A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions

Imtiaz Parvez, Student Member, IEEE, Ali Rahmati, Student Member, IEEE, Ismail Guvenc, Senior Member, IEEE, Arif I. Sarwat, Senior Member, IEEE, and Huaiyu Dai, Fellow, IEEE

Abstract—The fifth generation (5G) wireless network technology is to be standardized by 2020, where main goals are to improve capacity, reliability, and energy efficiency, while reducing latency and massively increasing connection density. An integral part of 5G is the capability to transmit touch perception type real-time communication empowered by applicable robotics and haptics equipment at the network edge. In this regard, we need drastic changes in network architecture including core and radio access network (RAN) for achieving end-to-end latency on the order of 1 ms. In this paper, we present a detailed survey on the emerging technologies to achieve low latency communications considering three different solution domains: RAN, core network, and caching. We also present a general overview of 5G cellular networks composed of software defined network (SDN), network function virtualization (NFV), caching, and mobile edge computing (MEC) capable of meeting latency and other 5G requirements.

Index Terms—5G, cloud, caching, haptic communications, latency, massive connectivity, real-time communication, SDN, tactile Internet, ultra-high reliability, ultra-low latency.

I. INTRODUCTION

The focus of next generation mobile communication is to provide seamless communication for machines and devices building the Internet-of-Things (IoT) along with personal communication. New applications such as tactile Internet, high-resolution video streaming, tele-medicine, tele-surgery, smart transportation, and real-time control dictate new specifications for throughput, reliability, end-to-end (E2E) latency, and network robustness [1]. Additionally, intermittent or always-on type connectivity is required for machine-type communication (MTC) serving diverse applications including sensing and monitoring, autonomous cars, smart homes, moving robots and manufacturing industries.

Several emerging technologies including wearable devices, virtual/augmented reality, and full immersive experience (3D) are shaping the demeanor of human end users, and they have special requirements for user satisfaction. Therefore, these use cases of the next generation network push the specifications of 5G in multiple aspects such as data rate, latency, reliability, device/network energy efficiency, traffic volume density, mobility, and connection density. Current fourth generation (4G) networks are not capable of fulfilling all the technical requirements for these services.

Fifth generation (5G) cellular network is the wireless access solution to fulfill the wireless broadband communication specifications of 2020 and beyond [3], [4]. In ITU, 5G ITU-R working group is working for development of 5G under the term IMT-2020 [5]. The vision of this work is to achieve one thousand times throughput improvement, 100 billion connections, and close to zero latency [1]. In particular, 5G will support enhanced mobile broadband (MBB) with end-user data rates of 100 Mbps in the uniform spatial distribution with peak data rates of 10-20 Gbps. Based on consensus, 5G will not only provide personal mobile service, but also massive machine type communications (MTC), and latency/reliability critical services. In mission critical communication (MCC), both the latency and reliability issues need to be addressed [6]. In many cases, the corresponding E2E latency as low as 1 ms needs to be met with reliability as high as 99.99% [7].

The evolution of latency requirement for different generations of cellular networks is presented in Fig. 1.

To achieve low latency for MCC, drastic changes in the network architecture need to be performed. Since the delay is contributed by radio access network (RAN) and core network along with backhaul between RAN and core network, new network topology involving software define network (SDN), network virtualized function (NFV), and mobile edge computing (MEC)/caching can be employed to reduce the latency significantly. This can happen due to the capability of seamless operation and independence from hardware functionality provided by these entities. Moreover, new physical air interface with small time interval transmission, small size packets, new waveforms, new modulation and coding schemes are the areas of investigation for attaining low latency. In addition, optimization of radio resource allocation, massive MIMO, carrier aggregation in millimeter wave, and priority of data transmission need to be addressed. All in all, a robust integration with existing LTE is necessary for 5G networks that will enable industries to deploy 5G quickly and efficiently when it is standardized and available.

In summary, 5G wireless access should be an evolution of LTE complemented with revolutionary architecture designs and radio technologies.

In the literature, surveys on 5G network including architecture [3], [4], SDN/NFV/MEC based core network [9], [10].

![Fig. 1: Latency for different generations of cellular networks](image-url)
In this paper, we present a comprehensive survey of latency reduction solutions particularly in the context of 5G wireless technology. We first present the sources and fundamental constraints for achieving low latency in a cellular network. We also overview an exemplary 5G network architecture with compliance to IMT-2000 vision. Finally, we provide an extensive review of proposed solutions for achieving low latency towards 5G. The goal of our study is to bring all existing solutions on the same page along with future research directions. We divide the existing solutions into three parts: (1) RAN solutions; (2) Core network solutions; (3) Caching solutions. However, detailed comparison of these solutions are beyond the scope of this work.

