IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 4, 2020, accepted September 17, 2020, date of publication September 24, 2020,
date of current version October 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3026540

Challenges on the Way of Implementing TCP Over 5G Networks

REZA POORZARE AND ANNA CALVERAS AUGÉ
Department of Network Engineering, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain
Corresponding author: Reza Poorzare (reza.poorzare@upc.edu)

This work was supported by the Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya under Grant 2017 SGR 376.

ABSTRACT 5G cellular communication, especially with its hugely available bandwidth provided by millimeter-wave, is a promising technology to fulfill the coming high demand for vast data rates. These networks can support new use cases such as Vehicle to Vehicle and augmented reality due to its novel features such as network slicing along with the mmWave multi-gigabit-per-second data rate. Nevertheless, 5G cellular networks suffer from some shortcomings, especially in high frequencies because of the intermittent nature of channels when the frequency rises. Non-line of sight state, is one of the significant issues that the new generation encounters. This drawback is because of the intense susceptibility of higher frequencies to blockage caused by obstacles and misalignment. This unique characteristic can impair the performance of the reliable transport layer widely deployed protocol, TCP, in attaining high throughput and low latency throughout a fair network. As a result, the protocol needs to adjust the congestion window size based on the current situation of the network. However, TCP is not able to adjust its congestion window efficiently, and it leads to throughput degradation of the protocol. This paper presents a comprehensive analysis of reliable end-to-end communications in 5G networks. It provides the analysis of the effects of TCP in 5G mmWave networks, the discussion of TCP mechanisms and parameters involved in the performance over 5G networks, and a survey of current challenges, solutions, and proposals. Finally, a feasibility analysis proposal of machine learning-based approaches to improve reliable end-to-end communications in 5G networks is presented.

INDEX TERMS 5G, end-to-end reliability, mmWave, TCP.

I. INTRODUCTION

Due to the rise in demand for higher data rates by appearing new features and services, the necessity for increasing the bandwidth in new generation mobile networks is inevitable. As an indicator, it can be said that a 56 percent increase in mobile traffic occurred only from the first quarter of 2019 to the first quarter of 2020, and it is expected to have 14 percent quarter-on-quarter growth in 2020, and 31 percent annual increase from 2019 to 2025. It is intriguing to mention that the global monthly mobile data traffic usage could reach 33 EB (ExaByte) in 2019 and is expected to attain 164 EB in 2025, in which 45 percent of it will be carried by 5G (Fifth Generation) networks. The largest monthly average mobile traffic will be for North America by reaching to 45 GB per month per smartphone [1]. Generally, the motivation behind this high growing demand can be categorized into three groups, enhanced device capabilities, cheaper data plans which lead to affordable services, and an increment in data-incentive content.

By transition from 4G to 5G, the transmission rate increases around 1000 times, and 5G is expected to handle around 30 percent of 8.9 billion mobile communication devices in 2025, which will include 7.5 billion smartphones [1]. This number is four percent lower than the previous prediction. The reason is the spread of the COVID-19 virus started in late 2019, which affected the speed of mobile telecommunication coverage progress and delayed some spectrum auctions. However, by the end of May 2020, more than 75 service providers could deliver 5G services. By expansion of 5G networks, 65 percent of the world population will go under the coverage in 2025, which was around 5 percent in late 2019. It is interesting to say that Switzerland had a significant share of this coverage by providing 5G

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

networks in more than 90 percent of the country at the end of 2019 [1].

The new generation enables three primary use cases [2], [3], eMBB (Enhanced Mobile Broadband), which provides high data rates, mMTC (massive Machine Type Communication) that supports up to 10^6 devices per square kilometer, and URLLC (Ultra-Reliable Low-Latency Communication) which aims to provide 1 ms latency for latency-critical communications such as V2X (Vehicle to Everything) [4]. Moreover, other applications requiring wireless access networks with low latency and high bandwidth, such as disastrous or remote healthcare ones, are also being a stakeholder for coming 5G technology. These three use cases' final goal is to come up with flexibility for networks and to connect everything, everywhere, anytime [5].

One of the significant upsides of using a mobile system is the eMBB feature, which provides connectivity and higher bandwidth for users and can cover a range of services such as hotspot and wide-area coverage. In the first one, a high data rate, large user density, and high capacity are the essential characteristics. While, in the second one, being connected in a seamless way and mobility are essential. The features for eMBB makes it to be categorized as human-centric communication [4].

URLLC aims to provide reliable communications with latencies close to zero. By the emergence of technologies such as autonomous driving, the necessity for reliable and low-latency services has become crucial. As a result, URLLC came into reality to fulfill the requirements. It has an essential role in covering both human-centric and machine-centric communications. In the latter one, latency, reliability, and high availability are critical in establishing connections, primarily in latency-critical communications such as V2Vs (Vehicle to Vehicles), which is categorized under machine-centric communication. For the human-centric ones, low-latency and higher data rates can be needed simultaneously in cases such as 3D gaming [2]. Nowadays, implemented 5G networks can provide 20-30 ms latency, which will be reduced to sub-10 ms in the near future [1].

When there are a lot of machine-centric devices with the need for transmitting a small amount of data, mMTC can be beneficial. Having a battery life up to ten years, a large number of devices, a low transmission rate, and not being delay-sensitive are the principal characteristics of this use case. When IoT (Internet of Thing) solutions based on NB-IoT (NarrowBand-IoT) [6] are deployed in places such as underground or inside other devices such as cars or dynamic traffic lights, being able to penetrate materials is critical. These features can be provided by the mMTC use case [4], [7].

5G features will allow having new and robust capabilities compared to past generations. A higher peak data rate of up to 20 Gbps for DL (DownLink) and 10 Gbps for UP (UpLink) are excellent advancements that emerged with 5G accompanying. These numbers are theoretical data rates and can be achieved in ideal conditions. However, the user-experienced data rate, which is one of the critical KPIs (Key Performance Indicators) in 5G, is 100 Mbps for DL and 50 Mbps for UL. The main difference between the user-experienced data rate and peak data rate is that the former one can be achieved in real-time for the majority of the UEs (User Equipment).

For attaining a higher data rate, high spectral efficiency is needed. Spectral efficiency refers to the achievable data rate over a specific bandwidth, and for 5G networks, it will be three times of International Mobile Telecommunications-Advanced Standard (IMT-Advanced Standard), so 30 bit/s/Hz in the downlink and 15 bit/s/Hz in the uplink are expected. We should consider that, by increasing the frequency, the spectral efficiency will decline.

Latency is another crucial KPI of 5G and will be significantly improved compared to the previous generations. For control plane latency, which is the time of transition from the idle state to the active one, the value equals 10 ms, and for user plane latency, it is 4 ms for eMBB and 1ms for URLLC.

One of the main goals of 5G is providing seamless connections for mobile UEs. Mobility interruption, which is the time that a device cannot have coverage of a gNB (gNodeB), which is the base station for 5G, for transmitting its data, can play an essential role in such a case. As a result, for having seamless communications, it aims to be zero in 5G.

From the aspect of mMTC, battery life is one of the most critical KPIs, and the predicted target for it in the coming generation is beyond ten years. Besides these KPIs, a 5G network needs to be reliable, supports up to 500 km/h mobility for a device, and 10^6 devices in a square kilometer. Moreover, consuming up to 100 times less energy compared to LTE and having area traffic capacity up to 10 Mbit/s/m^2 are other improvements [4], [7]–[9].

In the beginning steps, 5G NR (New Radio) established connections through the LTE core network called EPC (Evolved Packet Core), which was defined in early Rel-15 drop and called the non-standalone mode. Then, the following specifications completed the standalone mode, which made it possible to have a fully connected end-to-end 5G network and was defined in regular Rel-15 freeze. The detailed information about the standalone mode and the frequencies can be found in [10], [11]. Deploying gNBs and 5GCN (5G Core Network) together creates a fully 5G end-to-end communication called SA (Stand-Alone) [12]. In the former implementation, a connection between a UE and a gNB is established utilizing 5G RAN (Radio Access Network) while using EPC. However, in the latter one, the deployed RAN and the core network are fully 5G ones. The inceptive goal for introducing 5G SA is by deploying low-band frequencies. Then using third-generation chipsets in devices will help to gain optimized performance during 2021 [1].

The deployment of 5G and implementing the infrastructure have speeded up recently. Since 2018, when the first 5G device was launched, the deployment of 5G has been accelerating, and as time passes, the transformation from the old generations to the new ones gets more interest. For example, in 2019, GSM/EDGE had a large portion of the

India region. However, in 2025, LTE and 5G are predicted to have 64 and 18 percent of mobile communication in this region, respectively [1].

Having a detailed look on different regions proves that the penetration of new technologies is speeding up, and more nations intend to exploit cutting-edge ones in their mobile communication. As an example, in the Middle East and Africa, LTE had 23 percent of mobile communication by the end of 2019. However, in 2025 LTE and 5G are expected to have 52 and 9 percent of the market, respectively. In 2020, around 190 million users could connect to 5G networks, and this number is predicted to be 2.8 billion in 2025. Some countries like South Korea are moving at fast paces, and it aims to have nationwide coverage for 5G by 2021. Table 1 shows the penetrations of LTE and 5G in different regions [1].

TABLE 1. The penetration of LTE and 5G in different regions.

	LTE by the end of 2019	Expected LTE by the end of 2025	Expected 5G by the end of 2025
The Middle East and Africa	23%	52%	9%
Sub-Saharan Africa	11%	29%	3%
India	49%	64%	18%
Southeast Asia and Oceania	34%	63%	21%
Central and Eastern Europe	43%	66%	27%
Latin America	51%	68%	13%
North-East Asia	88%	37%	60%
Western Europe	68%	43%	55%
North America	92%	26%	74%

Because of the new features and capabilities of 5G, new networking terms have been introduced. Network slicing is one of the most significant terms that has been included in 5G and opened a new horizon in mobile communication. The importance of network slicing becomes obvious when realized that different use cases of 5G require various resources, and the capacity needed by the end-users must be delivered efficiently based on the requisites. For example, eMBB demands high bandwidth, mMTC needs ultra-dense connectivity, and for meeting URLLC necessities, providing low latency is paramount [8], [13], [14]. These unique features make 5G capable of delivering new services to Industrial IoTs, TSCs (Time-Sensitive Communications), NPNs (Non-Public Networks) [5], reliable communication between vehicles [15], novel services to industrial stakeholders (i.e., vertical industries) [16], location-based services [17], and NB-IoTs [18]. Combining these features satisfies the requirements to build a low latency, high speed, fully connected world, one of the aspirations for the 5G era.

For attaining high bandwidth and meeting 5G requirements, radio frequencies that have been used in the previous generations, such as LTE and 3G, seem obsolete to be exploited in the new generation or needed to be refarmed; as a result, there is a necessity for using more suitable spectrum for 5G networks.

Frequencies between 300 MHz and 3 GHz are called radio frequencies, from 3GHz to 30 GHz are microwave bands, and from 30 GHz to 300 GHz are named mmWave. Each frequency has a distinct characteristic behavior that separates it from the other ones.

Mobile telecommunication was using bands up to 2 GHz until 3G. However, by the expansion of the telecommunication technologies and the advent of 4G networks, higher frequencies up to 6 GHz were employed because the lower ones were not able to fulfill the new demands. By the recent advances in mobile devices, using mmWave frequencies is becoming available too, and the IMT 2020, i.e., 5G, has the capability of deploying higher frequencies, including mmWave bands.

Although deploying higher frequencies (i.e., mmWave) have some advantages, they suffer from some drawbacks. The most important ones are: 1) they cannot cover wide areas, 2) are unable to penetrate materials, and 3) are absorbed by rain. The critical difference between mmWave and other frequency bands is the wide range of frequencies, which for implementing it, both devices and base stations need novel technologies compared to the previous generations [4]. In this paper, we will make a special attention to mmWave due to its grand role in 5G networks.

The existing downsides in 5G mmWave can lead to some issues, such as blockages that can affect the performance of these networks. It means that obstacles such as buildings, cars, and human bodies can block the channel, which is in charge of data transmission and degrade the performance of the network [19]. Although some solutions, such as beamforming and handover, can mitigate the adverse effects to some levels, they cannot compensate the signal quality reduction [20].

These problems can be more intense when requiring a reliable end-to-end connection over 5G. The reason is that the end-to-end reliable transport layer's widely used protocol, TCP (Transmission Control Protocol), has a critical role in the performance of end-to-end connections, so there will be a necessity of making it compatible to 5G networks [21]. Nowadays, TCP is the defacto protocol for establishing end-to-end connections over the Internet. As a result, if mmWave flaws impair its functionality, it can harm the performance of the protocol in achieving its goal throughout the Internet.

In order to gain high performance in 5G networks, the first important issue is the blockage problem, which can degrade the strength of mmWave signals by interrupting the communication and affecting the TCP congestion control mechanism due to the reaction of TCP to packet losses [22]. TCP is unable to perform appropriately when frequent interruptions occur in the network because it cannot distinguish a packet loss is due to congestion or other shortcomings of the 5G network, such as blockages or misalignments [19], [21], [23]. Therefore, in order to improve end-to-end performance and having stable connections, problems such as blockage need to be addressed;

if not, these adverse effects can decline the performance of networks and prevent them from fulfilling the 5G requirements. Due to the characteristics of the high-frequency bands, this problem is more highlighted in mmWave.

This work presents a comprehensive analysis to evaluate reliable end-to-end communications in mmWave 5G networks. In particular, it provides: 1) the analysis of the effects of reliable TCP communications in 5G mmWave networks, 2) the discussion of TCP mechanisms and parameters involved in the performance of 5G networks, 3) a survey of current challenges, solutions, and proposals, and 4) a feasibility analysis proposal of machine learning-based approaches to apply to improve reliable end-to-end communications in 5G networks.

The rest of the paper is organized as follows. Section II brings a brief description of TCP and its fundamentals. The 5G network procedure and the important parameters that affect the functionality of the network are discussed in Section III. Section IV discusses the TCP specifications and their impact on the performance of 5G networks. Current challenges and solutions are presented in Section V. Different aspects of available simulation environments have been given in Section VI. New machine learning-based approaches are considered in Section VII. Finally, Section VIII concludes the paper.

II. FUNDAMENTALS OF TCP AND TCP VARIANTS

TCP [24] is the most widely used protocol for reliable end-to-end communications in the transport layer of the TCP/IP protocol stack. Apart from end-to-end reliability, TCP has a congestion control mechanism to handle the unacknowledged packets (i.e., packets in-flight) in order to utilize the available bandwidth and retransmit the lost ones. This mechanism is mainly controlled by a so-called congestion window (cwnd), which is used to adjust the sending rate.

