch-06. https://tools.ietf.org/html/draft-ietf-tsvwg-l4s-arch-06

[96] M. Menth, A. Stockmayer, and M. Schmidt. (Jul. 2015). *Lisp Hybrid Access*. [Online]. Available: https://www.ietf.org/archive/id/draft-menth-lisp-ha-00.txt

[97] K.-K. Yap, T.-Y. Huang, Y. Yiakoumis, S. Chinchali, N. McKeown, and S. Katti, "Scheduling packets over multiple interfaces while respecting user preferences," in *Proc. 9th ACM Conf. Emerg. Netw. Exp. Technol. (CoNEXT)*, New York, NY, USA, 2013, pp. 109–120.

[98] S. Singh, S.-P. Yeh, N. Himayat, and S. Talwar, "Optimal traffic aggregation in multi-RAT heterogeneous wireless networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, May 2016, pp. 626–631.

[99] I. G. Muriel, A. M. Heredia, and P. M. Gomez, "Testbed to experiment with LTE WiFi aggregation," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2019, pp. 506–511.

[100] A. Ford, C. Raiciu, M. Handley, and O. Bonaventure, *TCP Extensions for Multipath Operation With Multiple Addresses*, document RFC 6824, RFC Editor, Internet Requests for Comments, Jan. 2013. [Online]. Available: http://www.rfc-editor.org/rfc/rfc6824.txt

[101] M. Li, A. Lukyanenko, and Y. Cui, "Network coding based multipath TCP," in *Proc. IEEE INFOCOM Workshops*, Mar. 2012, pp. 25–30.

[102] Y. Cui, L. Wang, X. Wang, H. Wang, and Y. Wang, "FMTCP: A fountain code-based multipath transmission control protocol," *IEEE/ACM Trans. Netw.*, vol. 23, no. 2, pp. 465–478, Apr. 2015.

[103] P. Hurtig, K.-J. Grinnemo, A. Brunstrom, S. Ferlin, O. Alay, and N. Kuhn, "Low-latency scheduling in MPTCP," *IEEE/ACM Trans. Netw.*, vol. 27, no. 1, pp. 302–315, Feb. 2019.

[104] C. Diop, G. Dugue, C. Chassot, and E. Exposito, "QoS-oriented MPTCP extensions for multimedia multi-homed systems," in *Proc. 26th Int. Conf. Adv. Inf. Netw. Appl. Workshops*, Mar. 2012, pp. 1119–1124.

[105] J. Wu, C. Yuen, B. Cheng, M. Wang, and J. Chen, "Streaming high-quality mobile video with multipath TCP in heterogeneous wireless networks," *IEEE Trans. Mobile Comput.*, vol. 15, no. 9, pp. 2345–2361, Sep. 2016.

[106] Y. Cao, Q. Liu, G. Luo, Y. Yi, and M. Huang, "PR-MPTCP+: Context-aware QoE-oriented multipath TCP partial reliability extension for real-time multimedia applications," in *Proc. Vis. Commun. Image Process. (VCIP)*, Nov. 2016, pp. 1–4.

[107] A. Nikravesh, Y. Guo, F. Qian, Z. M. Mao, and S. Sen, "An in-depth understanding of multipath TCP on mobile devices: Measurement and system design," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, New York, NY, USA, 2016, pp. 189–201.

[108] C. Lee, S. Song, H. Cho, G. Lim, and J.-M. Chung, "Optimal multipath TCP offloading over 5G NR and LTE networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 293–296, Feb. 2019.

[109] A. Singh and A. L. N. Reddy, "Multi path PERT," in *Proc. 22nd Int. Conf. Comput. Commun. Netw. (ICCCN)*, Jul. 2013, pp. 1–9.