The rest of the paper is organized as follows. In Section II, an exemplary 5G architecture is presented with visions to meet the requirements of low latency, high reliability, very large throughput, and massive connectivity. Section III presents the latency critical services in 5G. The sources of latency in a cellular network are discussed in Section IV. Section V reviews the fundamental constraints and approaches for achieving low latency. Three key low latency solutions in RAN, core network, and caching are presented in Sections VI, VII, and VIII, respectively. Section IX presents the field tests, trials and experiments of low latency approaches. Finally, concluding remarks are provided in Section X.
input multiple-output (MIMO), advanced multiple access techniques, small cell deployment, and millimeter wave (mmWave) aggregation are proposed. On the other hand, non orthogonal multiple access, flexible packet structure, overloaded data transmission, and active user detection are being considered for mMTC. For low latency and reliability enhancement (i.e. uRLLC), modification in the physical layer along upper layers is required. An example candidate for 5G network is presented in Fig. 2 with proposed latency parameters tabulated in Table I.

Some enhancements including ultra-dense small cell deployments and new technologies such as massive MIMO, SDN, NFV, MEC/caching are the fundamental modifications in the proposed 5G network [8]. The dynamic deployment and scaling of functions can be achieved by SDN and NFV where access stratum (AS) and non access stratum (NAS) are integrated by intelligent protocol. Split of control plane (C-plane) and data plane (D-plane) in the architecture and the carrier frequency could be performed in order to have separate provision of capacity and coverage [21]. Special network functionality as a service (XaaS) which can be assumed as the relation between a radio network and the cloud will grant service on request/demand. Resource pooling can be an example of this XaaS. Moreover, intelligent data driven network structures such as mobile edge computing cloud/fog network will reduce latency and enhance the network throughput. The high data rate and reduced latency in radio interface will also lead to long battery life and energy savings.

In the following section, latency critical services in 5G networks are discussed in details.

### III. Low Latency Services in 5G

Latency is highly critical in some applications such as automated industrial production, control/robotics, transportation, health-care, entertainment, virtual reality, education, and culture (as illustrated in Fig. 3). In some cases, we need latency as low as 1 ms with packet loss rate no larger than $10^{-2}$. Several latency critical services which need to be supported by 5G are described as follows.

- **Factory Automation**: Factory automation includes real-time control of machine and system for quick production lines and limited human involvement. In these cases, the production lines might be numerous and contiguous. This is highly challenging in terms of latency and reliability. Therefore, the latency requirement for factory automation applications is between 0.25 ms to 10 ms with a packet loss rate of $10^{-9}$ [22], [23].

- **Intelligent Transportation Systems**: Autonomous driving and optimization of road traffic requires ultra reliable low latency communication. According to intelligent transportation systems (ITS), different cases including autonomous driving, road safety, and traffic efficiency services have different requirements [22], [24]. Autonomous vehicles require coordination among themselves for actions such as platooning and overtaking [25]. Road safety includes warnings about collisions or dangerous situations. Traffic efficiency services control traffic flow using the information of the status of traffic lights and local traffic situations. For these purposes, latency of 10 ms to 100 ms with packet loss rate of $10^{-3}$ to $10^{-5}$ is required.

- **Robotics and Telepresence**: In the near future, remote controlled robots will have applications in diverse sectors such as construction and maintenance in dangerous areas. A prerequisite for the utilization of robots and telepresence applications is remote-control with real-time synchronous visual-haptic feedback. In this case, system response times should be less than a few milliseconds including network delays [22], [26], [27]. Communication infrastructure capable of proving this level of real-time capacity, high reliability/availability, and mobility support is to be addressed in 5G networks.

- **Virtual Reality (VR)**: Several applications such as micro-assembly and tele-surgery require very high levels of sensitivity and precision for object manipulations. VR technology accommodates such services where several

### TABLE I: Default Latency Values for Physical Layer [7].

| Parameter                                              | value (ms) |
|--------------------------------------------------------|------------|
| User equipment (UE) processing time                    | 0.3        |
| eNB processing time                                    | 0.3        |
| Minimum time to transmit/transmission time interval (TTI)| 0.2        |
| Time required for uplink (UL) ACK/NACK transmission    | 0.06       |
|                                                       | 0.25       |
|                                                       | 0.30       |
|                                                       | 1.00       |
| Downlink (DL) resources for bidirectional TDD blocks with 1 ms duration | 0.80 (DL heavy block) |
|                                                       | 0.40 (balanced block) |
|                                                       | 0.20 (UL heavy block) |
| DL resources for bidirectional TDD blocks with 4 ms duration | 2.8 (DL heavy block) |

Fig. 3: Latency critical services in 5G.
users interact via physically coupled VR simulations in a shared haptic environment. Current networked communication does not allow sufficient low latency for stable, seamless coordination of users. Typical update rates of display for haptic information and physical simulation are in the order of 1000 Hz which allows round trip latency of 1 ms. Consistent local view of VR can be maintained for all users if and only if the latency of around 1 ms is achieved [26], [27].