TCP congestion control mechanism incorporates four different phases called slow start, congestion avoidance, fast retransmit, and fast recovery. In the slow start, conventionally, the congestion window size is increased by one segment per each received ACK (Acknowledgment), so it is doubled in every RTT (Round Trip Time). This process will continue until cwnd size is larger than a defined threshold (ssthresh) [25], a packet loss occurs in the network, or the window size exceeds the maximum transmission window announced by the receiver.

Duplicate ACKs can be created due to lost segments in a network. When a packet is lost, the receiver acknowledges the previous segment and causes the creation of duplicate ACKs. However, in reality, out-of-order delivered packets that need to be reordered can be another source for duplicate ACKs. As a result, TCP waits for at least three duplicate ACKs to be sure that a packet loss has occurred instead of a reordering process. In this case, without waiting for the Retransmission Time-Out (RTO) to be triggered, TCP resends the lost packet. This phase is called fast retransmit because it speeds up the retransmission process in the network. When the fast retransmit is finished, TCP enters the congestion avoidance phase, not the slow start, which is called fast recovery. The goal of fast recovery is attaining high throughput during moderate congestion in a network.

When no ACKs are received by a sender for a specific period of time, RTO is triggered, and TCP initiates the retransmission mechanism by initializing the cwnd value and entering the slow start phase. The primary goal of this procedure is to handle the congestion collapse in networks.

One of the most well-known congestion control mechanisms that has been serving for a long time is AIMD (Additive Increase Multiplicative Decrease), which is the default congestion control strategy in some TCPs, such as NewReno. This mechanism can perform adequately in networks with moderate congestion status but is not suitable for networks that need TCPs with aggressive approaches in increasing and decreasing the sending data rate. As a result, with the emergence of highspeed networks and ubiquitous wireless technology, new TCP variants emerged to adapt to these networks leading to the appearance of TCPs such as CUBIC, HighSpeed, and BBR.

Based on the congestion control mechanisms, today there are different types of TCPs including loss-based [26]–[32] such as NewReno, HighSpeed, and Cubic, delay-based [33]–[37] like Vegas, and loss-based with bandwidth estimation (i.e., hybrid) [38]–[42] such as Westwood and Jersey.

From the point of the network types, TCP can be divided into five categories. The first group, which is the basic specifications for the other TCPs, was striving to deal with the congestion collapse problem in the network. Congestion collapse is exceeding the sending rate above the network capacity, which leads to the packet losses after that rate. The protocols that belong to this group could somewhat solve the congestion problems. However, they created a new issue that is underutilization of the network resources. The first protocol in this group, which was introduced in the late 1980s, was TCP Tahoe. Other TCPs in this category that appeared after Tahoe tried to improve its functionality by making some modifications or establishing new concepts such as delay-based TCPs as in TCP Vegas [37], or Vegas+ [43], which is an extension of Vegas [22].

With the emergence of new networks such as MANETs (Mobile Ad hoc NETworks), packet reordering was another issue that TCP needed to deal with, which lead to the creation of the second TCP group. These TCPs aim to put a border between the loss packets and reordered packets in order to distinguish them from each other. Because the previous protocols behave with both of them as lost ones, a reordering process could cause a congestion window reduction without considering the fact that no packet has been lost [22]. TD-FR (Time Delayed Fast Recovery) [44], Eifel [45], [46], and DOOR (Detection of Out-of-Order and Response) [47] are categorized in this group.

The third group's focus is on the different services that can exist in the network. TCPs that belong to this group try to give

different priorities to various services. For example, if a background service such as an automatic update tries to initiate a connection, it gets less priority in comparison to foreground services. To be more precise, these protocols make an unfair network to give the network resource to services with higher priorities [22]. TCP Nice [48], which is based on Vegas, can be mentioned as an example in this group.

By the advent of wireless networks and appearing random packet losses due to the channel fluctuations and interferences between the different frequencies, the necessity for establishing new TCPs to handle this issue was inevitable. Because the base assumption of TCP, which thinks that packet losses are the indicators of the congestion, could not serve in these networks. When a packet is lost due to wireless channel characteristics, it is not a sign of buffer overflow in the network, so reducing the sending rate in order to drain the buffers is not a sensible action [22]. The protocols in this group are based on TCP Westwood [49] and Westwood+ [50].

By appearing faster networks and networks with long-delays, the fifth group of TCPs appeared. This group aims to deal with the BDP (Bandwidth-Delay Product) problem. The BDP is the product of a data link's capacity (in bits per second) and its round-trip delay time (in seconds). The BDP problem is relevant for the legacy TCPs working in networks with high bandwidth such as optical networks or large delays such as satellite ones. As TCP is based on a sliding window operation, the BDP results in an amount of data measured in bits (or bytes) that is equivalent to the maximum amount of data on the network that has been transmitted but not yet acknowledged. In terms of maximizing throughput, TCP sending rate and receiving buffers must be adapted to take advantage of the BDP. To be more precise, conventional TCPs can not utilize the high available bandwidth in networks such as optical ones. The main reason is deploying techniques like AIMD in the congestion avoidance phase, which makes TCP unable to use the available resources in this kind of networks. Previous TCPs prefer to use a conservative technique in discovering available resources in a network, which leads to the underutilization of higher bandwidths. AIMD technique is suitable for networks with small bandwidth or RTT, not for the ones that are capable of high BDPs. When a TCP wants to discover all the resources in a network, it takes at least an order of the BDP if there are not any packet drops in the network. As a result, as the bandwidth and RTT increase in a network, it takes more time for conventional TCPs to reach the highest available sending data rate. This situation can even be more complicated by occurring packet drops in the network. These flaws were the causes of highspeed TCPs emergence.

The protocols that belong to this group try to use a network's resources efficiently in a fair way and are able to react quickly to the network changes [22]. High-Speed TCP [27] was the first protocol proposed, and the others tried to improve its functionality.

Based on the popularity and implementation of current networks, main studies in 5G networks focusses on those summarized in the following subsections. The predominant reasons for choosing these TCPs are that, NewReno is one of the basic TCPs in designing other ones, CUBIC is the default protocol since Linux 2.6.26, HighSpeed is one of the primary candidates to be deployed in high-speed networks, and BBR is a cutting-edge protocol that is exploited in some of the Google's services due to its novel and suitable congestion control mechanism.

A. TCP NEWRENO

NewReno [51] follows an AIMD approach and is an extension to TCP Reno [25] with a slight modification in its fast recovery phase. The principal goal of NewReno was to overcome the problem that Reno had when several packets are lost during a single congestion window. This problem impairs the performance of Reno by a consecutive halving of the cwnd due to identical entering and exiting mechanism to fast recovery for each loss. This mechanism can have some issues in functioning properly because a single congestion event may cause all of these losses for a single congestion window. As a result, after the first cwnd having, the protocol should resend other lost ones without reducing the cwnd size. NewReno introduced a refined fast recovery to solve this problem by preventing multiple cwnd halving.

In the congestion avoidance phase, NewReno increases the cwnd size by 1/cwnd for each received ACK. Therefore, for increasing the cwnd size by one during congestion avoidance, the entire cwnd should be acknowledged. By perceiving three duplicate ACKs, NewReno reduces the cwnd by half and enters the fast retransmit phase.

B. TCP CUBIC

This variant is the default TCP [26] in Linux since Kernel 2.6.26, Android, and MAC operating systems [23], [52]. The key reason for CUBIC for being popular is in its mechanism in adjusting cwnd efficiently in high-BDP networks. Moreover, when CUBIC is deployed with other TCPs, it can attain an adequate value for fairness.

The mechanism that the CUBIC approaches the congestion problem is based on a cubic function, and there are two different ways in increasing or decreasing the size of the congestion window. The first one is a convene portion when the cwnd size ramps up quickly to the size before the last congestion event. Next is the convex mode, where CUBIC probes for more bandwidth slowly at first, then continues rapidly. CUBIC considers the time right before the last drop and tries to reach that capacity in fast paces during a short time.

C. TCP HIGHSPEED

This TCP variant [27] is suitable for the networks with high bandwidth-delay products and has been designed for networks in which a fast growth of cwnd is essential. The reason for developing this protocol is that previous TCPs perform deficiently in networks with large bandwidth-delay products. This protocol makes some slight changes to the congestion

control mechanism of the standard TCP to overcome this problem.

D. TCP BBR

TCP BBR (Bottleneck Bandwidth and Round-trip) is a cutting edge congestion control algorithm that was developed by Google in July 2017 [53] and is being used on Google, YouTube, and GCP (Google Cloud Platform). BBR, as the name indicates, tries to keep the most excellent cwnd size based on the current bottleneck bandwidth and RTT and tries to achieve higher bandwidth with low latency.

III. 5G MMWAVE NETWORK PROCEDURES AND PARAMETERS FOR RELIABLE END-TO-END COMMUNICATION

5G network parameters and their characteristics have a critical effect on the delivered performance to end-users. Procedures such as handover, techniques like beamforming, parameters like RLC (Radio Link Control) buffer size, system architecture, and ultra-lean design, plus how they are implemented, can play essential roles in 5G networks. This section aims to investigate these parameters and procedures and their impacts on the behavior of reliable end-to-end communications on 5G networks.

A. HANDOVER AND BEAMFORMING

Handover or handoff means the process of changing an ongoing data session from one cell to another. Beamforming focuses on sending powerful signals toward a particular device. The reason for deploying this technique is to prevent signal attenuation and prevent bandwidth degradation in situations such as a blockage. One of the main reasons for exploiting these techniques over 5G networks is because of compensating for the intermittent nature of mmWave signals. For example, when a UE exits from a cell's coverage and enters another cell's area of coverage, in order to prevent the connection termination, handover can be initialized to establish a connection with the new gNB. Moreover, when the capacity of a cell is reached, and a new device intends to connect to it, a handover process can be triggered in order to find another gNB that can serve the new UE.

One of the most important sources for handover initialization in 5G networks is due to blockage occurrence. When a connection is blocked by an obstacle, handover strives to find another cell to keep the connection on. The mentioned situations, i.e., blockage, outage, and reaching the maximum capacity of a gNB, can affect the performance of TCP in keeping the packet drops low. In this case, handover and beamforming can relieve the effect of negative factors on the performance of TCP and improve its functionality by preventing massive packet drops that may happen. If we could have a channel with an adequate amount of bandwidth with the help of these techniques, it could prevent the throughput reduction and RTT increment of TCP to some levels. However, it may not omit the adverse impacts in some situations.

There are two kinds of handovers, horizontal and vertical. Horizontal handover refers to gNB changes, i.e., changing from a gNB to another one to maintain the connectivity. However, when a wireless technology change occurs, for example, from 5G to 4G, it is called a vertical handover. Experiments in [54] exhibit that both handovers can degrade the performance of the network, though this effect can be severe in vertical ones.

B. BLOCKAGE AND MISALIGNMENT

Apart from high packet loss probability, there are some issues such as blockage and misalignment in 5G networks. Blockage means that high frequencies, i.e., mmWave cannot pass through obstacles, and misalignment happens due to non-matching beams of transmitters and receivers. The existence of these problems can degrade the performance of 5G mmWave networks. In particular, they have a dramatic impact on the performance of TCP, which is responsible for establishing reliable end-to-end connections. These problems make a transition from a LoS (Line-of-Sight) connection to a NLoS (Non-Line-of-Sight) one. In LoS, the data transmission can be performed through the established connection between the user and base station, but in NLoS modes, the reduced bandwidth of the channel can harm the performance of the network.

The reason for throughput degradation in NLoS mode is that SINR (Signal-to-Interference-Plus-Noise Ratio) cannot reach expected high values. As a result, the quality of the received signal by the UE becomes very low. Secondly, when NLoS connections exist, a lot of temporary disconnections may happen in the network, which can confuse TCP in adjusting its congestion window size. These interruptions can vary based on the size of obstacles and speed of UEs, which lead to different ones from short to long failures. Although both disconnections can affect the performance of TCP, the effects of the long ones are stronger due to the high probability of triggering the RTO, which leads to a congestion window initializing and slowing down the sending rate dramatically. This problem can even be worse when the network is not congested because initializing the cwnd size in this situation can degrade the throughput profoundly. This confusion in TCP functionality leads to a high end-to-end throughput degradation of 5G mmWave networks [19], [21]. There are some primary solutions, such as installing several gNBs in order to broaden the LoS regions or putting some relays in LoS areas to get the signals and reflect them to NLoS areas [19] and vice versa (capturing a UE's signals in NLoS areas to repeat them to a gNB in LoS area). That mitigates the adverse effects of a UE presence in NLoS conditions. However, they are expensive or can alleviate the problem to some levels, but are not good enough to eliminate it.

Blockage can happen because of different objects such as buildings, buses, cars, human bodies, and pillars. Almost anything except some thin materials like clear glass can create it. It can even be more emphasized when we know it is hard for 5G mmWave signals to penetrate a hand or human

body and makes the problem more difficult in urban areas. Another issue that makes the problem more severe is using UV-protective windows. UV (UltraViolet) rays are in the middle frequency ranges and mostly created by the sun and can be harmful to the human. These rays can lead to some diseases such as skin cancer, so it is important to protect humans' skin from being exposed to them. In order to prevent the damage that this ray can create, using UV-protective windows is becoming more common. Because these windows can block the middle frequency ranges, i.e., UV, they can also attenuate 5G signals and reduce the quality of the received ones, which leads to performance reduction.

Several parameters can strengthen or mitigate the effect of blockage on the performance of TCP. Small size obstacles have slighter effects compared to the large ones, especially when the UE is moving and can pass them quickly. However, when the size of the obstacles gets bigger, the chance for triggering the RTO becomes higher. As a result, large obstacles can create intense negative results. Blockage creates longer RTTs, higher packet loss probability, and triggers TCP RTO, which all of them degrade the performance of TCP over 5G mmWave networks.

When there is a blockage in a static situation, the chance of passing the obstacle is low, and the negative effect will be more stringent. As a result, in some cases, being dynamic can be beneficial by increasing the chance of reconnection between a UE and a gNB. The static mode can create persistent conditions and reduce the quality of the received signal by a UE intensely. Moreover, employing handover can reduce the negative effect of blockage by changing the associated gNB and keeping the connection on [19], [21], [54], [55].

In addition to obstacles, other factors can harm the performance of 5G too. Distance between a UE and gNB is one of these parameters that can play an essential role in the performance of the network. However, the negative impact that distance can have is low compared to the blockage. In addition to the blockage and the distance between a UE and gNB, the orientation between them is another effective player in the performance of 5G networks. In this case, a 90 degree one is the worst case, and a 0 degree one is the most favorable.

Misalignment is another severe issue in environments with high mobility compared to static conditions. This means that when a transmitter and receiver phases are not matched, a persistent connection cannot be established. As a consequence, a large number of packets will be lost and can cause a significant drop in the value of SINR, which makes the quality of the signal low.