[110] Q. D. Coninck, "First experiments with multipath quic," in *Proc. IETF 99*, Jul. 2017. [Online]. Available: https://datatracker.ietf.org/meeting/99/materials/slides-99-quic-sessb-first-experiments-with-multipath-quic/

[111] V. Singh, T. Karkkainen, J. Ott, S. Ahsan, and L. Eggert. (Jul. 2016). Multipath RTP (MPRTP), Working Draft, IETF Secretariat. [Online]. Available: https://tools.ietf.org/html/draft-ietf-avtcore-mprtp-03

[112] O. C. Kwon, "Multipath transport protocols for video streaming over heterogeneous wireless networks," Ph.D. dissertation, Dept. Comput. Sci. Eng., Pohang Univ. Sci. Technol., Pohang-si, South Korea, Dec. 2014.

[113] O. C. Kwon, Y. Go, Y. Park, and H. Song, "MPMTP: Multipath multimedia transport protocol using systematic raptor codes over wireless networks," *IEEE Trans. Mobile Comput.*, vol. 14, no. 9, pp. 1903–1916, Sep. 2015.

[114] Y. Hwang, B. O. Obele, and H. Lim, "Multipath transport protocol for heterogeneous multi-homing networks," in *Proc. ACM CoNEXT Student Workshop (CoNEXT)*, New York, NY, USA, 2010, pp. 5:1–5:2.

[115] W. Lei, S. Liu, and W. Zhang. (Feb. 2018). Multipath message transport protocol based on application-level relay (MPMTP-AR). Working Draft, IETF Secretariat. [Online]. Available: https://tools.ietf.org/html/draft-leiwm-tsvwg-mpmtp-ar-09

[116] K. R. Kashwan and S. Karthik, "The modified mobile concurrent multipath transfer for joint resource management," *Procedia Eng.*, vol. 30, pp. 963–969, Jan. 2012.

[117] L. P. Verma and M. Kumar, "An adaptive data chunk scheduling for concurrent multipath transfer," *Comput. Standards Interfaces*, vol. 52, pp. 97–104, May 2017.

[118] T. Zhu, D. Feng, F. Wang, Y. Hua, Q. Shi, Y. Xie, and Y. Wan, "A congestion-aware and robust multicast protocol in SDN-based data center networks," *J. Netw. Comput. Appl.*, vol. 95, pp. 105–117, Oct. 2017.

[119] E. Tsimbalo, A. Tassi, and R. J. Piechocki, "Reliability of multicast under random linear network coding," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2547–2559, Jun. 2018.

[120] X. Xiong and T. Chen, "MTM: A reliable multiple trees multicast for data center network," in *Proc. Int. Conf. Netw., Archit., Storage (NAS)*, Aug. 2017, pp. 1–7.

[121] K. Chi, L. Huang, Y. Li, Y.-H. Zhu, X.-Z. Tian, and M. Xia, "Efficient and reliable multicast using device-to-device communication and network coding for a 5G network," *IEEE Netw.*, vol. 31, no. 4, pp. 78–84, Jul. 2017.

[122] S. Roger, D. Martín-Sacristán, D. Garcia-Roger, J. F. Monserrat, P. Spapis, A. Kousaridas, S. Ayaz, and A. Kaloxylos, "Low-latency layer-2-based multicast scheme for localized V2X communications," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 8, pp. 2962–2975, Aug. 2018.

[123] X. Zhang, M. Yang, L. Wang, and M. Sun, "An OpenFlow-enabled elastic loss recovery solution for reliable multicast," *IEEE Syst. J.*, vol. 12, no. 2, pp. 1945–1956, Jun. 2018.

[124] K. Mahajan, D. Sharma, and V. Mann, "Athena: Reliable multicast for group communication in SDN-based data centers," in *Proc. 9th Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Jan. 2017, pp. 174–181.