- **Augmented Reality (AR):** In AR technology, the augmentation of information into the user’s field of view enables applications such as driver-assistance systems, improved maintenance, city/museum guides, telemedicine, remote education, and assistive technologies for police and firefighters [26]. However, insufficient computational capability of mobile devices and latency of current cellular network hinder the applications. In this case, latency as low as a few milliseconds is required.

- **Health care:** Tele-diagnosis, tele-surgery and tele-rehabilitation are a few notable healthcare applications of low latency tactile Internet. These allow for remote physical examination even by palpation, remote surgery by robots, and checking of patients’ status remotely. For these purposes, sophisticated control approaches with round trip latency of 1-10 ms and high reliability data transmission is mandatory [26], [27].

- **Serious Gaming:** The purpose of serious gaming is not limited to entertainment. Such games include problem-solving challenges, and goal-oriented motivation which can have applications in different areas such as education, training, simulation, and health. Network latency of more than 30-50 ms results in a significant degrade in game-quality and game experience ratings. Ideally, a round trip time (RTT) on the order of 1 ms is recommended for perceivable human’s interaction with the high-quality visualization [26].

- **Smart Grid:** The smart grid has strict requirements of reliability and latency. The dynamic control allows only 100 ms of end-to-end latency for switching suppliers (PV, windmill, etc.) on or off. However, in case of a synchronous co-phasing of power suppliers (i.e., generators), an end-to-end delay of not more than 1 ms is needed [26]. Latency more that 1 ms which is equivalent to a phase shift of about 18° (50 Hertz AC network) or 21.6° (60 Hertz AC network), may have serious consequence in smart grid and devices.

- **Education and Culture:** Low latency tactile Internet will facilitate remote learning/education by haptic overlay of teacher and students. For these identical multi-modal human-machine interfaces, round trip latency of 5-10 ms is allowed for perceivable visual, auditory, and haptic interaction [26], [27]. Besides that, tactile Internet will allow to play musical instruments from remote locations. In such scenarios, supporting network latency lower than few milliseconds becomes crucial [26].

Based on the applications and use case scenarios above, latency critical services in 5G networks demand an E2E delay of 1 ms to 100 ms. The latency requirements for various 5G services are summarized in Table II.

In the next section, the major sources of latency in a cellular network are discussed.

### IV. Sources of Latency in a Cellular Network

In the LTE system, the latency can be divided into two major parts: (1) user plane (U-plane) latency and (2) control plane (C-plane) latency. The U-plane latency is measured by one directional transmit time of a packet to become available in the IP layer between evolved UMTS terrestrial radio access network (E-UTRAN) edge/UE and UE/E-UTRAN node [28]. On the other hand, C-plane latency can be defined as the transition time of a UE to switch from idle state to active state. At the idle state, an UE is not connected with radio resource control (RRC). After the RRC connection is being setup, the UE switches from idle state into connected state and then enters into active state after moving into dedicated mode. Since the application performance is dependent mainly on the U-plane latency, U-plane is the main focus of interest for low latency communication.

In the U-plane, the delay of a packet transmission in a cellular network can be contributed by the RAN, backhaul, core network, and data center/Internet. As referred in Fig. 5, the total one way transmission time [29] of current LTE system can be written as

\[ T = T_{Radio} + T_{Backhaul} + T_{Core} + T_{Transport} \]  

where

- \( T_{Radio} \) is the packet transmission time between eNB and UEs and is mainly due to physical layer communication;
- \( T_{Backhaul} \) is the time for building connections between eNB and the core network (i.e., EPC). Generally, the core network and eNB are connected by copper wires or microwave or optical fibers;
- \( T_{Core} \) is the processing time taken by the core network;
- \( T_{Transport} \) is the delay to data communication between the core network and Internet/cloud.

The end-to-end delay, \( T_{E2E} \) is then approximately given by \( 2 \times T \). The \( T_{Radio} \) is the sum of transmit time, propagation latency, processing time (channel estimation, encoding and decoding time for first time), and retransmission time (due to packet loss). In particular, the \( T_{Radio} \) for a scheduled user [30], [31] can be expressed as:

\[ T_{Radio} = t_Q + t_{FA} + t_{tx} + t_{bhp} + t_{mpt} \]
where

- $t_Q$ is the queuing delay which depends on the number of users that will be multiplexed on same resources;
- $t_{FA}$ is the delay due to frame alignment which depends on the frame structure and duplexing modes (i.e., frequency division duplexing (FDD) and time division duplexing (TDD));
- $t_{tx}$ is the time for transmission processing, and payload transmission which uses at least one TTI depending on radio channel condition, payload size, available resources, transmission errors and retransmission;
- $t_{bsp}$ is the processing delay at the base station;
- $t_{mpt}$ is the processing delay of user terminal. Both the base station and user terminal delay depend on the capabilities of base station and user terminal (i.e., UE), respectively.