Although there are some techniques such as beam sweeping [19], which tries to match the pairs after misalignment occurrence, these techniques lose their efficiency when the UE keeps its mobility. If beam sweeping finds a matched pair between the UE and gNB, there are not any guarantees that the communications will remain consistent because misalignment can happen frequently.

These issues can be worse when frequent initializing occurs due to a mobile UE because of several handovers, which causes a reduction in the performance of TCP. In this case, the high bandwidth of the 5G networks can be wasted, and TCP cannot reach its high throughput [19].

1) BLOCKAGE EFFECT ON DIFFERENT DEPLOYMENT SCENARIOS

The deployed scenario plays a vital role in the effect of the blockage on the performance of reliable end-to-end communications on mmWave 5G networks. Because reliable end-to-end communication highly depends on the functionality of TCP, the protocol performance will differ in various deployment scenarios. In situations that the blockage effect is low, TCP can work more efficiently. In contrast, in the circumstances with a high number of blockages, it will have some difficulties in maintaining the high throughput and needs more attention. As a result, knowing the effect of blockage on different deployment scenarios can give an insight and provide a clear vision of creating new protocols.

Generally, there are twelve defined scenarios [7] including indoor hotspot, dense urban, rural, urban macro, high-speed, extreme long-distance coverage in low-density areas, urban coverage for massive connections, highway scenario, the urban grid for connected cars, commercial air to ground, light aircraft scenario, and satellite extension to terrestrial.

TABLE 2. Impact of the blockage on different deployment scenarios.

Deployment Scenario	Carrier Frequency	Number of Obstacles	Blockage Effect
Indoor Hotspot	4, 30, and 70 GHz	High	High
Dense Urban	4 and 30 GHz	High	High
Rural	700 MHz and 4 GHz	Low	Low
Urban Macro	2, 4, and 30 GHz	High	Medium
High-Speed	4, 30, and 70GHz	Low	Low
Extreme Long Distance Coverage in Low Density	3 GHz and below 1 GHz	Low	Low
Urban Coverage for Massive Connections	700 and 2100 MHz	High	Low
Highway	6 GHz	Low	Low
Urban Grid For Connected Cars	6 GHz	Medium	Low
Commercial Air to Ground	Below 4 GHz	Low	Low
Light Aircraft	Below 4 GHz	Low	Low
Extension to Terrestrial	1.5, 2, 20, 40, and 50 GHz based on the deployments scenario	Low	Low

Table 2 compares the different scenarios and the level of blockage effect on each one, as Low, Medium, and High. Low means that the scenario is almost immune to blockages, medium indicates moderate effects of blockages, and high means blockages can impair the communication. Considering the used carrier frequency, layout, user distribution, and speed, the effect of blockage can be different.

In indoor hotspot, which focuses on high throughput for users inside buildings and intends to use high carrier frequencies, the blockage effect can be intense. The first reason is existing an enormous number of obstacles that create interruptions in the connections and make the channels intermittent. The second reason is using high frequencies because we know when the frequency increases, the signal is attenuated easily. As a result, in this situation, having consistent communication is almost impossible.

Dense urban deployment focus is to provide high throughput for a large number of users in a downtown or dense areas inside a city. In this scenario, besides data rate, coverage is another factor. Using higher frequencies is one of the key characteristics of this scenario in order to support high bandwidth and data rate, which makes this scenario sensitive to obstacles. Besides high frequencies, there are a large number of obstacles like building, cars, buses, and humans in this deployment, which can have both positive and negative effects. As the number of obstacles grows, the probability of having an interruption because of a blockage increases. On the other hand, existing large obstructions like buildings in a high number can reflect the signals which have been sent by gNBs. As a result, it can help to mitigate the negative effect of the blockage. In this scenario, the majority of users are inside buildings or moving at a speed of 3 km/h, which make it hard to have constant connections.

The rural deployment aims to cover large areas. In this scenario, the most crucial factor is supporting mobile vehicles in broad areas. Because of using low frequencies around 700 MHz and 4 GHz, and areas with a few numbers of obstacles, the blockage cannot have a substantial effect in this scenario.

Like the rural deployment, the urban macro scenario focuses on the coverage of wide areas but inside a city. It uses both higher and lower frequencies based on the requirements. Most of the UEs are considered to be inside buildings, which makes it hard to reach them. Although blockage can have adverse effects in this scenario, it is less than indoor hotspot deployment. Because in the first one, all of the users are considered inside buildings, and 70 GHz frequency can be used, which is highly sensitive to obstacles. However, in this scenario, some of the users are outside, and frequencies around 30 GHz or even lower are deployed, which mitigate blockage effects.

The high-speed scenario strives to cover UEs inside high-speed trains. High mobility up to 500 km/h is the key characteristic of this scenario. For supporting all of the users, a lot of small cells (i.e., gNBs) are deployed along tracks. By using handover techniques, the blockage effect can be eliminated entirely because a UE is connected to a gNB all the time. However, existing frequent handovers can have its own adverse effect.

For large areas with a slight number of users, extreme long-distance coverage in low density can be the primary candidate. In this scenario, macrocells with frequencies below 3 GHz are used to provide extensive coverage with moderate bandwidth because a high sending data rate is not a priority. As a result, the blockage is not an essential issue in this scenario, especially when frequencies under 1 GHz are deployed.

For fulfilling mMTC requirements, the urban scenario for a massive number of connections can be exploited. The most important parameter here is the high number of devices which can be indoor, outdoor, or inside vehicles. However, by considering the penetrating power of low frequencies, which are used in this scenario, it is almost immune to the blockage problem. Frequencies around 700 MHz or 2 GHz are favorite ones in this scenario, which can satisfy the requirements.

There is a scenario similar to high-speed but with lower mobility supporting up to 300 km/s, which is called highway and focuses on supporting mobile vehicles in highways. Having used a lot of small cell, not existing obstacles, open area of connection, and using frequencies around 6 GHz are the main characteristic of this scenario which mitigate adverse effects of the blockage.

When freeways end into cities, they can cause heavy traffic with a large number of cars. For supporting this scenario, the urban grid for connected cars is the leading candidate. The aim of this deployment is providing reliable and available connections with an acceptable latency for cars. Like high way scenario, this one is deploying macrocells with frequencies around 6 GHz too, which makes it somewhat immune to blockages.

Both commercial air to ground and light aircraft are for supporting machines on the air. In the first case, the goal is providing connections for UEs boarded on airplanes and in the second one for UEs boarded on helicopters and small airplanes. The main goal in both of them is supporting a large area of coverage upward. Throughput and user density are not KPIs here, and providing basic data and voice services is convenient by using frequencies below 4 GHz. Existing almost zero obstacles and using lower frequencies omit the negative effects of the blockage. However, it was not a severe problem from the beginning.

Satellite extension to terrestrial is the last deployment scenario. It is useful for supporting those areas where providing terrestrial services are impossible or not noteworthy to be deployed. Because of using satellites for broadcasting, the blockage is not an issue in this scenario.

C. 5G CORE NETWORK AND ARCHITECTURE

5GCN is based on the EPC with three novel improvements: service-based architecture, network slicing support, and SDN (Software-Defined Networking)/NFV (Network Function Virtualization).

Being based on service-based architecture means that the concentration is on the provided services and functionalities by the core network. Network slicing is a new term introduced in 5G and means, instead of separating a network into different physical parts, it is divided into some logical parts based on the service demands and necessities. In such a case, different slices are run on the same physical infrastructure, but from the user view, they seem separate. Control-plane/user-plane

separation, based on SDN/NFV, is one of the new features supported by 5GCN in order to use different capacities within them. As an example, it is possible to use more capacity for the user-plane without affecting the control-plane.

5G Core network and architecture must be analyzed in detail in future research in order to see the impact to improve TCP functionality. The service-based architecture can be an enabler in exploiting different TCPs for different services based on its needs. For services with high data rate necessity, high-speed TCPs can be used, or for the ones with delay sensitivity, the appropriate TCPs can be deployed. Other schemes can include using proper TCPs in different slices to seek the optimal functionality for the network. Finally, By separating the control-plane and user-plane, the deployed TCPs for each one can be distinct, and the ideal one can be chosen.

Deploying 5GCN eases the path to SA 5G networks and makes it possible to use the full privileged of the new generation. New Features, i.e., service-based architecture, network slicing, and SDN/NFV, can be enabled based on a service requirement, which leads to an improved end-to-end user experience [12].

D. RLC BUFFER SIZE

RLC buffer size can have a crucial effect on the performance of 5G networks by masking the packet losses to higher layers' protocols. Although for attaining higher performance exploiting large buffers can be beneficial, it leads to higher latency values. There are two main reasons that large buffers yield higher performances. First, when big buffers are used, the chance of packet drops due to buffer overflow becomes low. However, deploying big buffers can make long queues and causes packets to wait longer in buffers, which leads to bufferbloating issue. Bufferbloat caused by large buffers leads to higher latency values in a network and can be alleviated by reducing buffers size.

On the other hand, Reliable transport layer protocols, such as loss-based TCPs, will be affected intensely from a high number of losses when buffers sizes are reduced. Secondly, the network becomes less sensitive to high link variations of the 5G mmWave channel. Because a small buffer size can be filled up quickly in NLoS states and start to drop packets sooner than a buffer with a large size. As the number of packet drops goes higher, the sending rate will be decreased, and it will affect the performance of the network.

In contrast, when small buffers are used, latency can be decreased but at the cost of declined performance. Maintaining a tradeoff between performance and latency is critical, especially when remote servers are deployed and can affect TCP severely. This tradeoff can be attained by using AQM (Active Queue Management) techniques, such as CoDel [56] and Fq-CoDel [57]. However, these techniques require some modifications to be adapted to 5G networks [21].

E. ULTRA-LEAN DESIGN

One of the most critical problems that current mobile communication has is the existence of ''always-on'' signals, especially in highly dense areas with an extreme load of traffic. The presence of these signals is regardless of user traffic and can occupy a portion of the bandwidth in the network. Signals such as base station detection, system information broadcast, and channel estimation reference are categorized under always-on signals. The ultimate goal of TCP is to handle the congestion issue in networks and having fairness between users and flows. However, always-on signals can increase the traffic in a network and affect TCP performance. As a result, deploying ultra-lean design can help to mitigate the amount of traffic in networks, reduces the congestion events, and improves TCP functionality.

Moreover, energy consumption is another negative aspect of these signals, and they can create interference too. The ultra-lean design in 5G networks strives to reduce the use of always-on signals. As a result, it can enhance energy consumption and prevent bandwidth wastage, and mitigate signal interferences in the network, by turning on the signals when they are needed and turning off when they are not.

Furthermore, reducing channel interference and bandwidth usage can lead to enhanced user experience as the transport layer protocols will encounter less sudden changes in the network.

F. LATENCY

One of the most stringent requirements for 5G networks is the value of latency. Latency is the time interval from when a source sends a packet to the time the destination receives. Latency will be improved significantly in 5G networks, especially for critical latency devices that exploit URLLC use case of 5G. The value of latency can be damaged in 5G networks due to the existence of adverse impacts, and the occurrence of blockage and misalignment can create long latencies in the network. Therefore, this parameter must be a performance optimization one.

Generally, conventional TCPs can benefit from reduced delays in a network. One of the main efforts in improving TCPs functionality is to bring the servers close to users by using techniques such as CDN (Content Delivery Network), in order to reduce the delay, which can improve TCP's functionality, especially the loss-based one. The reason for reducing the delay in a network is that most of the TCPs increase the congestion window in every RTT; as a result, reduced delays lead to shorter loops and faster reactions for them. In this case, the reduced latency in 5G networks can help TCP in ramping up to the high sending rates and having a better functionality. In addition to faster reaction, by improving latency and having similar values, fairness between different flows can be enhanced.

To enhance latency, one of the practical factors to be taken into account is the location of nodes and servers, which are affected by the architecture of the 5G network, especially 5GCN (5G Core Network).

One solution for the user-plane side for reducing the latency can be improving the functionality of UPF (User Plane Function), which is responsible for routing and

forwarding the packets and is the gateway between the RAN and external networks. How this function operates can have a direct impact on the value of latency. Another solution can be using edge computing by the support of network slicing and running part of the user-plane applications near the core network. By deploying edge computing in this way, the value of latency can be declined.

IV. TCP MECHANISMS AND PARAMETERS INVOLVED IN THE PERFORMANCE OF 5G NETWORKS

The transport layer has a significant role in determining end-to-end performance in a network. Although the new mobile generation provides high bandwidth, without an effective transport layer, which is able to utilize the available bandwidth of mmWave in 5G networks and deal with the existing issues such as blockage and misalignment, this bandwidth will be wasted, and reaching high data rates will be challenging. As we said, TCP is the most widely used protocol in the transport layer and is the key player in end-to-end reliability. Various TCP mechanisms, such as congestion control and loss detection, can have a great effect on the delivered performance to the final user. This section aims to give an overview of the different mechanisms, parameters, and analyzing their effects on the performance of 5G networks.

A. TCP PACKET SIZE

Adapting MSS (Maximum Segment Size) to MTU (Maximum Transmission Unit), and optimizing its values for 5G networks is a challenge. The default value of MTU has been used for a long time and has been performing properly in the previous generations because the moderate bandwidth of them did not need big MSS to deliver high throughput. On the other hand, MSS conventional size has a couple of adverse effects on the performance of TCP over 5G networks. First of all, the small size of MSS degrades the performance because TCP cannot utilize the high capacity of the network. This small size is a hurdle on the way of TCP in utilizing the high bandwidth of 5G mmWave networks. An investigation on the impact of the size of MSS was done in [21], and the results show that loss-based TCPs such as NewReno, adds up to their congestion window sizes slowly when the standard value for MSS is used, which makes protocols underutilize the high bandwidth of 5G mmWave networks. If the size of MSS increases, it leads to faster growth of the sending rate, so a higher performance will be achieved in a shorter time, and it can also help when recovering from congestion states. When RTO triggers, the sending rate is initialized, and TCP enters the slow start phase. If the MSS is small, reaching to high sending rates can take more time. However, having a large one can help to recover faster. By increasing the size of MSS, TCP can quit the slow start quickly, because of the exponential growth in this phase. As a result, larger MSS can have a positive effect on the slow start phase. In the congestion avoidance phase, cwnd is added linearly, so the increased value of MSS can help the protocol to ramp up quickly and utilize the available bandwidth. As a result, having larger MSS in a network with a high data rate can assist the protocol to attain higher throughput. We should notice that increasing MSS can compensate for the throughput degradation issue to some levels, and it is not a perfect solution. The main focus should be on adapting the congestion control mechanism to 5G networks.