[125] J. Postel, *Transmission Control Protocol*, document RFC 0793, RFC Editor, Internet Requests for Comments, Sep. 1981. [Online]. Available: http://www.rfc-editor.org/rfc/rfc793.txt

[126] P. B. Beskow, A. Petlund, G. A. Erikstad, C. Griwodz, and P. Halvorsen, "Reducing game latency by migration, core-selection and TCP modifications," *Int. J. Adv. Media Commun.*, vol. 4, no. 4, pp. 343–363, 2010.

[127] A. Bujari, M. Massaro, and C. E. Palazzi, "Vegas over access point: Making room for thin client game systems in a wireless home," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 2002–2012, Dec. 2015.

[128] L. S. Brakmo, S. W. O'Malley, and L. L. Peterson, "TCP vegas: New techniques for congestion detection and avoidance," in *Proc. Conf. Commun. Archit., Protocols Appl.*, 1994, vol. 24, no. 4, pp. 24–35.

[129] S. Bensley, D. Thaler, P. Balasubramanian, L. Eggert, and G. Judd, "Data center TCP (DCTCP): TCP congestion control for data centers," document RFC 8257, RFC Editor, Internet Requests for Comments, Oct. 2017.

[130] D. B. Oljira, K.-J. Grinnemo, A. Brunstrom, and J. Taheri, "MDTCP: Towards a practical multipath transport protocol for telco cloud datacenters," in *Proc. 9th Int. Conf. Netw. Future (NOF)*, Nov. 2018, pp. 9–16.

[131] P. Bellavista and C. Giannelli, "Internet connectivity sharing in multipath spontaneous networks: Comparing and integrating network- and application-layer approaches," in *Proc. Int. Conf. Mobile Wireless Middleware, Oper. Syst., Appl.* Berlin, Germany: Springer, 2010, pp. 84–99.

[132] Y.-C. Chen, Y.-S. Lim, R. J. Gibbens, E. M. Nahum, R. Khalili, and D. Towsley, "A measurement-based study of multipath tcp performance over wireless networks," in *Proc. Conf. Internet Meas. Conf. (IMC)*, New York, NY, USA, 2013, pp. 455–468. [Online]. Available: http://doi.acm.org/10.1145/2504730.2504751

[133] C. Paasch, G. Detal, F. Duchene, C. Raiciu, and O. Bonaventure, "Exploring mobile/WiFi handover with multipath TCP," in *Proc. ACM SIGCOMM Workshop Cellular Netw., Oper., Challenges, Future Design (CellNet)*, New York, NY, USA, 2012, pp. 31–36.

[134] S. C. Nguyen and T. M. T. Nguyen, "Evaluation of multipath TCP load sharing with coupled congestion control option in heterogeneous networks," in *Proc. Global Inf. Infrastruct. Symp. (GIIS)*, Aug. 2011, pp. 1–5.

[135] S. C. Nguyen, X. Zhang, T. M. T. Nguyen, and G. Pujolle, "Evaluation of throughput optimization and load sharing of multipath TCP in heterogeneous networks," in *Proc. 8th Int. Conf. Wireless Opt. Commun. Netw.*, May 2011, pp. 1–5.

[136] *Technical Specification Group Services and System Aspects; System Architecture for the 5G System (5GS);Stage 2 (Release 16)*, document TS 23.501, Version 16.3.0, 3GPP, Dec. 2019. [Online]. Available: http://www.3gpp.org/ftp//Specs/archive/23_series/23.501/23501-g30.zip

[137] Tessares. (2019). *Tessares' 5G ATSSS Solution*. [Online]. Available: https://www.tessares.net/solutions/5g-atsss-solution/

[138] O. Dharmadhikari. (Apr. 2019). 5G link aggregation with multipath TCP (MPTCP). CableLabs. [Online]. Available: https://www.cablelabs.com/the-10g-converged-optical-network

[139] Q. De Coninck and O. Bonaventure, "Multipath quic: Design and evaluation," in *Proc. 13th Int. Conf. Emerg. Netw. Exp. Technol. (CoNEXT)*, New York, NY, USA, 2017, pp. 160–166.