In compliance with ITU, $T_{Radio}$ should not be more than 0.5 ms for low latency communication [32]. In this regard, radio transmission time should be designed to be on the order of hundreds of microseconds while the current configuration in 4G is 1 ms. For this, enhancement in various areas of RAN such as packet/frame structure, modulation and coding schemes, new waveform designs, transmission techniques, and symbol detection need to be carried out. In order to reduce the delay in $T_{Backhaul}$, approaches such as advanced backhaul techniques, caching/fog enabled networks, and intelligent integration of AS and NAS can provide potential solutions. For $T_{Core}$, new core network consists of SDN, NFV, and various intelligent approaches can reduce the delay significantly. For $T_{Transport}$, MEC/fog enabled Internet/cloud/caching can provide reduced latency.

In the following section, we discuss the constraints and approaches for achieving low latency.

V. CONSTRAINTS AND APPROACHES FOR ACHIEVING LOW LATENCY

There are major fundamental trade-offs between capacity, coverage, latency, reliability, and spectral efficiency in a wireless network. Due to these fundamental limits, if one metric is optimized for improvement, this may result in degradation of another metric. In the LTE system, the radio frame is 10 ms with the smallest TTI being 1 ms. This fixed frame structure depends on the modulation and coding schemes for adaptation of the transmission rate with constant control overhead. Since latency is associated with control overhead (cyclic prefix, transmission mode, and pilot symbols) which occupies a major portion of transmission time of a packet (approximately 0.3-0.4 ms per packet transmission), it is not wise to consider a packet with radio transmission time less than 1 ms. If we design a packet with time to transmit of 0.5 ms, more than 60% of the resources will be used by control overhead [29]. Moreover, retransmission per packet transmission takes around 8 ms, and removal of retransmission will affect packet error significantly. As a result, we need radical modifications and enhancements in packet/frame structure and transmission strategy. In this regard:

- First, a novel radio frame reinforced by limited control overhead and smaller transmission time is necessary to be designed. For reduction of control overhead, procedures for user scheduling, resource allocation, and channel training can be eliminated or merged.
- Second, packet error probability for first transmission should be reduced with new waveforms and transmission techniques reducing the retransmission delay.
- Third, since latency critical data needs to be dispatched immediately, techniques for priority of data over normal data need to be identified.

![Fig. 4: Categories of different solutions for achieving low latency in 5G.](image-url)
• Fourth, synchronization and orthogonality are the indispensable aspects of OFDM that are major barriers for achieving low latency. Even though asynchronous mode of communication is more favorable over synchronized operation in terms of latency, it requires additional spectrum and power resources [33].

• Fifth, since the latency for data transmission also depends on the delay between the core network and the BS, caching networks can be used to reduce latency by storing the popular data at the network edge.

Researchers proposed various techniques/approaches for achieving low latency in 5G. As summarized in Fig. 4, we divided the existing solutions into three major categories: (1) RAN solutions, (2) core network solutions, and (3) caching solutions. The RAN solutions include new/modified frame or packet structure, waveform designs, multiple access techniques, modulation and coding schemes, transmission schemes, control channels enhancements, low latency symbol detection, mmWave aggregation, cloud RAN, reinforcing QoS and QoE, energy-aware latency minimization, and location aware communication techniques. On the other hand, new entities such as SDN, NFV, MEC and fog network along with new bachaul based solutions have been proposed for the core network. The solutions of caching can be subdivided into caching placement, content delivery, centralized caching, and core network. The solutions of caching can be subdivided into caching placement, content delivery, centralized caching, and distributed caching, while bachaul solutions can be divided into general and mmWave bachaul. In the following sections, these solutions are described in further details.

VI. RAN SOLUTIONS FOR LOW LATENCY

To achieve low latency, various enhancements in the RAN have been proposed. Referring to Table III, RAN solutions/enhancements include frame/packet structure, advanced multiple access techniques/waveform designs, modulation and coding scheme, diversity and antenna gain, control channel, symbol detection, energy-aware latency minimization, carrier aggregation in mmWave, reinforcing QoS and QoE, cloud RAN and location aware communication. In what follows, the detailed overview for each of these solutions is presented.

A. Frame/packet structure

In the RAN solutions, modification in the physical air interface has been considered as an attractive choice. In particular, most of the proposed solutions are on the physical (PHY) and medium access control (MAC) layers.