Moreover, when small MSS is exploited, more overhead is forced to the network because of the need for a large number of headers. As a result, using a large size for MSS reduces the number of overheads in the network. Finally, small MSS means transmitting more number of segments, which leads to a higher number of ACKs in the networks, which can exhaust the network. MSS can play a significant role in the performance of NewReno, CUBIC, and HighSpeed, but not on BBR. The reason is that loss-based TCPs have different congestion avoidance mechanisms compared to BBR. Loss-based TCPs, as the name indicates, rely on packet drops and try to adjust the sending rate when detecting a loss in the network or increase the sending rate in non-congested conditions based on the size of the MSS. For example, as it was mentioned, NewReno reduces its sending rate to half when it receives three duplicate ACKs during the congestion avoidance phase and increases it by 1/cwnd for every received ACK. In both cases, it performs based on the order of cwnd. However, BBR is a model-based TCP, and packet drops do not affect its functionality, and it does not try to adjust the sending rate based on the order of cwnd. It strives to measure the bottleneck bandwidth and the minimum RTT of the network and works based on them. The goal of this protocol is to deliver the maximum possible packets during the shortest available delay. Because of this mechanism, BBR tries to work close to the bottleneck bandwidth without considering the size of MSS.

In [21], the authors compared 1500 bytes MTU (about 1.4 KB size for MSS) with 14 KB MSS. In the latter case, results showed a better performance. The reason is that when loss-based TCPs are in the congestion avoidance phase, they increase the size of cwnd by one MSS in every RTT, which is done when all packets are appropriately acknowledged.

Although standard MSS size has been performing efficiently for years, employing small ones may underutilize the high potential of 5G networks. Therefore, one challenge is to adapt its size to 5G networks.

B. INITIAL CONGESTION WINDOW SIZE

When TCP starts sending data, the first phase is the slow start. In this phase, congestion windows size starts from the minimum value, which can be one, two, or four [58], and then by receiving every ACK, TCP adds up cwnd size by one. Although this mechanism aims to probe the link and can be efficient in networks before 5G, it seems not suitable for the new generation. The first reason is that the sending data rate can be tremendously huge in 5G networks because of the high bandwidth. However, the small starting number for cwnd and increasing it even in this exponential way can take a long time to utilize the full potential of 5G networks. Secondly, when

RTO triggers and TCP enters the slow start again, initializing and beginning from one in a network that can support high data rates is astonishingly wrong. The first step could be using high values for the initial congestion window, so TCP can get the benefit of it in getting higher sending rates by doubling it in the slow start phase. In this case, we should be careful about a premature transition from the slow start to the congestion avoidance. As a result, it may be necessary to modify the slow start threshold by considering the initial congestion window.

In addition to increasing the initial values of the congestion window, it seems that new approaches are essential to be proposed to deal with this issue. These approaches can be as simple as testing new values or proposing smart and intelligent solutions considering machine learning techniques to set the initial value of the cwnd based on the network parameters such as loss probability, available bandwidth, cwnd size in the last packet drop, and time intervals between drops.

C. EXPONENTIAL BACKOFF RETRANSMISSION TIME-OUT

RTO effects can be severe in long time disconnections. When long failures occur, the probability of triggering RTO is high, which can affect the performance of 5G networks negatively. The cause of this degradation is triggering RTO when the cause is not congestion. There are circumstances where the network is not congested and performing well, but having an RTO triggered by an obstacle can also lead to initializing the cwnd size and entering the slow start, causing a dramatic reduction in the performance. Moreover, small obstacles can create duplicate ACKS, which cause TCP to start fast retransmit, and then fast recovery and prevent the protocol from functioning adequately. These issues can be severe when static situations or long distance between a UE and a gNB exist in the network because the chance of triggering RTO in each blockage will be high, and techniques such as handover seem useless in these situations. As a result, for preventing the performance degradation caused by RTOs during blockage, some solutions need to be proposed.

Link-layer retransmission [4], [59] is a method that can help the reduction of the number of TCP retransmissions by hiding some of the losses from the transport layer. In this case, other layers of the 5G protocol stack, such as MAC (Medium Access Layer) and RLC layer, try to mask some of the losses from the upper layers. Hybrid Automatic Repeat reQuest (HARQ) is the deployed method in the MAC layer. In addition to the MAC layer, the RLC layer, which resides on top of the MAC, can do retransmission to some levels when the AM (Acknowledged Mode) is enabled. By considering the limited number of attempts in the MAC layer in recovering the lost packets, RLC can compensate it and help in retransmitting more lost packets. If the UM (Unacknowledged Mode) is activated in RLC, this procedure will be halted. Being timeliness is the advantage of the mentioned methods. However, the existence of some limitations is the downside of the link-layer retransmission compared to TCP one. For example, the number of retransmissions in the MAC layer is usually limited to three attempts [59]. However, these retransmissions can aid TCP, especially in NLoS mode, in which there is a high probability of losing packets.

The most important reason for link-layer retransmission and hiding losses from the upper layers is because of giving some guarantees in delivering packets. However, this can force shuffling in TCP packets order and can lead to the reordering problem. Moreover, these retransmissions increase delay, so TCP RTO, in some cases, can also expire. In conclusion, tuning these parameters is also a challenge in improving the performance of TCP over 5G networks.

There is a comparison of the TCP functionality over mmWave 5G networks with and without link-layer retransmission in [59]. The results showed that in LoS mode, the distance between a UE and a gNB has a significant impact on the throughput of TCP. In this case, when the distance is low, deploying link-layer retransmissions do not have a significant effect, however, by distance increment, the throughput of TCP declines in the absence of link-layer retransmissions. In a nutshell, the principal reason for the throughput decrement without link-layer retransmission is that TCP cannot handle all of the retransmissions efficiently by itself, especially in NLoS mode. We should notice that it is beneficial to have a trade-off between throughput and latency as the link-layer retransmissions can harm the value of latency. As the results in [59] showed, the best value for latency is for the time that only TCP retransmits the lost packets, i.e., no HARQ plus RLC UM. To sum up, deploying link-layer retransmission can help to increase the throughput of TCP, especially in higher distances at the cost of increased latency.

D. TCP LOSS DETECTION

As it was said in section two, conventionally, there are two ways for TCP to detect a loss in a network, three duplicate ACKs for indicating moderate congestions or triggered RTOs for heavy congestions. These mechanisms could perform properly in wired networks. However, existing issues such as blockages in 5G mmWave networks can damage their functionality, and more attention is needed to improve loss detection mechanisms. These efforts can focus on distinguishing packet losses due to congestion from other losses that can have sources except congestion.

Moreover, by deploying TCP Selective Acknowledgements (SACK) [60] option, the sender will be informed of the successfully transmitted segments, then retransmit only the lost ones. This mechanism prevents the sender from resending the correct ones. The use of the TCP SACK option increases the amount of packet overhead by improving the retransmission mechanism. However, being restricted to 40 bytes for the TCP option field forced by TCP specification is a hurdle on the way of implementing SACK in large BDP networks.

E. FAIRNESS

One of the principal achievements of TCP is reliable connections through fair networks. It means that when there are

several flows in a network, they get the same proportions. However, retaining fairness in 5G mmWave networks is a challenging issue. The reason is that this feature is directly connected to RTT values [52]. The value of RTT (which can be reduced by reducing latency) has a direct impact on fairness. Imagine two flows are deploying NewReno, one with an RTT value of 20 ms the other 25 ms. NewReno (like most of TCPs) updates the sending rate by receiving ACKs. For example, when it is in the congestion avoidance phase, by getting each ACK, it increases the sending rate equals to 1/cwnd, i.e., one every round trip time. As a result, the flow with a low value for latency can increase the sending rate quickly, and it leads to utilizing more portion of the network resources and leading to unfairness.

As a consequence, if a flow has a shorter RTT, it can ramp up to higher throughput quickly and gets more shares of the available bandwidth; as a result, forcing unfairness to the network. This unfairness, which is due to increased RTTs caused by NLoS states, can be intense in the existence of scenarios with a lot of hurdles, such as urban deployments. The reason is that when the number of LoS to NLoS transition increases, it leads to increments in the RTT value. It can be more severe while a UE can see the gNB, and another one cannot establish a proper connection because of being behind an obstacle. This NLoS state is going to increase the value of the RTT for the corresponding flows, as a consequence, the share of the user from the bandwidth will decline dramatically, and it damages the fairness intensely. Therefore, the proposals that improve reliable end-to-end communications in 5G networks must take into account fairness among UEs.

To sum up, the unique characteristics of 5G mmWave networks are barriers on the way of implementing TCP and having reliable end-to-end communication. As a result, making some modifications to TCP in order to make it suitable for 5G mmWave is a necessity.

The next section aims to present the challenges of implementing TCP over 5G networks, the made efforts on the way of deploying TCP such as new schemes and investigations, and recent advances.

V. RELATED WORK

As it was mentioned in the previous sections, using TCP over 5G networks can be challenging. One of the most promising approaches to mitigate or eliminate the adverse effects of 5G mmWave networks such as blockages, is modifying or adapting some mechanisms of TCP. Another alternative can be designing a new protocol from scratch and replacing the existing protocols. Both of the mentioned techniques can improve the performance of 5G networks and have their advantages and disadvantages. When designing a new protocol, existing issues can be addressed in detail, solved more efficiently, and the chance of resulting in an improved performance becomes high. However, there is no guarantee that it will work with other protocols, and there is a probability of having some problems such as fairness when coexisting with other ones. Moreover, testing environments may need to be modified in order to be compatible to evaluate new protocols accurately.

In addition, the main hurdle on the way of creating a new protocol is that it is almost impossible to replace a protocol in the Internet Stack because the existing ones are widespread on the Internet and have been around for a long time. For that, as an example, QUIC (Quick UDP Internet Connections) [61], the protocol developed by Google, is based on UDP (User Datagram Protocol), which intends to reduce end-to-end latency.

There exist several investigations on TCP over 5G, especially 5G mmWave networks [19], [21], [23], [55], [59], [63], [71]. The most significant motivation in modifying or optimizing TCP and making it capable of being deployed in high-speed networks, especially 5G mmWave, is its end-to-end reliability. The proposals of TCP over 5G need to overcome a variety of constraints like throughput degradation, latency increment, fluctuation in adjusting congestion window, and fairness issue when several flows co-exist [21], [62]. However, the first step to address these issues is to detect them and then set the goals. This section first addresses a more in-depth investigation of the general TCP proposal. Then, TCP based throughput enhancement are addressed. Other important groups of investigations are explained as they are focused on latency and fairness, and multi flows versus a single flow.

A. A DEEPER INVESTIGATION

In order to approach a problem, the first step is providing a clear insight into it. As a result, for finding the characteristics of different TCP variants, a thorough investigation of TCP over 5G mmWave was done in [21]. There, different TCPs in various situations were analyzed to have a more in-depth view on the functionality of the protocol. The aspects that they evaluated were, deploying edge servers (with minimum RTTs on the order of 4 ms) versus remote servers (with minimum RTTs on the order of 40 ms), handover and mobility effects, different congestion control algorithms and their impacts, TCP packet size, and RLC buffer size effects. They were analyzed in two different scenarios, including a high-speed scenario where a UE is inside a moving train and a dense urban environment. Four different versions of TCPs (NewReno, HighSpeed, CUBIC, and BBR) have been analyzed throughout the simulations. The results revealed that when edge servers are deployed, it can improve loss-based TCPs because of the short control loop feature for them. However, there are some exceptions to this conclusion, and by using small buffers, the goodput for CUBIC and HighSpeed in remote server mode is higher than the edge sever one. Among the four analyzed TCPs, BBR shows the best performance along with using big buffers. However, it cannot reach the saturated achievable goodput, which is 2 Gbps for 28 GHz spectrum. We should consider that high goodput can be attained at the cost of higher latency; however, by deploying edge servers with small buffers, this negative impact can be compensated for up to some levels. Among these loss-based

TCPs, the best goodput is for HighSpeed, for it increases the size of cwnd aggressively in high BDP areas. Among NewReno and CUBIC, in remote server mode, CUBIC can perform better, however, in the edge server one, the opposite is correct. In the urban deployment scenario, all of the TCPs can attain the same average cell goodput, but the RTT values for each one can be different significantly. Especially when we have NLoS or inside building UEs, loss-based TCPs suffer from higher latencies. Evaluations exhibit that for satisfying 5G requirements (i.e., goodput larger than 100 Mb/s and latency lower than 10 ms) in an urban deployment, only BBR can perform well in accompanying an edge server deployment and under desirable channel conditions [21]. Simulation results revealed that TCP generally could benefit from edge servers due to the shorter response time. However, in the edge server mode, CUBIC has the lowest throughput, as this value is for NewReno when remote servers are deployed. In addition to the location of servers, MSS size can have effects on the functionality of TCP, especially the loss-based ones. For example, by increasing the size of MSS, CUBIC gets more benefits. In contrast, it does not have any effects on the performance of BBR. Moreover, if big buffers are exploited, HighSpeed can reach higher performance at the cost of latency. As a result, Implementing HighSpeed needs using some techniques of AQM to reduce latency. In contrast, BBR prefers small buffers where the performance of HighSpeed will experience a reduction, but its latency will be improved.

In addition to urban and high-speed scenarios, the performance of TCP in indoor environments, like train stations, is another KPI to be analyzed. As a consequence, an analysis has been conducted in [55] to answer this question, evaluate the effect of the human body as a blocker of 5G mmWave communication, and how using TCP-FSO (Free-Space Optical), which is one of the candidates for long-distance high-speed wireless communications [69] can affect the performance. To attain this goal, an indoor train station scenario was simulated by using MATLAB 5G library. We should notice that, although TCP-FSO has some similarities to CUBIC, some modifications have been adapted, such as the retransmission has been improved, the congestion control mechanism has become delay-based ACK, and improved ACK retransmission control has been used. Thorough information about TCP-FSO can be found in [69]. In the first step, the effect of the human-body blockage was evaluated. It was assumed that passengers in a train station act as blockers in low, medium, and high-density environments. Some other obstacles, such as pillars and walls, could block communications too. Results showed that when the UE is close to the gNB, the number of obstacles, which indicates the number of blockages, had a minor negative impact on the performance. In the second step, the value of SNR was calculated at 28 GHz carrier frequency. For this, the actual channel at Haneda International Airport Terminal was observed. Results indicated that SNR could be degraded drastically by the blockages caused by human bodies. Moreover, the distance between a UE and gNB is another factor that can play a significant role in the quality of the received signal. The third step was the 5G network bandwidth calculation. The MATLAB 5G library was used to estimate the bandwidth of the 5G network downlink. Finally, the evaluation of TCP throughput was done by considering the 5G network bandwidth simulation results. To sum up, an overall look on the results reveals that TCP-FSO can reach higher throughput compared to CUBIC when exploited over 5G networks. Moreover, the number of blockers and the distance between a UE and base station have important impacts on the performance of TCP. When there are a few numbers of obstacles, and the distance is low, TCP can function more efficiently.