[140] L. Weimin, W. Zhang, and S. Liu. (Jul. 2013). *A Framework of Multipath Transport System Based on Application-Level Relay (MPTS-AR)*. [Online]. Available: https://tools.ietf.org/html/draft-leiwm-tsvwg-mpts-ar-00

[141] J. Iyengar, K. Shah, P. Amer, and R. Stewart, "Concurrent multipath transfer using SCTP multihoming," in *Proc. SPECTS*, 2004, pp. 951–964.

[142] C. A. García-Pérez and P. Merino, "Enabling low latency services on LTE networks," in *Proc. IEEE 1st Int. Workshops Found. Appl. Self Syst. (FASW)*, Sep. 2016, pp. 248–255.

[143] C. A. García-Pérez and P. Merino, "Experimental evaluation of fog computing techniques to reduce latency in LTE networks," *Trans. Emerg. Telecommun. Technol.*, vol. 29, no. 4, p. e3201, Apr. 2018.

[144] K. Zhang, S. Leng, Y. He, S. Maharjan, and Y. Zhang, "Mobile edge computing and networking for green and low-latency Internet of Things," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 39–45, May 2018.

[145] M. J. Piran, S. M. R. Islam, and D. Y. Suh, "CASH: Content- and network-context-aware streaming over 5G HetNets," *IEEE Access*, vol. 6, pp. 46167–46178, 2018.

[146] S.-Q. Lee and J.-U. Kim, "Local breakout of mobile access network traffic by mobile edge computing," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2016, pp. 741–743.

[147] G. Cattaneo, F. Giust, C. Meani, D. Munaretto, and P. Paglierani, "Deploying CPU-intensive applications on MEC in NFV systems: The immersive video use case," *Computers*, vol. 7, no. 4, p. 55, Oct. 2018.

[148] A. Huang, N. Nikaein, T. Stenbock, A. Ksentini, and C. Bonnet, "Low latency MEC framework for SDN-based LTE/LTE–A networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[149] J. Heinonen, P. Korja, T. Partti, H. Flinck, and P. Pöyhönen, "Mobility management enhancements for 5G low latency services," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, May 2016, pp. 68–73.

[150] E. Schiller, N. Nikaein, E. Kalogeiton, M. Gasparyan, and T. Braun, "CDS-MEC: NFV/SDN-based application management for MEC in 5G systems," *Comput. Netw.*, vol. 135, pp. 96–107, Apr. 2018. [Online]. Available: http://www.eurecom.fr/publication/5461

[151] B. Yang, W. K. Chai, Z. Xu, K. V. Katsaros, and G. Pavlou, "Cost-efficient NFV-enabled mobile edge-cloud for low latency mobile applications," *IEEE Trans. Netw. Service Manage.*, vol. 15, no. 1, pp. 475–488, Mar. 2018.

[152] R. Cziva and D. P. Pezaros, "On the latency benefits of edge NFV," in *Proc. ACM/IEEE Symp. Archit. Netw. Commun. Syst. (ANCS)*, May 2017, pp. 105–106.

[153] S. Nunna, A. Kousaridas, M. Ibrahim, M. Dillinger, C. Thuemmler, H. Feussner, and A. Schneider, "Enabling real-time context-aware collaboration through 5G and mobile edge computing," in *Proc. 12th Int. Conf. Inf. Technol. New Generat.*, Apr. 2015, pp. 601–605.

[154] S. Dutta, T. Taleb, P. A. Frangoudis, and A. Ksentini, "On-the-fly QoE-aware transcoding in the mobile edge," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.

[155] T. Taleb, S. Dutta, A. Ksentini, M. Iqbal, and H. Flinck, "Mobile edge computing potential in making cities smarter," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 38–43, Mar. 2017.