In LTE cellular network, the duration of a radio frame is 10 ms. Each frame is partitioned into 10 subframes of size 1 ms which is further divided into 0.5 ms units that are referred as a resource block (RB). Each RB spans 0.5 ms (6 or 7 OFDM symbols) in time domain and 180 KHz (12 consecutive subcarriers, each of which 15 KHz) in frequency domain. Based on this, the subcarrier spacing Δf is 15 KHz, the OFDM symbol duration T_{OFDM} = 1/Δf = 66.67µs, the FFT size is 2048, the sampling rate f_s is Δf × N_{FFT} = 33.72 MHz and the sampling interval T_s = 1/f_s.

To reduce TTI for achieving low latency, the subcarrier spacing Δf can be changed to 30 KHz [37]. This results the corresponding OFDM symbol duration T_{OFDM} to be 33.33 µs and the FFT size N_{FFT} to become 1024 while sampling rate f_s is kept 30.72 MHz similar to LTE systems. The frame duration T_{f} = 10 ms can be divided into 40 subframes in which each subframe duration T_{sf} is 0.25 ms and contains 6 or 7 symbols. Two types of cyclix prefixs (CPs) can be employed in this configuration with durations

\[ T_{cp1} = 5/64 × N_{FFT} × T_s ≈ 2.604 \mu s, \]  \[ T_{cp2} = 4/64 × N_{IFT} × T_s ≈ 2.083 \mu s. \]

An exemplary physical air frame and conventional LTE radio frame with equal sized RB are illustrated in Fig. 5(a) and Fig. 5(b), respectively.

In [34], an extensive analysis of the theoretical principles that regulates the transmission of small-scale packets with low latency and high reliability is presented with metrics to assess their performance. The authors emphasize control overhead optimization for short packet transmission. In [35], small-scale data such as 10 – 20 bytes are proposed for transmission in order to achieve latency as low as 0.5 – 1 ms and reliability (probability of failed packet delivery) of not higher than 10^{−9}. In this regard, various forms of diversity could be employed such as spatial diversity, time diversity, and frequency diversity. The spatial, time, and frequency diversity is accomplished by means of several transmit and receive antennas, time slots of independent coefficients, and multiple RBs of independent fading coefficients, respectively.

In [45], a flexible 5G radio frame structure is introduced in which the TTI size is configurable in accordance with the requirement of specific services. At low offered load, 0.25 ms TTI is an attractive choice for achieving low latency due to low control overhead. However, for more load, control overhead increases which affects reliability and packet recovery mechanism resulting in increased latency. This study argues to employ user scheduling with different TTI sizes in the future 5G networks. In [44], the numerology and subframe structure are defined considering diverse carrier frequencies and bandwidths for low latency 5G networks. Cyclic prefix, FFT size, subcarrier spacing, and sampling frequency were expressed as a function of the carrier frequency. In [41], software defined radio (SDR) platform based 5G system implementation with
### TABLE III: OVERVIEW OF TECHNIQUES IN RAN FOR LOW LATENCY.