In addition to simulations, practical testings are paramount in achieving a clear view of a problem, which could happen after the implementation of 5G networks. As a result, one of the first practical evaluations of the commercial 5G mmWave networks was done in [54]. The test was conducted through the first world's commercial 5G in Chicago and Minneapolis provided by Verizon since April 2019. This network is operating at 28 GHz carrier with 400 MHz subcarriers. This evaluation was done at four different locations by downloading a TCP bulk to emulate different deployment scenarios of 5G. The results showed considerable enhancements for 5G compared to 4G in terms of throughput, which in some cases, it was ten times more than 4G. Although 5G has a much higher throughput, it showed many fluctuations, even in LoS connections. The reason behind these fluctuations is that different layers, such as the transport layer, are not ready to be deployed in 5G networks. In terms of RTT, 5G could not exhibit significant improvements, and it showed slight enhancements compared to 4G. It is because of the NSA (Non-Stand Alone) mode that makes most of the used infrastructure in 5G borrowed from 4G. It will be improved dramatically if SA (Stand Alone) mode is implemented in the coming future. Experiments in the presence of blockages such as human bodies, pillars, and trains showed that except for some thin materials like backpacks, cardboard boxes, or clear glass, most of them caused a drastic reduction in the performance of 5G.

The simulation results in [21], [55] and the practical testing output in [54] revealed that TCP's functionality could be impaired in several aspects including, throughput, latency, and congestion windows adjusting. As a consequence, for benefiting 5G mmWave full potential, some efforts should be made. These efforts can include wide rages from non-intelligence based schemes to complex intelligence-based algorithms in improving different aspects of TCP.

B. THROUGHPUT ENHANCEMENT

By the advent of 5G networks, the backbone traffic will increase intensely, and a need for a protocol to handle it efficiently is inevitable. One of the important efforts of designing a novel TCP for 5G networks was TCP Ohrid [63], which aims to improve the throughput by attaining 400 Gbps data

rates in the core network. The main purpose of TCP Ohrid is to manage the backhaul traffic to prevent collapse due to heavy congestion. The design of TCP Ohrid is based on High-Speed TCP, with this main difference that it strives to have different responses to different speeds. As a consequence, the behavior of TCP Ohrid is up to the current speed of the network. The goal of TCP Ohrid is to achieve at least 5 Gbps for mobile users under heavy mobility and 400 Gbps data rate for the backhaul. Because of the different mechanisms of TCP Ohrid in approaching congestion in a network, it can reach to larger congestion window size compared to NewReno. However, the results revealed that it could not outperform High-Speed TCP in terms of the congestion window, so it means that HighSpeed TCP can reach larger cwnd sizes compared to TCP Ohrid. The principal advantage of TCP Ohrid over HighSpeed is being more friendly to existing protocols and achieving comparable data rates to HighSpeed TCP, which makes this protocol suitable for being deployed in mobile communication and backhaul transmission [63].

When a catastrophic circumstance occurs, the necessity for establishing an instant wireless communication to transmit current videos in order to evaluate the conditions on the site is inevitable. The best candidate to be deployed in these situations is 5G mmWave due to its high bandwidth and extremely low data latency. However, collapsed buildings, broken trees, and other obstacles prevent 5G mmWave from functioning adequately. For that, another attempt to modify TCP in order to have a new scheme that is suitable for disastrous situations called DL-TCP (Deep-Learning TCP), was proposed in [19]. The main effort of TCP Ohrid was improving 5G network performance by mainly increasing the backhaul data rates, then the fronthaul by a less priority. However, DL-TCP aims to improve the fronthaul functionality by adjusting the congestion window (cwnd) size efficiently during disconnection occurrences in the network caused by blockages or misalignments and prevents the sending rate from being initialized wrongly in disastrous situations.

In DL-TCP, an ML (Machine Learning) framework was developed to put a threshold between RTOs caused by congestion and the ones created by the blockage and misalignment. DL-TCP employs some parameters to divide the network into three parts, long-time failure, short-time failure, and congestion. When long-time failures happen, the network is going through a long time disconnection. In such a case, DL-TCP uninitializes the cwnd size and prevents the resetting process of the sending rate. When short-time failures occur in the network, they are indications for short interruptions, so the algorithm intends to maintain the cwnd size and retransmits the most recent transmitted packets. The goal of this process is retransmitting recently lost packets caused by short disconnections. When congestion is detected, the algorithm decreases the cwnd size and enters the steady-state to work in its normal way.

The used parameters in the deep neural network for DL-TCP for estimating the mentioned states are: ''time'' that is the time of used SNRs (Signal to Noise Ratio), ''location'' which is the location information of the TCP sender, ''velocity,'' which is the current speed of the TCP sender, and ''SNR,'' which is the SNR value received by the TCP sender and shows the signal quality. Authors in [19] evaluated the performance by means of simulations in two different scenarios (small and big obstacles) and two different mobility modes.

The simulation results showed that the proposed TCP could outperform other TCPs. In terms of RTT, DL-TCP, NewReno, and CUBIC are similar, but BBR has high RTT values compared to the others. Comparing different cwnd sizes indicates that all of the protocols are experiencing intense fluctuations because of the intermittent nature of the channels. However, DL-TCP prevents cwnd initializing during the interruptions and lowers it in case of a congestion event and can help to adjust cwnd size more efficiently [19].

C. LATENCY AND FAIRNESS

Besides throughput, latency and fairness are two other KPIs, which need to be improved in order to adapt TCP to 5G mmWave networks. Latency is one of the critical features in 3gpp specifications for 5G networks, which pursues considerably low values close to zero. By improving latency, fairness will automatically be enhanced due to its direct correlation to latency because shorter latencies lead to faster paces in increasing the sending rate, so senders with shorter latencies can reach larger sending rates compared to the ones with higher values. As a consequence, having a fair network could be hard in situations that latencies differ drastically.

A simulation analysis of TCP over 5G networks to see the impact of parameters such as RLC buffer size and RTO on the functionality of TCP has been made in [23]. RCL buffer size can have a significant impact on the latency and throughput by having the capability of masking losses to higher layers, especially the transport layer. At first, the effect of the RLC buffer size on the performance of higher layer protocols was analyzed, and the results indicated that exploiting buffers at the size of 1 MB, which was enough in the previous 3gpp mobile networks, is not good enough for 5G networks. As a result, they suggest deploying 7 MB buffer size with an RTO of 200 ms to replace the conventional 1 second to improve the functionality of TCP.

In order to analyze multiple-flows, different UEs using various applications were analyzed. In such a case, one of the UEs is generating the heaviest traffic and moving around the environment to trigger the handover and affects the other UEs' performances. In this case, both blockage and long flows can exist at the same time. Results showed that most of the UEs have fewer retransmitted packets in YeAH compared to CUBIC, and generally, multiple flows can perform better when YeAH is deployed. However, when a UE is not affected by a heavy flow and is served by one gNB in the entire period (i.e., no handover is triggered), the retransmissions number is much less for CUBIC compared to YeAH. On the other hand, when a UE is served by several gNBs, it can affect CUBIC more than YeAH. For static users, the performances

are similar, but the number they need for retransmissions is different. When different flows exist together, CUBIC will retransmit more, but it can reach higher performance.

From the buffer using point, when medium-size flows are not affected by long ones, both protocols use the same buffer size, and ARQ (Automatic Repeat reQuest) at the MAC layer can mask packet losses caused by wireless errors to the transport layer. However, when a long flow exists, CUBIC deploys more buffers than YeAH and can attain a higher rate. Protocols like CUBIC that try to utilize the capacity of the link and have quick recovery mode perform well during NLoS disconnections but not very well when long NLoS disconnections exist. On the other hand, protocols with a hybrid mechanism like YeAH (which uses packet loss and RTTs) have fewer performance variations. Moreover, A comparison between throughput and RTT for different TCPs can be found in [23], which indicates different reactions of each TCP to various delays.

As the authors in [23] suggested some simple mechanisms such as modifying RLC buffer size and RTO value, then tried to analyze TCP functionality, the authors in [52] strived to enhance latency and fairness by leveraging sophisticated and straightforward schemes.

They sought the root of the problem in quick paces of buffers fillings in NLoS states when queue sizes are large. In contrast, deploying a small buffer leads to an underutilization of TCP performance. The base suggested solution to handle the problem is deploying AQM techniques such as CoDel [56] and Fq-CoDel [57], which drop packets before the queue is full. However, these techniques need some modifications in order to work appropriately in 5G networks. For tackling the fairness issue, the first choice is exploiting Fq-Codel, which behaves each flow differently in queuing, and tries to maintain fairness among them. However, this technique is not able to perform properly in 5G mmWave networks. This malfunction of Fq-CoDel in 5G networks is due to the harsh effects of NLoS disconnections, especially long failures during static conditions. Moreover, dropping many packets during a NLoS period by AQM techniques forces TCP to enter the fast retransmit or the slow start phase, then after switching to LoS, it takes a long time for TCP to gain the possible high performance.

The first proposed solution in [52] is called on-off. In this case, when the network goes through a NLoS situation, CoDel and Fq-Codel will be disabled and are not able to drop packets. This approach prevents massive packet drops throughout the NLoS period. This can be achieved by setting the target parameter to five seconds to mimic a disabled state. The second scheme, which is more complicated than the first one, can perform even better. In this case, the RTT for each flow needs to be estimated, then based on the estimated values, the target parameter for each flow is calculated and used. Results showed that better fairness can be achieved in the second approach compared to the on-off one. Using CoDel + on-off, Fq-CoDel + on-off, and tuning (i.e., the RTT estimation approach) in accompanying with NewReno and CUBIC could lead to almost constant fairness during different LoS/NLoS conditions even when the distance between the UE and gNB increases. Especially, Fq-CoDel + on-off exhibits nearly the same fairness independent of NLoS time.

In addition to fairness, exploiting these approaches can affect the value of the delay parameter, and when CoDel + on-off mode is deployed, the average delays for different flows can be almost the same. However, the delay values for the three approaches are different, and in most cases, the best value is for Fq-CoDel + tuning. By deploying the three suggested schemes in [52], fairness can be improved at the cost of 10 ms of more delay. However, when Fq-CoDel + tuning is used, this number can be reduced to 5 ms by negatively affecting the fairness.

The principal cause for improving fairness is that it is one of the ultimate goals of TCP congestion control algorithms that is desired to be obtained along with high throughput while preventing congestion in the network. In order to increase the performance of networks, buffers are used to prevent dropping the packets that are experiencing short-lived traffic peaks. In general, buffers suffer from two drawbacks, a weakness in managing the queues and TCP congestion control failure, which both can lead to higher latencies and underutilizing the available bandwidth of the network. Moreover, full buffers in a network which are reasons for higher latencies, end up in bufferbloat problem, one of the most significant issues in deploying buffers. This problem can be intense when we know a tremendous number of buffers have been installed throughout the Internet without having efficient strategies in controlling the queues. These buffers can degrade the performance of TCP and be hurdles on the way of this protocol in accomplishing its aims [64], [65]. The existing techniques encounter some challenges in 5G networks and need to be renovated to adapt to these networks.

On the one hand, the proposed solutions in [23], [52] can enhance latency and fairness to some levels. On the other hand, they are not adequate enough to meet 5G mmWave networks desired values. As a consequence, some advanced algorithms should be proposed. One solution can be giving intelligence to AQM techniques by using ML approaches to make the dropping mechanism more effective, so that they can handle the issues caused by buffer size and packet dropping in the queues more accurately. In such a case, the AQM techniques' static mechanism will be modified and replaced with smart schemes, so they will look at the existing parameters, and based on them, decide to drop a packet or not. The ultimate goal of new algorithms can be predicting the behavior of a network and drop beforehand in order to provide a tradeoff between throughput and latency.

One of the most promising ML algorithms that can be convenient in redesigning AQM techniques might be RNN (Recurrent Neural Network) because it can use prior states and is capable of predicting the behavioral patterns based on past functionalities, which might be useful in controlling packet drops in a buffer. The reason is that queue controlling can be a sequential action in its nature, and exploiting

feedback from the previous results can be beneficial. Moreover, we assume that having some information from the previous moments can be useful in making decisions to drop a packet or not. In this case, the exploitation of memory in RNN or LSTM (Long Short-Term Memory) type of RNN can assist the dropping algorithm in the estimating of the loss probability. When deploying these techniques, the way of producing data, training time, length of history, and other hyper parameters like learning rate and the number of layers as well as the memory, i.e., the feedback of the prior outputs should be chosen by elaborate analysis and testing. In addition to more in-depth analysis, LSTM, which is an architecture of RNN in the field of deep learning and developed to overcome the gradient vanishing problem, can be used to increase the memory and buffer of the network. This algorithm is one of the appropriate techniques for classification and regression problems. As a result, by deploying these techniques, especially the LSTM, coming packets can be predicted to be dropped or not. However, these schemes need more analysis in order to evaluate their efficiency.

Another technique can be bringing the cloud close to UEs, which is called fog networking [66], [67]. In this case, a node such as APs, small cells, or routers can be a fog node in which provides services to other UEs. One of the most critical questions in fog networking is the location of the fog nodes (i.e., which nodes are the best candidates to be selected as the fog nodes), especially in heterogeneous networks, in the combination of HPNs (High Power Nodes) with LPNs (Low Power Nodes), where some LPNs are selected to be upgraded to fog nodes in order to improve the performance of the network. One way could be dividing the nodes into some clusters, then choose a node in each group as the leader. ML techniques can be beneficial to be exploited in order to reach this purpose. As a result, an unsupervised ML approach was proposed in [68] to answer the central question in fog networking, which LPNs should be upgraded to become fog nodes. This algorithm is based on an unsupervised soft clustering machine learning. In this case, all of the LPNs are divided into separate groups, and then in each group, a node is selected as the head of the group. After that, all of the heads turn into fog nodes. One of the ultimates of this approach is improving the k-means hard clustering, in which each node chooses the corresponding fog node based on the closest Euclidean distance. This approach, which is executed by deploying the Voronoi Tessellation model, can lead to a poor channel connection because there is no guarantee that the channel between the node and the closest fog node has the best quality. As a result, performance and latency will be degraded. By using an unsupervised ML approach, the proposed algorithm is able to enhance the latency, which is one of the most critical issues on the way of deploying TCP over 5G networks.