[156] E. Li, Z. Zhou, and X. Chen, "Edge intelligence: On-demand deep learning model co-inference with device-edge synergy," in *Proc. Workshop Mobile Edge Commun. (MECOMM)*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 31–36, doi: 10.1145/3229556.3229562.

[157] M. Maier and A. Ebrahimzadeh, "Towards immersive tactile Internet experiences: Low-latency FiWi enhanced mobile networks with edge intelligence [invited]," *J. Opt. Commun. Netw.*, vol. 11, no. 4, p. B10, Apr. 2019.

[158] J. Liu and Q. Zhang, "Offloading schemes in mobile edge computing for ultra-reliable low latency communications," *IEEE Access*, vol. 6, pp. 12825–12837, 2018.

[159] F. Giust *et al.*, "MEC deployments in 4G and evolution towards 5G," Eur. Telecommun. Standards Inst. (ETSI), Sophia Antipolis, France, ETSI White Paper 24, Feb. 2018. [Online]. Available: https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp24_MEC_deployment_in_4G_5G_FINAL.pdf

[160] J. Page and J.-M. Dricot, "Software-defined networking for low-latency 5G core network," in *Proc. Int. Conf. Mil. Commun. Inf. Syst. (ICMCIS)*, May 2016, pp. 1–7.

[161] J. Costa-Requena, A. Poutanen, S. Vural, G. Kamel, C. Clark, and S. K. Roy, "SDN-based UPF for mobile backhaul network slicing," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2018, pp. 48–53.

[162] J. Wang and D. Li, "Adaptive computing optimization in software-defined network-based industrial Internet of Things with fog computing," *Sensors*, vol. 18, no. 8, p. 2509, Aug. 2018.

[163] E. Lakiotakis, C. Liaskos, and X. Dimitropoulos, "Application-network collaboration using SDN for ultra-low delay teleorchestras," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jul. 2017, pp. 70–75.

[164] S. Garg, K. Kaur, S. H. Ahmed, A. Bradai, G. Kaddoum, and M. Atiquzzaman, "MobQoS: Mobility-aware and QoS-driven SDN framework for autonomous vehicles," *IEEE Wireless Commun.*, vol. 26, no. 4, pp. 12–20, Aug. 2019.

[165] G. Wang, Y. Zhao, J. Huang, and Y. Wu, "An effective approach to controller placement in software defined wide area networks," *IEEE Trans. Netw. Service Manage.*, vol. 15, no. 1, pp. 344–355, Mar. 2018.

[166] K.-K. Yap, T.-Y. Huang, M. Kobayashi, Y. Yiakoumis, N. McKeown, S. Katti, and G. Parulkar, "Making use of all the networks around us: A case study in Android," in *Proc. ACM SIGCOMM Workshop Cellular Netw., Oper., Challenges, Future Design (CellNet)*, New York, NY, USA, 2012, pp. 19–24.

[167] T. Hu, P. Yi, J. Zhang, and J. Lan, "Reliable and load balance-aware multi-controller deployment in SDN," *China Commun.*, vol. 15, no. 11, pp. 184–198, Nov. 2018.

[168] M. T. Raza and S. Lu, "Enabling low latency and high reliability for IMS-NFV," in *Proc. 13th Int. Conf. Netw. Service Manage. (CNSM)*, Nov. 2017, pp. 1–9.

[169] M. T. Raza, S. Lu, M. Gerla, and X. Li, "Refactoring network functions modules to reduce latencies and improve fault tolerance in NFV," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 10, pp. 2275–2287, Oct. 2018.

[170] L. Qu, C. Assi, K. Shaban, and M. J. Khabbaz, "A reliability-aware network service chain provisioning with delay guarantees in NFV-enabled enterprise datacenter networks," *IEEE Trans. Netw. Service Manage.*, vol. 14, no. 3, pp. 554–568, Sep. 2017.