| Case/Area | Reference | Approach | Summary |
|-----------|-----------|----------|---------|
| Frame/Packet structure | [40]-[42] | Transmission of small packets | Transmission of small scale data is investigated for packet loss rate of 10^{-6} and latency as low as 1 ms. |
| | [37] | Subcarrier spacing | Subcarrier spacing is enlarged to shorten the OFDM symbol duration, and the number of OFDM symbols is proposed to keep unchanged in each subframe. |
| | [35], [38], [40] | Flexible OFDMA based TDD subframe | TDD numerology is optimized for dense deployment with smaller cell sizes and larger bandwidth in the higher carrier frequencies. |
| | [45] | Modification of physical subframe | Different control and data part patterns for consecutive subframes, TX and RX control parts are proposed to be separated from each other, and from the data symbols with a GP, leading to total number of 3 GPs per subframe. |
| | [46] | Short frame design numerology for local area | The main focus in designing physical parameter is FFT, frame duration and physical channel for local area based communication. |
| | [41], [43] | Numerology and flexible sub frame | Numerology and subframe structure are defined considering diverse carrier frequencies and bandwidths to envision 5G including low latency. Cyclic prefix, FFT size, subcarrier spacing, and sampling frequency were expressed as the function of carrier frequency. |
| | [47], [49] | mmWave MAC layer frame structure | A new mmWave MAC layer frame structure is proposed with several improvements such as adaptable and smaller transmission times, extended messaging, dynamic locations for control signals, and the capacity to multiplex directional control signals (dynamic HARQ placement) efficiently. |
| Advanced multiple access/Waveform | [47], [48], [49] | Filtered CP-OFDM, UFMC and FBMC | UFMC outperforms over OFDM by about 10% in case of both large and small packets, FBMC demonstrates better performance in transmitting long sequences; however, it suffers during the transmission of short bursts/frames. |
| | [40] | Polar coding | Based on simulation and field test, polar coding has been proposed for 5G, outperforming over turbo coding in case of small packet transmission. |
| Modulation and coding | [42] | Turbo decoding with combined sliding window algorithm and cross parallel window (CPW) algorithm | A highly-parallel architecture for the latency sensitive turbo decoding is proposed combining two parallel algorithms: the traditional sliding window algorithm and cross parallel window (CPW) algorithm. |
| | [47] | New IFFT design with butterfly operation | Input signal of IFFT processor corresponding to guard band are assigned as null revealing the existence of numerous zeros (i.e., 0). If the sequence of OFDM symbol data which enter the IFFT is adjusted, the memory depth can be reduced from 1024 to 176. |
| | [41] | Sparse code multiple access (SCMA) | A dynamic shrunk square searching (DSSS) algorithm is proposed, which cuts off unnecessary communication control port (CCP) calculation along with utilization of both the noise characteristic and state space structure. |
| | [50] | Priority to latency critical data | A latency reduction approach by introducing TDM of higher priority ultra-low latency data over other less time critical services is proposed which maps higher priority user data during the beginning of a subframe followed by the normal data. |
| | [51] | Balanced truncation | Balanced truncation is applied for the model reduction in the linear systems that are being coupled over arbitrary graphs under communication latency constraints. |
| | [52] | Finite block length bounds and coding | Recent advances in finite-block length information theory are utilized in order to demonstrate optimal design for wireless systems under strict constraints such as low latency and high reliability. |
| Transmission | [55] | Asymmetric window | Asymmetric window is proposed instead of well-known symmetric windows for reduction of cyclic prefix by 30%. This technique suppresses OOB emission but makes the system more susceptible to channel induced ISI and ICI. |
| | [50] | Transmission power optimization | Transmission power is optimized by steepest descent algorithm considering transmission delay, error probability and queuing delay. |
| | [51] | Path-switching method and a packet-recovery method | Low latency packet transport system uses a quick path-switching method and a packet-recovery method are introduced for a multi-radio-access technology (multi-RAT) environment. |
| | [52] | Diversity | Diversity could be employed through various approaches such as spatial diversity, time diversity, and frequency diversity. |
| Control channel | [53] | Control channel sparse encoding (CCSE) | CCSE is introduced in order to provide the control information using non-orthogonal spreading sequences. |
| | [54] | Scaled control channel design | A scaled-LTE frame structure is proposed assuming the scaling factor to be 5 with a dedicated UL CCHs for all sporadic-traffic users in each transmission time interval with possible smallest SR size. |
| | [55] | Symbol-level frequency hopping and sequence-based sPUCCH | A sequence-based sPUCCH (SS-PUCCH) incorporating two SC-FDMA symbols is introduced in order to meet a strict latency requirement. Symbol-level frequency hopping technique is employed to achieve frequency diversity gain and reliability enhancement. |
| | [56] | Radio bearer and S1 bearer management | Establishment of radio bearer and S1 bearer in parallel are proposed where eNB and mobility management element (MME) manages and controls radio bearer and S1 bearer, respectively. The eNB sends only single control signal in order to configure radio bearers such as SRB1, SRB2 and DRBs, that decreases the signaling interaction rounds between the UE and the eNBs. |
| | [57] | Outer-loop link adaptation (OLL A) scheme | The proposed scheme controls the size of the compensation in the estimated SINR based on the time elapsed after a UE transits from an idle state to an active state, which helps to reduce latency for small packet applications. |
TABLE III: OVERVIEW OF TECHNIQUES IN RAN FOR LOW LATENCY (CONTINUED).

| Case/Area          | Reference | Approach                                      | Summary                                                                                                                                 |
|--------------------|-----------|-----------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|
| Symbol detection   | [66]      | SM-MIMO detection scheme with ZF and MRC-ZF    | A low-complexity and low latency massive SM-MIMO detection scheme is introduced and validated using SDR platforms. The low complexity detection scheme is proposed with a combination of ZF and MRC-ZF. |
|                    | [67]      | Linear MMSE                                    | A linear MMSE receiver is presented for low latency wireless communications using ultra-small packets.                                     |
|                    | [68]      | Space-time encoding and widely linear estimator | Space-time encoding is introduced within a GFDM block for maintaining overall low latency in the system. On the other hand, a widely linear estimator is used to decode the GFDM block at the receiver end, which yields significant improvements in gain over earlier works. |
|                    | [69]–[71] | Compressed sensing                             | Compressed sensing has been proposed to be effective in reducing latency of networked control systems if the state vector can be assumed to be sparse in some representation. |
|                    | [72]      | Low complexity receiver design                 | A low complexity receiver is designed and using this, the performance of an SCMA system is verified via simulations and real-time prototyping. This approach triples the whole system throughput while maintaining low latency similar to flexible orthogonal transmissions. |
| mmWave aggregation | [46], [73]–[75] | mmWave based air interface                     | Physical layer air interface is proposed using mmWave aggregation. Large bandwidth along with various approaches such as small frame structure, mmWave backhaul and beamforming can help to achieve low latency. |
| Cloud RAN          | [76]      | Virtualized cloud RAN (CRAN)                   | A CRAN over passive optical network (PON) architecture is introduced called virtualized-CRAN (V-CRAN) which can dynamically associate any radio unit (RU) to any digital unit (DU) so that several RUs can be coordinated by the same DU, and the concept of virtualized BS (V-BS) that can jointly transmit common signals from multiple RUs to a user. |
| Location           | [36], [40], [77] | Location information                           | Issues and research challenges of 5G are discussed followed by the conclusion that 5G networks can exploit the location information and accomplish performance gains in terms of throughput and latency. |