D. MULTI-FLOWS VERSUS A SINGLE FLOW

Because new cellular devices intend to use several interfaces, deploying MP-TCP (Multi-Path TCP) [70] can have some advantages compared to other TCPs. The key feature of MP-TCP is in its capability with multipath communication, which means when a socket establishes TCP connections, it can handle more than one interface related to different applications. These interfaces can incorporate various types of communications, such as Wi-Fi, a cellular network, and an Ethernet connection. The significant feature of MP-TCP that makes it different from conventional TCPs is how it handles the cwnd size in different subflows, which can be coupled or uncoupled. When deploying the latter one, each subflow is treated independently, and cwnd sizes for them are adjusted separately. In contrast, in the coupled mode, all of the cwnds are adjusted in a correlated way. As a result, the congestion control algorithm of MP-TCP includes two different approaches. The ultimate goal of this separation is an attempt to achieve the main purposes of MP-TCP, which are: 1) the minimum performance of MP-TCP should be as good as a single-path TCP; 2) the deployed resources by MP-TCP should not be more than conventional TCPs; and 3) it should be capable of navigating more packets to uncongested paths. By considering the aims, deploying MP-TCP can be one of the solutions for having a trade-off between throughput and latency.

The analysis of MP-TCP over 5G and LTE networks was done in [71]. The simulation results showed that MP-TCP could outperform SP-TCP (Single-Path TCP) about 30-40 percent in the LoS conditions when deploying in 5G networks. However, it is not true when 5G and LTE coexist together. We should consider that in contrast to 5G mmWave, in LTE, the distance between a UE and a gNB is not a key factor. Thus, in higher distances, MP-TCP can perform better in the coexistence of LTE and mmWave compared to the time it is deployed only in 5G mmWave networks. Moreover, Simulation results revealed that by increasing the distance between a UE and gNB, the value of latency gets higher.

It is worth mentioning that MP-TCP with a coupled congestion control mechanism showed a poor performance compared to CUBIC and cannot fulfill the first goal of the MP-TCP. The cause behind it is that in the congestion control process, MP-TCP assumes mmWave as a congested path due to its high loss probability and tries to transmit packets through the LTE links. In contrast, this issue does not exist in the uncoupled mode. Another problem is that when the uncoupled MP-TCP coexists with SP-TCP, due to the unfriendly nature of them, it leads to an unfair network. All the mentioned problems indicate that for designing an MP-TCP in order to fulfill all the goals, more efforts need to be made.

VI. SIMULATION ENVIRONMENTS AND USE CASES

To evaluate the design of a new protocol or improve a specific part of a network, we need to perform detailed analysis, which can be achieved in real scenario tests, analytics models, or simulation environments. In this section, we focus on different simulating tools that can be chosen based on

the fundamental necessities of a research on TCP over 5G networks.

One of the popular simulating tools is called LENA [72], an LTE-EPC network simulator. LENA is an open-source product-oriented LTE/EPC network simulator that supports LTE small/macrocell vendors to create and test Self Organized Network (SON) algorithm applications. Target applications for LENA include the design and performance evaluation of DownLink and UpLink Schedulers, Radio Resource Management algorithms, Inter-cell Interference Coordination solutions, Load Balancing and Mobility Management, Heterogeneous Networks (HetNets) solutions, end-to-end QoE (Quality of Service) provisioning, Multi-RAT network solutions, and Cognitive LTE systems. LENA is based on the well-known ns-3 [73] discrete-event network simulator. LENA simulator covers different layers such as RRC (Radio Resource Control), PDCP (Packet Data Convergence Protocol), RLC, physical layer, MAC layer, different channels, and antenna models.

Another powerful open-source tool for simulating 5G mmWave networks, which has been built up on LENA [72], [74] and is an extension of ns-3, is ns-3 mmWave [75], [76]. The channel model implementation was proposed in [77], and the dual connectivity was explained in [78], [79]. A comprehensive description of ns-3 mmWave is available in [80]. This module is a powerful and accessible tool that can simulate various aspects of 5G mmWave, such as the corresponding layers, and channels defined in 3GPP specifications. One of the main aspects of this module is its availability of connecting it to a Direct Code Execution [81], so the Linux stack TCP/IP can be run as the TCP/IP stack of ns-3 nodes. Moreover, it has a wide range of selection from 6-100 GHz channels, which is the official 3GGP channel model [82].

Another tool for simulating 5G mmWave networks is using a library supported by MATLAB [83], which can provide the main features of 5G networks. It includes standard-compliance functions and reference examples that can be used in order to model, simulate, and verify 5G communication systems. By using this toolbox, configuration, simulation, measurement, and analysis of an end-to-end connection is available.

There is another tool for simulating 5G networks, which is called K-SimNet proposed by Seoul National University [84]. It is an extension of ns-3 that can support 5G NR, 5G core, multi-RAT protocols, traffic management on multi-connectivity, SDN/NFV, and other features of 5G, which make it capable of simulating 5G end-to-end networks.

To sum up, ns-3-mmWave is the primary candidate and the most used simulator to evaluate a 5G network. This ns-3 based software includes the majority of the 3GPP features such as channel modeling, supporting dual connectivity, and more. Moreover, being around for a long time is the compelling aspect of the ns series.

VII. OPEN ISSUES IN MACHINE LEARNING-BASED APPROACHES

One of the most convenient ways of enhancing TCP for 5G networks and providing intelligence is using ML techniques. ML techniques are generally categorized into supervised, unsupervised, and reinforcement learning. Commonly, the exploited methods for modifying TCP can be categorized into two groups, LP-TCPs (Loss Predictor TCPs) and RL-TCPs (Reinforcement TCPs) [85]. LP-TCP can use different ML techniques to learn the network status and then predict the behavior of the network. Most of the LP-TCPs use supervised algorithms in order to attain their goals. In the following, there are some LP-TCP and RL-TCP proposals that can be deployed as guidelines to solve the performance degradation of TCP over 5G networks.

A. NEURAL NETWORKS, DEEP NEURAL NETWORKS, AND REINFORCEMENT LEARNING PROPOSALS

Neural networks and deep neural networks [86] are state-of-the-art techniques for many applications and can model non-linear and sophisticated systems. Applied to TCP congestion control mechanisms and optimization parameters, the goal of using these techniques is to classify a network into different categories, then control the congestion based on the state of the network deducted from different deployment parameters. For example, in order to enhance the congestion control mechanism, a network can be grouped into five states, including long disconnections, short disconnections, congestion, normal, i.e., the network is neither congested nor utilized, and empty, i.e., low capacity of the network is used. Afterward, the cwnd size can be adjusted based on the state in which the network is performing in. These techniques can give flexibility to the network; thus, the protocol can adapt itself to the existing situation based on the current output of the ML engine. Dividing to five states is one of the ways of classifying the network, and other approaches can be dividing it into more or fewer parts based on the aims of the scheme. As a consequence, the models for organizing the network can be diverse. For example, instead of five states, we can have three ones incorporating short disconnections, moderate disconnections, and long disconnections based on the duration of the disconnections, and then behave by considering the current state.

To sum up, neural networks and deep neural networks algorithms can be promising candidates to be deployed in order to adapt TCP to 5G mmWave networks and mitigate the adverse effects. Obligating TCP to function in more intelligent approaches enhances its functionality in making congestion control decisions. As a result, TCP can react appropriately to different circumstances, which leads to performance improvement.

When a topology change occurs in the network, loss predictor TCPs have a drawback and cannot function autonomously. In this case, training over is essential, and makes it impossible for these algorithms to function

autonomously. Retraining the ML engine over and over can exhaust the resources and reduces the functionalities of new algorithms. This problem can be solved by using an ML technique called Reinforcement Learning (RL). The main question is that, is it possible for a TCP to adaptively learn and respond in a topology that can experience changes in its parameters [85]?

RL has an agent that iteratively interacts with its environment to learn the state and select an action to change the state. After that, the agent gets feedback, and based on this feedback, tries to find the best sequence of actions or policies to maximize a cumulative reward. If the system performs well, the action is rewarded; otherwise, it is penalized. As a result, the training set in RL includes a set of state-action pairs and rewards or penalties. DQN (Deep Q-Network) is a common way in which a deep learning model is created to find the actions and an agent that can result in the best rewards.

This technique is suitable for problems such as decision making, planning, and scheduling tasks, so if we want a TCP to react to the network changes in a dynamic way, RL algorithms can be deployed in the network to modify the TCP to make it suitable over 5G networks. By exploiting this technique, TCP can learn the network and will have a proper reaction to problems such as blockages and can enhance the performance of the network. There is a successful implementation of reinforcement learning for TCP in [85], and it may be possible to adapt this learning algorithm for 5G mmWave networks. The proposed approach in [85], which is called RL-TCP, tries to improve the performance of TCP over a dumbbell topology. They employ three components in the network called the sensing engine, the learner, and the actuator. The sensing engine is responsible for updating a length-5 state vector, which is given to the learner after the individual received ACKs by the sensing engine. Then the actuator decides the action, which is adjusting cwnd based on the output of the learner. The primary responsibility of the learner is to learn how good or bad is the current action at the present state, which is called Q(s, a) function. In order to learn Q-function, SARSA (State–Action–Reward–State–Action), which is an on-policy learning algorithm, was exploited. The simulation results showed improvements in terms of throughput and delay. Moreover, a higher sending rate could be achieved by using RL-TCP. However, in terms of fairness, it could not compete NewReno. It should be considered that in a dynamic environment with many changes, RL-TCP may lose its optimal functionality. For adapting this technique to mmWave networks, more detailed analysis and experiments are needed.

B. RANDOM FOREST ALGORITHM

Random forests or random decision forests [87] is a learning algorithm suitable for classification, regression, and other kinds of problems that can be handled by making a multitude of decision trees. This algorithm creates some trees for making decisions on new data. These trees use data samples in order to be trained. After the training, for making predictions on new data, the outputs from all trees are gathered and based on the voting, i.e., the output of each tree, the best solution is chosen. The most significant power of this algorithm is in the classification model. As a result, it can be employed to model a network and divide it into different categories, and then, based on the current state, future states can be predicted. In 5G networks, one of the parameters that can be exploited to determine the network's behavior is blockage time, which can be divided into different categories. It means different blockers such as human bodies, buses, and buildings can be distinguished, and therefore, a TCP that behaves differently in each case can be implemented. As an example, obstacles can be classified based on their sizes to different groups, such as small, average, and big ones.

Related to congestion control in TCP, when a human body is predicted, TCP can retransmit the recent packets and then increases the cwnd size when the UE is moving or keep the cwnd size fixed when the UE is still. This mechanism can be adapted for all of the small obstacles to handle short disconnections. When there are obstacles with average sizes such as buses and cars, cwnd size can be maintained fixed during the blockage time without any changes by resending the recently sent packets. For big obstacles such as buildings, RTO can be triggered but not by changing the size of cwnd to one. By having these three categories, TCP can have a suitable strategy in encountering disconnections and reacting to failures.

Finally, when the congestion is detected in the network, TCP sets the cwnd size to one and enters the slow start phase, or instead, a new phase can be designed from scratch. The goal of designing a new phase can be preventing the sending rate from being initialized and promoting the network to utilize the high bandwidth of 5G networks. In conventional TCPs, when RTO is triggered, cwnd is set to one and starts over. This approach may not be suitable for 5G networks because the channel bandwidth is large, and they can support high data rates. As a result, when TCP sets cwnd to one and starts over, it can take a long time to ramp up the high potential of 5G networks.

Moreover, because of high packet loss probability in 5G mmWave networks, traditional loss detection strategies cannot perform well in 5G networks, and it is not wise to have a TCP, which detecting a lost packet only based on three received duplicate ACKs or triggered RTO. For the 5G network, deploying new and sophisticated strategies that are not working only based on packet losses is inevitable. It seems that it is time for new players such as RTT, the time interval between two drops, cwnd sizes before packet drops, and instantaneous throughput to be exploited in new TCPs in order to have robust and intelligent mechanisms in tackling various issues in 5G networks. As a result, it is convenient to replace loss-based TCPs with novel intelligent model-based TCPs. Random Forest algorithm can be one of the candidates to be exploited in designing new approaches for congestion detection because of its power in classification problems.

In addition to the new congestion detection approaches, in designing new protocols, when the RTO is triggered, cwnd values and the increasing pace can also be defined based on the output of the ML engine, and instead of setting it to one, new values can be defined. This approach can save much time in reaching the high potential of 5G networks.

Another problem of the slow start phase is that the used pace for increasing the sending rate is inefficient for 5G networks. As a consequence, when the RTO is triggered and the protocol enters the slow start phase, in addition to new cwnd size, a novel scheme for increasing the cwnd size can be deployed too. This approach can be as simple as increasing the initial window size and pacing based on the slow start threshold or as sophisticated as functioning based on the output of the ML engine. As a result, using the Random Forest algorithm can be one of the most convenient techniques in dealing with the blockage and misalignment problems in 5G networks.

C. BEING MORE FAIR

For tackling the fairness issue, the protocol should be intelligent enough to predict the variations and prevent the network from being unfair. Different ML techniques, such as neural networks and deep neural networks, can model the networks to handle this problem. In this case, ML techniques with the capability of dividing the network by deploying efficient ways can be appropriate.

D. RETRANSMISSION TIME-OUT AND THE POSSIBLE SOLUTIONS

The value of RTO in 5G networks and how it is adjusted can be a sensitive procedure. Both small and large values for RTO can harm the functionality of TCP over 5G networks, so this value should be set carefully.

One solution can be deploying classification ML techniques to predict whether the coming RTO is due to issues such as blockage, misalignment, or congestion in the network. In this case, if the RTO is triggered by a blockage, the size of the cwnd will be maintained fixed, and uninitializing process will be prevented by resending the recently sent packets. As a consequence, an enhancement on the performance of TCP over 5G networks can be achieved by maintaining the desired cwnd size. If we can distinguish RTOs caused by blockages, misalignments, or congestion, a significant improvement will be made in the performance of TCP over 5G networks. Neural Networks and Deep Neural Networks, which are able to classify a system in an efficient way, can be the leading ML techniques to be deployed in order to handle this issue. In this case, RTT, throughput, and SINR values can be nominations of being deployed as the training set accompanying other parameters such as positions and velocities of the UEs in order to train an ML algorithm. By deploying these techniques, RTOs can be classified as blockage and congestion based ones, and then based on the output, proper decisions will be made.

In this section, deploying machine learning techniques to improve TCP's functionality was analyzed. However, other approaches can be used to enhance 5G networks performance by improving the functionality of the other layers of the protocol stack. One of the appealing schemes is improving the physical layer functionality by techniques such as network densification, especially by means of microcells that are one of the enablers for the network densification. Network densification, which is one of the undemanding approaches for service providers, can be implemented by increasing the number of cell sites in order to enlarge the available capacity. This technique can be useful in areas that the available spectrum can be drained by the high number of UEs such as dense urban areas. The most important advantage of this technique is the straightforward and quick implementation that can boost the available services wherever needed. Moreover, mmWave small area coverage can be compensated for by using this technique.