[171] P.-V. Mekikis, K. Ramantas, A. Antonopoulos, E. Kartsakli, L. Sanabria-Russo, J. Serra, D. Pubill, and C. Verikoukis, "NFV-enabled experimental platform for 5G tactile Internet support in industrial environments," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 1895–1903, Mar. 2020.

[172] W. Ding, H. Yu, and S. Luo, "Enhancing the reliability of services in NFV with the cost-efficient redundancy scheme," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[173] M. Nascimento, T. Primini, E. Baum, P. Martucci, F. Cabelo, and L. Mariote, "Acceleration mechanism for high throughput and low latency in NFV environments," in *Proc. IEEE 18th Int. Conf. High Perform. Switching Routing (HPSR)*, Jun. 2017, pp. 1–6.

[174] D. Cho, J. Taheri, A. Y. Zomaya, and P. Bouvry, "Real-time virtual network function (VNF) migration toward low network latency in cloud environments," in *Proc. IEEE 10th Int. Conf. Cloud Comput. (CLOUD)*, Jun. 2017, pp. 798–801.

[175] C. Sun, J. Bi, Z. Zheng, H. Yu, and H. Hu, "NFP: Enabling network function parallelism in NFV," in *Proc. Conf. ACM Special Interest Group Data Commun. (SIGCOMM)*. New York, NY, USA: Association for Computing Machinery, 2017, pp. 43–56, doi: 10.1145/3098822.3098826.

[176] J. Fan, Z. Ye, C. Guan, X. Gao, K. Ren, and C. Qiao, "GREP: Guaranteeing reliability with enhanced protection in NFV," in *Proc. ACM SIGCOMM Workshop Hot Topics Middleboxes Netw. Function Virtualization (HotMiddlebox)*. New York, NY, USA: Association for Computing Machinery, 2015, pp. 13–18, doi: 10.1145/2785989.2786000.

[177] O. Bekkouche, M. Bagaa, and T. Taleb, "Toward a UTM-based service orchestration for UAVs in MEC-NFV environment," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[178] P. Valsamas, P. Papadimitriou, I. Sakellariou, S. Petridou, L. Mamatas, S. Clayman, F. Tusa, and A. Galis, "Multi-PoP network slice deployment: A feasibility study," in *Proc. IEEE 8th Int. Conf. Cloud Netw. (CloudNet)*, Nov. 2019, pp. 1–6.

[179] Y. Yao, Q. Cao, J. Chase, P. Ruth, I. Baldin, Y. Xin, and A. Mandal, "Slice-based network transit service: Inter-domain L2 networking on ExoGENI," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2017, pp. 736–741.

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

[181] G. Carofiglio, L. Mekinda, and L. Muscariello, "LAC: Introducing latency-aware caching in information-centric networks," in *Proc. IEEE 40th Conf. Local Comput. Netw. (LCN)*, Oct. 2015, pp. 422–425.

[182] G. Carofiglio, L. Mekinda, and L. Muscariello, "FOCAL: Forwarding and caching with latency awareness in information-centric networking," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1–7.

[183] G. Carofiglio, L. Mekinda, and L. Muscariello, "Joint forwarding and caching with latency awareness in information-centric networking," *Comput. Netw.*, vol. 110, pp. 133–153, Dec. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128616303176

[184] Z. Zhang, C.-H. Lung, I. Lambadaris, and M. St-Hilaire, "When 5G meets ICN: An ICN-based caching approach for mobile video in 5G networks," *Comput. Commun.*, vol. 118, pp. 81–92, Mar. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140366417305066

[185] M. Sardara, L. Muscariello, and A. Compagno, "A transport layer and socket API for (h)ICN: Design, implementation and performance analysis," in *Proc. 5th ACM Conf. Inf.-Centric Netw. (ACM ICN)*, Sep. 2018, pp. 137–147.