Fig. 6: Physical air interface (a) Conventional LTE radio frame, (b) Exemplary 5G radio frame with flexible time and frequency division for low latency [78] (GB: guard band; LLC: low latency communication).
strict latency requirement is presented. The scalability of the proposed radio frame structure is validated with E2E latency less than 1 ms.

In [36], a 5G flexible TDD is proposed for local area radio interface (5GETLA) with FFT size of 256 and 512, and short frame structure to achieve latency of lower than 1 ms. In particular, RTT of 0.25 ms can be achieved for the packets smaller than 50 Kbits. In [37], the proposed subcarrier spacing is enlarged to shorten the OFDM symbol duration, and the number of OFDM symbols in each subframe is kept unchanged in the new frame structure for TDD downlink. The subcarrier spacing is changed to 30 KHz resulting the corresponding OFDM symbol duration \(T = 33.33 \mu s\). The fast Fourier transform (FFT) size \(N\) is 1024, while the sampling rate \(f_s\) is kept same as 30.72 MHz. The frame duration \(T_s\) is still 10 ms with 40 subframes.

In [38], in order to have fully flexible allocations of different control and data RB in the consecutive subframes, TX and RX control RBs are proposed to be separated from each other and also from the data RB by guard periods (GPs). This leads to total number of 3 GPs per subframe which separates them. Assuming symmetrical TX and RX control parts with \(N_{ctrl,s}\) symbols in each and defining that same subcarrier spacing is used for control and data planes, with \(N_{data,s}\) being the number of data symbols and \(T_{sym}\) being the length of an OFDM symbol, the subframe length \(T_{sf}\) can be determined as

\[
T_{sf} = \left(2N_{ctrl,s} + N_{data,s}(T_{symbol} + T_{GP})\right) + 2T_{GP}. \tag{5}
\]

In [38], the fundamental limits and enablers for low air interface latency are discussed with a proposed flexible OFDM based TDD physical subframe structure optimized for 5G local area (LA) environment. Furthermore, dense deployment with smaller cell sizes and larger bandwidth in the higher carrier frequencies are argued as notable enablers for air interface latency reduction. In [7], a new configurable 5G TDD frame design is presented, which allows flexible scheduling (resource allocation) for wide area scenarios. The radical trade-offs between capacity, coverage, and latency are discussed further with the goal of deriving a 5G air interface solution capable of providing low latency, high reliability, massive connectivity, and enhanced throughput. Since achieving low latency comes at cost of lower spectral efficiency, the proposed solution of the study includes control mechanisms for user requirement, i.e. whether the link should be optimized for low latency or high throughput.

A 5G flexible frame structure in order to facilitate users with highly diversified service requirements is proposed in [39]. Although, in-resource physical layer control signaling is the basis of this proposed radio frame, it allows the corresponding data transmission based on individual user requirements. For this, it incorporates adaptable multiplexing of users on a shared channel with dynamic adjustment of the TTI in accordance with the service requirements per link. This facilitates optimization of the fundamental trade-offs between latency, spectral efficiency, and reliability for each link and service flow. In [40], a new radio interface design along with the corresponding numerology is presented for future local area system by using flexible TDD, and coexistence with overlay LTE-A network, sleeping modes, interference statistics, contention based data channel, and channel quality indicator. Here, the channel quality and traffic statistics are accumulated from the small cells which help to boost throughput and to reduce latency.

A new mmWave MAC layer frame design with adaptable and smaller transmission times, extended messaging, dynamic locations for control signals, and the capacity to multiplex directional control signals (dynamic HARQ placement) efficiently is presented in [46]. The granular transmission time, and dynamic and directional control signals ensure low latency. PHY and MAC layer solutions [77] including the compressed sensing based multi user detection, sparse code multiple access and continuous phase modulation are developed by the METIS project to address low latency and high node density challenges. In [79], a scheme that reserves resources for re-transmission for a group of ultra reliable low latency communication UEs is presented. The optimum dimensioning of groups and block error rate (BLER) target can reduce the probability of contention for the shared retransmission resources. Moreover, the unused resources can be utilized for non-grouped UEs resulting in overall efficiency enhancement.