In contrast, being costly is the foremost hurdle on the deployment path of coverage area expansion methods. There have been some efforts to use this technique more efficiently, such as [88]. This paper aims to deploy a large number of base stations by exploiting ultra-densification in order to alleviate the adverse effect of mmWave. Using ultra-densification can improve the performance of the network at the cost of co-channel interferences. It means in this case, two or more devices might be competing on utilizing the same frequency, which can increase the waiting time by the increment of the number of devices and degrade the functionality. The reason is the queue creation for devices in deploying the same channel. The main upside of the scheme is in lowering the access distance and boosting the choice of base stations for individual UEs. One of the most critical considerations of network densification should be the geographical distribution of users and the location of blockers, which can affect the placement of base stations. The simulation results showed that deploying base stations in a large dense and amorphous way can reduce the negative effect of mmWave networks. However, another issue called co-channel interference needs to be handled in this situation.

Although network densification deployment has some advantages in enhancing 5G mmWave network functionality, we should notice that having many gNBs can trigger handover in dense areas and it can have its own defects on the performance of TCP. The reason is that existing many handovers can confuse TCP in attaining an optimal functionality in achieving the highest available throughput through a fair network. Moreover, Synchronizing a large number of base stations and finding the optimal number for them are other issues that need to be addressed. As a result, other schemes such as cross-layer designs might be analyzed as a prominent candidate to solve the performance reduction of 5G mmWave networks. Cross-layers schemes can be obtained by sharing the information between the different layers of the protocol stack. The main reason for deploying cross-layer approaches is that the different layers have been designed for

wired networks. However, this assumption cannot be correct for wireless networks, in which the concept of the link is entirely different from the one in wired networks. In this case, the physical layer and transport layer collaboration can be beneficial in making decisions and adjusting the sending rate. However, the detailed analysis of network densification and cross-layer design is out of this paper's scope and can be investigated in future work.

To sum up, based on the intended aims, different schemes can be deployed to improve the performance of 5G mmWave network. One can be improving the transport layer protocols functionality by means of various techniques such as machine learning algorithms. Another can be improving link-level directional beamforming schemes, which its focus is on amplifying the transmission power. Some physical layer solutions such as network densification, which aims to solve the limited coverage problem of mmWave by exploiting many base stations in order to expand the connectivity area can be another candidate. In addition to solutions that focus on a particular level, cross-layer strategies can provide an environment for collaboration of different layers and exchanging information between them in order to make efficient decisions. Each one of these solutions has advantages and disadvantages and should be chosen based on the necessities. Modifying and changing the transport layers can lead to issues such as unfair networks or bufferbloat. Moreover, co-working the new protocol and the previous ones can be an important question to be answered. Deploying network densification can be expensive, lead to the creation of co-channel interference, trigger frequent handovers in the network, and needs base stations synchronization [88]. Cross-layer implementation lacks a clear and comprehensive architecture, are hard to design and implement, establishing a framework for the collaboration of different layers needs elaborate efforts, and without detail analysis and experimentations, can lead to chaos [89].

VIII. CONCLUSION

The different layers of the protocol stack, mainly the widely used transport protocol TCP, encounter new issues when deployed in 5G networks, especially along with higher frequencies such as mmWave. The main challenge of TCP is due to the intermittent nature of mmWave channels, which are sensitive to blockage and misalignment. These problems cause fluctuations in the functionality of the congestion control mechanisms of different TCP variants, which lead to degradation of the measuring factors, including throughput, latency, and fairness. In this paper, we have done a thorough review of 5G technology and its different aspects, TCP and its functionality over 5G networks, and the analysis of if it is better replacing this protocol with novel ones in the coming future or adapting it. The conclusions indicate that the 5G network is a promising telecommunication infrastructure that will revolute various aspects of communication. However, different parts of the Internet, such as its regulations and protocol stack, will face new challenges, which need to be solved in order to exploit 5G to its full potential, and without smart rules and intelligent protocols, the high bandwidth of 5G, especially 5G mmWave will be wasted. Moreover, existing challenges, new areas of research, and novel proposals as the guidelines for future work have been discussed in detail, which can show key directions for researchers. Finally, the investigations revealed that providing intelligence for TCP can be one of the fundamental pillars of establishing a new protocol to function adequately in 5G networks. This smartness can be attained by deploying machine learning techniques. The type of exploited technique can be diverse based on the issue that is needed to be addressed. Supervised classification techniques such as Neural Networks, Deep Neural Networks, and Random Forest Algorithm can be beneficial in categorizing a network into different working areas when a sufficient number of training set parameters are available. However, when being autonomous is crucial, Reinforcement Learning can be noteworthy. Besides, some questions are needed to be answered to improve the functionality of TCP over 5G networks:

- Is the time arrived to have new congestion control techniques based on new parameters?
- Is it necessary to modify some conventional aspects of TCP, such as initial congestion windows or retransmission time-out?
- To what extent exploiting a new protocol such as QUIC can improve the functionality of the transport layer if replaced with the conventional ones, i.e., TCP and UDP?
- Can changing the default value of the MTU and adapting it to larger MSS sizes be beneficial in increasing the performance of TCP over 5G networks?
- To what extent deploying machine learning approaches in 5G procedures and parameters can improve the functionality of the network?

REFERENCES

[1] Several Authors, Fredrik Jejdling, Ericsson. (Jun. 2020). *Ericsson Mobility Report*. [Online]. Available: https://www.ericsson.com/49da93/assets/local/mobility-report/documents/2020/june2020-ericsson-mobility-report.pdf

[2] Several Authors. (Apr. 2016). *5G PPP Use Cases and Performance Evaluation Models, Version 1.0*. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-use-cases-and-performance-evaluation-modeling_v1.0.pdf

[3] Several Authors. *5G Network Support of Vertical Industries in the 5G Public-Private Partnership Ecosystem*. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2020/03/5PPP_VTF_brochure_v2.1.pdf

[4] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The Next Generation Wireless Access Technology*, 1st ed. Amsterdam, The Netherlands: Elsevier, Aug. 2018.

[5] A. Ghosh, A. Maeder, M. Baker, and D. Chandramouli, "5G evolution: A view on 5G cellular technology beyond 3GPP release 15," *IEEE Access*, vol. 7, pp. 127639–127651, Sep. 2019, doi: 10.1109/ACCESS.2019.2939938.

[6] Several Authors, GSMA, London, U.K. (Jun. 2019). *NB-IoT Deployment Guide to Basic Feature Set Requirements*. [Online]. Available: https://www.gsma.com/iot/wp-content/uploads/2019/07/201906-GSMA-NB-IoT-Deployment-Guide-v3.pdf

[7] *Study on Scenarios and Requirements for Next Generation Access Technologies, V14.2.0*, document 3GPP, TR 38.913, Sophia Antipolis, France, 2017. [Online]. Available: https://www.etsi.org/deliver/etsi_tr/138900_138999/138913/14.02.00_60/tr_138913v140200p.pdf

[8] V. P. Kafle, Y. Fukushima, P. Martinez-Julia, and T. Miyazawa, "Consideration on automation of 5G network slicing with machine learning," in *Proc. ITU Kaleidoscope, Mach. Learn. 5G Future (ITU K)*, Santa Fe, Argentina, Nov. 2018, pp. 1–8, doi: 10.23919/ITU-WT.2018.8597639.

[9] *IMT Vision-Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond*, document ITU-R M.2083-0, M. Series 2015.

[10] *User Equipment (UE) Radio Transmission and Reception; Part 1: Range 1 Standalone, V15.5.0*, document TS 38.101-1, 3GPP, Sophia Antipolis, France, 2019.

[11] *User Equipment (UE) Radio Transmission and Reception; Part 2: Range 2 Standalone, V15.5.0*, document TS 38.101-2, 3GPP, Sophia Antipolis, France, 2019.

[12] GSMA, London, U.K. (Apr. 2018). *Road to 5G: Introduction and Migration*. [Online]. Available: https://www.gsma.com/futurenetworks/wp-content/uploads/2018/04/Road-to-5G-Introduction-and-Migration_FINAL.pdf

[13] D. Sattar and A. Matrawy, "Optimal slice allocation in 5G core networks," *IEEE Netw. Lett.*, vol. 1, no. 2, pp. 48–51, Jun. 2019, doi: 10.1109/LNET.2019.2908351.

[14] C. V. Murudkar and R. D. Gitlin, "Optimal-capacity, shortest path routing in self-organizing 5G networks using machine learning," in *Proc. IEEE 20th Wireless Microw. Technol. Conf. (WAMICON)*, Cocoa Beach, FL, USA, Apr. 2019, pp. 1–5, doi: 10.1109/WAMICON.2019.8765434.

[15] K. Katsaros and M. Dianati, "Evolution of Vehicular Communications within the Context of 5G Systems," in *Enabling 5G Communication Systems to Support Vertical Industries*. Hoboken, NJ, USA: Wiley, 2019, pp. 103–126, doi: 10.1002/9781119515579.ch5.

[16] A. Tzanakaki, M. P. Anastasopoulos, and D. Simeonidou, "Converged optical, wireless, and data center network infrastructures for 5G services," *J. Opt. Commun. Netw.*, vol. 11, no. 2, p. A111, Feb. 2019, doi: 10.1364/JOCN.11.00A111.

[17] Q. Liu, R. Liu, Z. Wang, and Y. Zhang, "Simulation and analysis of device positioning in 5G ultra-dense network," in *Proc. 15th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Tangier, Morocco, Jun. 2019, pp. 1529–1533, doi: 10.1109/IWCMC.2019.8766743.

[18] S. Martiradonna, A. Grassi, G. Piro, L. Grieco, and G. Boggia, "An open source platform for exploring NB-IoT system performance," in *Proc. IEEE Eur. Wireless Conf. (EW)*, Catania, Italy, May 2018, pp. 174–179.

[19] W. Na, B. Bae, S. Cho, and N. Kim, "DL-TCP: Deep learning-based transmission control protocol for disaster 5G mmWave networks," *IEEE Access*, vol. 7, pp. 145134–145144, Oct. 2019, doi: 10.1109/ACCESS.2019.2945582.

[20] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmWave frequencies," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 173–196, 1st Quart., 2019, doi: 10.1109/COMST.2018.2869411.

[21] M. Zhang, M. Polese, M. Mezzavilla, J. Zhu, S. Rangan, S. Panwar, and M. Zorzi, "Will TCP work in mmWave 5G cellular networks?" *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 65–71, Jan. 2019, doi: 10.1109/MCOM.2018.1701370.

[22] A. Afanasyev, N. Tilley, P. Reiher, and L. Kleinrock, "Host-to-Host congestion control for TCP," *IEEE Commun. Surveys Tuts.*, vol. 12, no. 3, pp. 304–342, 3rd Quart., 2010, doi: 10.1109/SURV.2010.042710.00114.

[23] P. J. Mateo, C. Fiandrino, and J. Widmer, "Analysis of TCP performance in 5G mm-wave mobile networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Shanghai, China, May 2019, pp. 1–7, doi: 10.1109/ICC.2019.8761718.

[24] J. Postel, "Transmission control protocol," document STD 7, RFC 793, RFC1122, RFC 3168, RFC 6093, RFC 6528, RFC 793, Sep. 1981. [Online]. Available: https://tools.ietf.org/html/rfc793, doi: 10.17487/RFC0793.

[25] M. Allman, V. Paxson, and E. Blanton, *TCP Congestion Control*, document RFC 5681, Sep. 2009. [Online]. Available: https://tools.ietf.org/html/rfc5681

[26] S. Ha, I. Rhee, and L. Xu, "CUBIC: A new TCP-friendly high-speed TCP variant," *ACM SIGOPS Oper. Syst. Rev.*, vol. 42, no. 5, pp. 64–74, Jul. 2008, doi: 10.1145/1400097.1400105.

[27] S. Floyd, *HighSpeed TCP for Large Congestion Windows*, RFC 3649, Dec. 2003. [Online]. Available: https://tools.ietf.org/html/rfc3649.

[28] T. Kelly, "Scalable TCP: Improving performance in highspeed wide area networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 33, no. 2, pp. 9–83, Apr. 2003, doi: 10.1145/956981.956989.

[29] L. Xu, K. Harfoush, and I. Rhee, "Binary increase congestion control (BIC) for fast long-distance networks," in *Proc. IEEE INFOCOM*, Hong Kong, Mar. 2004, pp. 2514–2524, doi: 10.1109/INFCOM.2004.1354672.

[30] D. Leith. (Oct. 2008). H-TCP: TCP congestion control for high bandwidth delay product paths. IETF Internet Draft. Accessed: Oct. 25, 2017. [Online]. Available: https://tools.ietf.org/html/draft-leith-tcp-htcp-06

[31] G. Marfia, C. Palazzi, G. Pau, M. Gerla, M. Y. Sanadidi, and M. Roccetti, "TCP Libra: Exploring RTT-fairness for TCP," UCLA Comput. Sci. Dept., Los Angeles, CA, USA, Tech. Rep. UCLA-CSD TR-050037, 2005.

[32] C. Caini and R. Firrincieli, "TCP hybla: A TCP enhancement for heterogeneous networks," *Int. J. Satell. Commun. Netw.*, vol. 22, no. 5, pp. 547–566, Sep. 2004. [Online]. Available: https://dl.acm.org/doi/10.1002/sat.799.

[33] A. Baiocchi, A. P. Castellani, and F. Vacirca, "YeAH-TCP: Yet another highspeed TCP," in *Proc. PFLDnet, ISI, Marina Del Rey*, Los Angeles, CA, USA, Feb. 2007, pp. 1–6.

[34] R. King, R. Baraniuk, and R. Riedi, "TCP-africa: An adaptive and fair rapid increase rule for scalable TCP," in *Proc. IEEE 24th Annu. Joint Conf. IEEE Comput. Commun. Societies*, Miami, FL, USA, Mar. 2005, pp. 1838–1848, doi: 10.1109/INFCOM.2005.1498463.

[35] K. Tan, J. Song, Q. Zhang, and M. Sridharan, "A compound TCP approach for high-speed and long distance networks," in *Proc. 25TH IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Barcelona, Spain, Apr. 2006, pp. 1–12, doi: 10.1109/INFOCOM.2006.188.

[36] S. Liu, T. Başar, and R. Srikant, "TCP-illinois: A loss- and delay-based congestion control algorithm for high-speed networks," *Perform. Eval.*, vol. 65, nos. 6–7, pp. 417–440, Jun. 2008, doi: 10.1016/j.peva.2007.12.007.

[37] L. S. Brakmo, S. W. O'Malley, and L. L. Peterson, "TCP vegas: New techniques for congestion detection and avoidance," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 24, no. 4, pp. 24–35, Oct. 1994, doi: 10.1145/190809.190317.