[186] C. Dannewitz, D. Kutscher, B. Ohlman, S. Farrell, B. Ahlgren, and H. Karl, "Network of information (NetInf)—An information-centric networking architecture," *Comput. Commun.*, vol. 36, no. 7, pp. 721–735, Apr. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140366413000364

[187] Z. Wang, H. Luo, H. Zhou, and J. Li, "R2T: A rapid and reliable hop-by-hop transport mechanism for information-centric networking," *IEEE Access*, vol. 6, pp. 15311–15325, 2018.

[188] S. Vakilinia and H. Elbiaze, "Latency control of ICN enabled 5G networks," *J. Netw. Syst. Manage.*, vol. 28, no. 1, pp. 81–107, Jan. 2020.

[189] M. Scharf and A. Ford, *Multipath TCP (MPTCP) Application Interface Considerations*, document RFC 6897, Internet Requests for Comments, RFC Editor, Mar. 2013.

[190] B. Hesmans, G. Detal, S. Barre, R. Bauduin, and O. Bonaventure, "SMAPP: Towards smart multipath TCP-enabled applications," in *Proc. 11th ACM Conf. Emerg. Netw. Exp. Technol. (CoNEXT)*, New York, NY, USA, 2015, pp. 28:1–28:7.

[191] B. Hesmans and O. Bonaventure, "An enhanced socket API for multipath TCP," in *Proc. Appl. Netw. Res. Workshop (ANRW)*, New York, NY, USA, 2016, pp. 1–6.

[192] K.-J. Grinnemo, T. Jones, G. Fairhurst, D. Ros, A. Brunstrom, and P. Hurtig, "Towards a flexible Internet transport layer architecture," in *Proc. IEEE Int. Symp. Local Metrop. Area Netw. (LANMAN)*, Jun. 2016, pp. 1–7.

[193] N. Khademi, D. Ros, M. Welzl, Z. Bozakov, A. Brunstrom, G. Fairhurst, K.-J. Grinnemo, D. Hayes, P. Hurtig, T. Jones, S. Mangiante, M. Tuxen, and F. Weinrank, "NEAT: A platform- and protocol-independent Internet transport API," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 46–54, Jun. 2017.

[194] T. Pauly, B. Trammell, A. Brunstrom, G. Fairhurst, C. Perkins, P. Tiesel, and C. Wood. (Jul. 2020). An architecture for transport services, working draft, IETF Secretariat. [Online]. Available: https://datatracker.ietf.org/doc/draft-ietf-taps-arch/. https://datatracker.ietf.org/doc/draft-ietf-taps-arch/

[195] J. J. Nielsen, R. Liu, and P. Popovski, "Ultra-reliable low latency communication using interface diversity," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1322–1334, Mar. 2018.

[196] B. D. Higgins, A. Reda, T. Alperovich, J. Flinn, T. J. Giuli, B. Noble, and D. Watson, "Intentional networking: Opportunistic exploitation of mobile network diversity," in *Proc. 16th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, 2010, pp. 73–84.

[197] R. Davoli and M. Goldweber, "Msocket: Multiple stack support for the Berkeley socket API," in *Proc. 27th Annu. ACM Symp. Appl. Comput. (SAC)*, 2012, pp. 588–593.

[198] P. S. Schmidt, T. Enghardt, R. Khalili, and A. Feldmann, "Socket intents: Leveraging application awareness for multi-access connectivity," in *Proc. 9th ACM Conf. Emerg. Netw. Exp. Technol. (CoNEXT)*, New York, NY, USA, 2013, pp. 295–300.

[199] R. Kapoor, G. Porter, M. Tewari, G. M. Voelker, and A. Vahdat, "Chronos: Predictable low latency for data center applications," in *Proc. 3rd ACM Symp. Cloud Comput.*, 2012, p. 9.

[200] A. Belay, G. Prekas, A. Klimovic, S. Grossman, C. Kozyrakis, and E. Bugnion, "IX: A protected dataplane operating system for high throughput and low latency," in *Proc. 11th USENIX Symp. Oper. Syst. Design Implement. (OSDI)*. Broomfield, CO, USA: USENIX Association, 2014, pp. 49–65.

[201] A. Belay, G. Prekas, M. Primorac, A. Klimovic, S. Grossman, C. Kozyrakis, and E. Bugnion, "Corrigendum to 'the ix operating system: Combining low latency, high throughput and efficiency in a protected dataplane,'" *ACM Trans. Comput. Syst.*, vol. 35, no. 3, pp. 10:1–10:1, Dec. 2017.

[202] A. A. Siddiqui and P. Mueller, "A requirement-based socket API for a transition to future Internet architectures," in *Proc. 6th Int. Conf. Innov. Mobile Internet Services Ubiquitous Comput.*, Jul. 2012, pp. 340–345.

[203] S. Bocking, "Sockets++: A uniform application programming interface for basic level communication services," *IEEE Commun. Mag.*, vol. 34, no. 12, pp. 114–123, Dec. 1996.

[204] P. G. S. Florissi, Y. Yemini, and D. Florissi, "QoSockets: A new extension to the sockets API for end-to-end application QoS management," in *Proc. Integr. Netw. Manage. Distrib. Manage. Netw. Millennium, 6th IFIP/IEEE Int. Symp. Integr. Netw. Manage.*, May 1999, pp. 655–668.

[205] H. Abbasi, C. Poellabauer, K. Schwan, G. Losik, and R. West, "A quality-of-service enhanced socket API in GNU/Linux," in *Proc. 4th Real-Time Linux Workshop*, Sep. 2004, p. 31.

[206] B. Reuther, D. Henrici, and M. Hillenbrand, "DANCE: Dynamic application oriented network services," in *Proc. 30th Euromicro Conf.*, 2004, pp. 298–305.

[207] D. A. Chekired, M. A. Togou, L. Khoukhi, and A. Ksentini, "5G-slicing-enabled scalable SDN core network: Toward an ultra-low latency of autonomous driving service," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 8, pp. 1769–1782, Aug. 2019.

[208] M. Emmelmann *et al.*, "Deliverable D6.1: Trials and experimentation (cycle 1), version 2.0," 5Genesis, 5G-PPP, Tech. Rep., 2019. [Online]. Available: https://5genesis.eu/wp-content/uploads/2019/12/5GENESIS_D6.1_v2.00.pdf

[209] E. Balevi and R. D. Gitlin, "Unsupervised machine learning in 5G networks for low latency communications," in *Proc. IEEE 36th Int. Perform. Comput. Commun. Conf. (IPCCC)*. Los Alamitos, CA, USA: IEEE Computer Society, Dec. 2017, pp. 1–2. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/PCCC.2017.8280492

[210] M. E. Morocho-Cayamcela, H. Lee, and W. Lim, "Machine learning for 5G/B5G mobile and wireless communications: Potential, limitations, and future directions," *IEEE Access*, vol. 7, pp. 137184–137206, 2019.

DELIA RICO received the B.Sc. degree in telematics engineering and the M.Sc. degree in telematics engineering and telecommunication networks from the University of Malaga, Spain, in 2017 and 2018, respectively, where she is currently pursuing the Ph.D. degree. Her current research interests include cellular networks, protocols, and middleware solutions.

PEDRO MERINO is currently a Professor with the University of Malaga (UMA). His research interests include new generation Internet, 5G networks, and automated methods for software reliability. He has led more than 30 national and international research projects, most of them in collaboration with industry. He leads a 4G/5G outdoor testbed at Malaga city. He was the Chair of the ERCIM WG on Formal Methods for Industrial Critical Systems, and a member of the Executive Committee of ERCIM. He represents the University of Malaga in Networld2020 ETP and 5G Industrial Association. He is also the Coordinator of EuWireless project, a Technical Manager of 5GENESIS project, and the Director of the ITIS Software Research Institute, University of Malaga.