[38] J. Sing and B. Soh, "TCP new Vegas: Improving the performance of TCP Vegas over high latency links," in *Proc. 4th IEEE Int. Symp. Netw. Comput. Appl.*, Cambridge, MA, USA, Jul. 2005, pp. 73–80, doi: 10.1109/NCA.2005.52.

[39] K. Yamada, R. Wang, M. Y. Sanadidi, and M. Gerla, "TCP westwood with agile probing: Dealing with dynamic, large, leaky pipes," in *Proc. IEEE Int. Conf. Commun.*, Paris, France, Jun. 2004, pp. 1070–1074, doi: 10.1109/ICC.2004.1312665.

[40] D. Kliazovich, F. Granelli, and D. Miorandi, "Logarithmic window increase for TCP Westwood+ for improvement in high speed, long distance networks," *Comput. Netw.*, vol. 52, no. 12, pp. 2395–2410, Aug. 2008. [Online]. Available: https://dl.acm.org/doi/abs/10.1016/j.comnet.2008.04.018.

[41] H. Shimonishi, T. Hama, and T. Murase, "TCP-adaptive reno for improving efficiency-friendliness tradeoffs of TCP congestion control algorithm," in *Proc. 4th Int. Wksp. Protocols Fast Long Distance Netw.*, Feb. 2006, pp. 87–91.

[42] K. Kaneko, T. Fujikawa, Z. Su, and J. Katto, "TCP-Fusion: A hybrid congestion control algorithm for high-speed networks," in *Proc. Int. Wksp. PFLDnet, ISI, Marina Del Rey*, Los Angeles, CA, USA, Apr. 2007, pp. 31–36.

[43] G. Hasegawa, K. Kurata, and M. Murata, "Analysis and improvement of fairness between TCP reno and vegas for deployment of TCP Vegas to the Internet," in *Proc. Int. Conf. Netw. Protocols*, Osaka, Japan, Nov. 2002, pp. 177–186, doi: 10.1109/ICNP.2000.896302.

[44] V. Paxson, "End-to-end Internet packet dynamics," *IEEE/ACM Trans. Netw.*, vol. 7, no. 3, pp. 277–292, Jun. 1999, doi: 10.1109/90.779192.

[45] R. Ludwig and R. H. Katz, "The eifel algorithm: Making TCP robust against spurious retransmissions," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 30, no. 1, pp. 30–36, Jan. 2000. [Online]. Available: https://dl.acm.org/doi/10.1145/505688.505692.

[46] R. Ludwig and A. Gurtov, *The Eifel Response Algorithm for TCP*, RFC 4015, Feb. 2005. [Online]. Available: https://tools.ietf.org/html/rfc4015

[47] F. Wang and Y. Zhang, "Improving TCP performance over mobile ad-hoc networks with out-of-order detection and response," in *Proc. 3rd ACM Int. Symp. Mobile Ad Hoc Netw. Comput. (MobiHoc)*, New York, NY, USA, 2002, pp. 217–225, doi: 10.1145/513800.513827.

[48] A. Venkataramani, R. Kokku, and M. Dahlin, "TCP Nice: A mechanism for background transfers," *Oper. Syst. Rev.*, vol. 36, pp. 329–344, Dec. 2002, doi: 10.1145/844128.844159.
[49] S. Mascolo, C. Casetti, M. Gerla, M. Y. Sanadidi, and R. Wang, "TCP westwood: Bandwidth estimation for enhanced transport over wireless links," in *Proc. 7th Annu. Int. Conf. Mobile Comput. Netw. MobiCom*, Rome, Italy, 2001, pp. 287–297, doi: 10.1145/381677.381704.
[50] L. A. Grieco and S. Mascolo, "Performance evaluation and comparison of Westwood+, new reno, and vegas TCP congestion control," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 34, no. 2, pp. 25–38, Apr. 2004. [Online]. Available: https://dl.acm.org/doi/10.1145/997150.997155.
[51] T. Henderson, S. Floyd, A. Gurtov, and Y. Nishida, *The NewReno Modification to TCP's Fast Recovery Algorithm*, document RFC 6582, Apr. 2012. [Online]. Available: https://tools.ietf.org/html/rfc6582
[52] M. Pieska, A. Kassler, H. Lundqvist, and T. Cai, "Improving TCP fairness over latency controlled 5G mmWave communication links," in *Proc. 22nd Int. ITG Wksp. Smart Antennas (WSA)*, Bochum, Germany, Jun. 2018, pp. 1–8.
[53] N. Cardwell, Y. Cheng, C. S. Gunn, S. H. Yeganeh, and V. Jacobson, "BBR: Congestion-based congestion control," *Queue*, vol. 14, no. 5, pp. 20–53, Oct. 2016, doi: 10.1145/3012426.3022184.
[54] A. Narayanan, E. Ramadan, J. Carpenter, Q. Liu, Y. Liu, F. Qian, and Z.-L. Zhang, "A first look at commercial 5G performance on smartphones," 2019, *arXiv:1909.07532*. [Online]. Available: http://arxiv.org/abs/1909.07532
[55] M. Okano, Y. Hasegawa, K. Kanai, B. Wei, and J. Katto, "TCP throughput characteristics over 5G millimeterwave network in indoor train station," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Marrakesh, Morocco, Apr. 2019, pp. 1–6, doi: 10.1109/WCNC.2019.8886119.
[56] K. Nichols, V. Jacobson, A. McGregor, and J. Iyengar, *Controlled Delay Active Queue Management*, document RFC 8289, Jan. 2018. [Online]. Available: https://tools.ietf.org/html/rfc8289
[57] T. Hoeiland-Joergensen, P. McKenney, D. Taht, J. Gettys, and E. Dumazet, *The Flow Queue CoDel Packet Scheduler and Active Queue Management Algorithm*, document RFC 8290, Jan. 2018. [Online]. Available: https://tools.ietf.org/html/rfc8290.
[58] M. Allman, S. Floyd, and C. Partridge, *Increasing TCP's Initial Window*, RFC 3390, Oct. 2002. [Online]. Available: https://tools.ietf.org/html/rfc3390.
[59] 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)*, Atlanta, GA, USA, May 2017, pp. 343–348, doi: 10.1109/INFCOMW.2017.8116400.
[60] M. Mathis, J. Mahdavi, S. Floyd, and A. Romanow, *TCP Selective Acknowledgment Options*, document RFC 2018, Oct. 1996. [Online]. Available: https://tools.ietf.org/html/rfc2018
[61] J. Iyengar and M. Thomson. (Feb. 2020). QUIC: A UDP-based multiplexed and secure transport. Work in Progress. draft-ietf-quic-transport-29. [Online]. Available: https://datatracker.ietf.org/doc/draft-ietf-quic-transport/
[62] J. Rodriguez, *Fundamentals of 5G Mobile Networks*, 1st ed. Hoboken, NJ, USA: Wiley, Apr. 2015, doi: 10.1002/9781118867464.
[63] I. Petrov and T. Janevski, "Design of novel 5G transport protocol," in *Proc. Int. Conf. Wireless Netw. Mobile Commun. (WINCOM)*, Fez, Morocco, Oct. 2016, pp. 29–33, doi: 10.1109/WINCOM.2016.7777186.
[64] D. Scholz, B. Jaeger, L. Schwaighofer, D. Raumer, F. Geyer, and G. Carle, "Towards a deeper understanding of TCP BBR congestion control," in *Proc. IFIP Netw. Conf. (IFIP Networking) Workshops*, Zurich, Switzerland, May 2018, pp. 1–9, doi: 10.23919/IFIPNetworking.2018.8696830.
[65] J. Gettys and K. Nichols, "Bufferbloat: Dark buffers in the Internet," *Commun. ACM*, vol. 55, no. 1, pp. 57–65, Jan. 2012, doi: 10.1145/2063176.2063196.
[66] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the Internet of Things," in *Proc. 1st Ed. MCC Workshop Mobile cloud Comput. (MCC)*, New York, NY, USA, Aug. 2012, pp. 13–16. [Online]. Available: https://dl.acm.org/doi/10.1145/2342509.2342513
[67] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854–864, Dec. 2016, doi: 10.1109/JIOT.2016.2584538.
[68] 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)*, Dec. 2017, pp. 1–2, doi: 10.1109/PCCC.2017.8280492.
[69] Y. Hasegawa and J. Katto, "A transmission control protocol for long distance high-speed wireless communications," *IEICE Trans. Commun.*, vol. E101.B, no. 4, pp. 1045–1054, 2018, doi: 10.1587/transcom.2017EBP3229.
[70] A. Ford, C. Raiciu, M. Handley, O. Bonaventure, and C. Paasch, *TCP Extensions for Multipath Operation With Multiple Addresses*, document RFC 8684, Mar. 2002. [Online]. Available: https://tools.ietf.org/html/rfc8684.
[71] M. Polese, R. Jana, and M. Zorzi, "TCP and MP-TCP in 5G mmWave networks," *IEEE Internet Comput.*, vol. 21, no. 5, pp. 12–19, Sep. 2017, doi: 10.1109/MIC.2017.3481348.
[72] N. Baldo, M. Miozzo, M. Requena-Esteso, and J. Nin-Guerrero, "An open source product-oriented LTE network simulator based on NS-3," in *Proc. 14th ACM Int. Conf. Modeling, Anal. Simulation Wireless Mobile Syst. (MSWiM)*, Miami Beach, FL, USA, 2011, pp. 293–298, doi: 10.1145/2068897.2068948.
[73] *Network Simulator 3*. Accessed: Feb. 2, 2020. [Online]. Available: https://www.nsnam.org/
[74] CTTC. (Jan. 2020). *LTE-EPC Network Simulator*. [Online]. Available: http://networks.cttc.es/mobile-networks/software-tools/lena/
[75] M. Mezzavilla, S. Dutta, M. Zhang, M. R. Akdeniz, and S. Rangan, "5G mmWave module for the NS-3 network simulator," in *Proc. 18th ACM Int. Conf. Modeling, Anal. Simulation Wireless Mobile Syst. (MSWiM)*, Cancun, Mexico, Jun. 2015, pp. 283–290. [Online]. Available: https://dl.acm.org/doi/10.1145/2811587.2811619
[76] M. Zhang, M. Mezzavilla, J. Zhu, S. Rangan, and S. Panwar, "TCP dynamics over mmWave links," in *Proc. IEEE 18th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Sapporo, Japan, Jul. 2017, pp. 1–6, doi: 10.1109/SPAWC.2017.8227746.
[77] M. Zhang, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "NS-3 implementation of the 3GPP MIMO channel model for frequency spectrum above 6 GHz," in *Proc. Workshop NS-3 (WNS3)*, Porto, Portugal, Jun. 2017, pp. 71–78. [Online]. Available: https://dl.acm.org/doi/10.1145/3067665.3067678
[78] M. Polese, M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Improved handover through dual connectivity in 5G mmWave mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2069–2084, Sep. 2017, doi: 10.1109/JSAC.2017.2720338.
[79] M. Polese, M. Mezzavilla, and M. Zorzi, "Performance comparison of dual connectivity and hard handover for LTE-5G tight integration," in *Proc. 9th EAI Int. Conf. Simulation Tools Techn. (SIMUTOOLS)*, Prague, Czech Republic, Aug. 2016, pp. 118–123. [Online]. Available: https://dl.acm.org/doi/10.5555/3021426.3021445
[80] M. Mezzavilla, M. Zhang, M. Polese, R. Ford, S. Dutta, S. Rangan, and M. Zorzi, "End-to-End simulation of 5G mmWave networks," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2237–2263, 3rd Quart., 2018, doi: 10.1109/COMST.2018.2828880.
[81] H. Tazaki, F. Uarbani, E. Mancini, M. Lacage, D. Camara, T. Turletti, and W. Dabbous, "Direct code execution: Revisiting library OS architecture for reproducible network experiments," in *Proc. 9th ACM Conf. Emerg. Netw. Exp. Technol. (CoNEXT)*, Santa Barbara, CA, USA, Dec. 2013, pp. 217–228. [Online]. Available: https://dl.acm.org/doi/10.1145/2535372.2535374
[82] *Study on Channel Model for Frequency Spectrum Above 6 GHz, V14.2.0*, document TR 38.900, 3GPP, Sophia Antipolis, France, 2017.
[83] MATLAB. *5G Library for LTE System Toolbox*. Accessed: Feb. 20, 2020. [Online]. Available: https://www.mathworks.com/products/5g.html
[84] S. Choi, J. Song, J. Kim, S. Lim, S. Choi, T. T. Kwon, and S. Bahk, "5G K-SimNet: End-to-end performance evaluation of 5G cellular systems," in *Proc. 16th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, Las Vegas, NV, USA, Jan. 2019, pp. 1–6, doi: 10.1109/CCNC.2019.8651686.
[85] Y. Kong, H. Zang, and X. Ma, "Improving TCP congestion control with machine intelligence," in *Proc. Workshop Netw. Meets AI ML (NetAI)*, Budapest, Hungary, Aug. 2018, pp. 60–66. [Online]. Available: https://dl.acm.org/doi/10.1145/3229543.3229550
[86] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015, doi: 10.1016/j.neunet.2014.09.003.
[87] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[88] W. Feng, Y. Wang, D. Lin, N. Ge, J. Lu, and S. Li, "When mmWave communications meet network densification: A scalable interference coordination perspective," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1459–1471, Jul. 2017, doi: 10.1109/JSAC.2017.2698898.

[89] G. N. Vivekananda and P. C. Reddy, "Critical analysis of cross-layer approach," in *Proc. Int. Conf. Green Comput. Internet Things (ICGCIoT)*, Noida, India, Oct. 2015, pp. 12–16, doi: 10.1109/ICGCIoT.2015.7380419.

REZA POORZARE received the B.S. and M.S. degrees in computer engineering from the Azad University of Iran, in 2010 and 2014, respectively. He is currently pursuing the Ph.D. degree in network engineering with the Universitat Politècnica de Catalunya. His research interests include 5G, mmWave, TCP, wireless mobile networks, and artificial intelligence.

ANNA CALVERAS AUGÉ was born in Barcelona, Spain, in 1969. She received the Ph.D. degree in telecommunications engineering from the Universitat Politècnica de Catalunya, Spain, in 2000. She is currently an Associate Professor with the Wireless Networks Group (WNG), Department of Computer Networks, Universitat Politècnica de Catalunya. Her research interests include the design, evaluation, and optimization of communications protocols and architectures for cellular, wireless multihop networks, ad-hoc networks, wireless sensor networks, the Internet of Things, and application domains, such as smart cities, building automation, satellite, and emergency environments. She has been involved in several national and international research or technology transfer projects. She has published in international and national conferences and journals.