In [80], fundamental trade-offs among three KPIs (reliability, latency, and throughput) in a 4G network is characterized, and an analytical framework is derived. In cases where the theory can not be extended via mathematical formulations due to complexity of scenario in hand, some guidelines are

### TABLE IV: PHY AND MAC LAYER BASED RADIO INTERFACE SOLUTIONS FOR LOW LATENCY.

| References | Approach/Area | PHY layer | MAC layer |
|------------|---------------|-----------|-----------|
| [45]       | Short TTI     | ✓         |           |
| [34], [44] | Numerology and sub frame structure | ✓         |           |
| [37]       | Subcarrier    | ✓         |           |
| [36], [38] | Flexible subframe | ✓         |           |
| [41]       | Flexible subframe implementation with SDR platform | ✓         |           |
| [43]       | Allocation of control and data RB | ✓         | ✓         |
| [40]       | Radio frame using local statistics | ✓         | ✓         |
| [7]        | Radio frame and scheduling | ✓         | ✓         |
| [39]       | Flexible TTI and multiplexing | ✓         | ✓         |
| [46]       | mmWave based MAC layer | ✓         | ✓         |
| [79]       | Reservation of resources | ✓         |           |
| [71]       | Radio frame with compressed-sensing-based multi-user detection | ✓         | ✓         |
| [80]       | Calculating the fundamental trade-offs among three KPIs | ✓         | ✓         |
provided to make the problem tractable. In order to improve
the aforementioned trade-offs between these KPIs in future 5G
systems different candidate techniques are proposed.

The above approaches of frame/packet structure to achieve
low latency at the RAN level are tabulated in Table IV.

B. Advanced Multiple Access Techniques/Waveform

Different kinds of candidate multiple access (MA) tech-
niques and waveforms including orthogonal, non orthogonal
and asynchronous access have been proposed for low latency
communication [33], [47]–[49]. Since synchronization and
orthogonality (integral to OFDM) is a hindrance for achieving
low latency, asynchronous non orthogonal multiple access
techniques have been discussed in [33]. Reduction of symbol
duration to 67 μs is not a promising solution in critical time
budgeting. In this regard, interleave division multiple access
(IDMA) has been introduced in [81], [82] for generating
signal layers. The IDMA is a variant of the CDMA technique
which uses specific interleaving for user segregation in lieu of
using a spread sequence to the individual user. Here, channel
coding, forward error correction coding and spreading are
combined into a single block by a low rate encoder. The
spreading can not be considered as a distinct and special
task. Interleaving usually utilizes a simpler iterative multiuser
identification approach. However, this approach needs further
rigorous investigation.

In order to supplement synchronization and orthogonality,
sparse code multiple access (SCMA) and non orthogonal
multiple access (NOMA) have been presented in [83] for
5G scenarios. In SCMA, symbol mapping and spreading are
combined together, and the mapping of multi dimensional
codeword over incoming bits is performed directly from
SCMA codebook. SCMA is comparatively simpler and has
superior performance over low density version of CDMA.
Another modulation technique that aims to reduce latency,
is referred as the generalized frequency division multiplexing
(GFDM) is introduced in [68], [84]. The flexibility of covering
both the cyclic prefix OFDM (CP-OFDM) and single carrier
frequency domain equalization (SC-FDE), and block structure
of GFDM help to achieve low latency. A typical mapping
structure of GFDM, OFDM and SC-FDM is illustrated in
Fig. 7. The overall comparison among IDMA, SCMA and
GFDM is presented in Table V.

Filter bank multi carrier (FBMC) has been a strong candi-
date waveform for 5G [3], [85]. FBMC demonstrates better
performance in case of transmitting long sequences; however,
it suffers during the transmission of short bursts/frames. For
usage of cyclic prefix, wide frequency guards and more
required coordination, OFDM may be inefficient in case of low
latency communication [84]. Universal filtered multi-carrier
(UFMC) [84], [86] is upgraded version of FBMC which
offsets the disadvantage of FBMC. It outperforms OFDM by
about 10% in cases of time frequency efficiency, inter carrier
interference (ICI) and transmissions of long or short packets.
Additionally, UFMC preforms better than FBMC in the case
of very short packets while demonstrating similar performance
for long sequences. These make UFMC as the one of the best
choices for next generation low latency communication.

In case of UFMC, the time domain transmit vector [86] for
a user is superposition of sub-band wise filtered components.
The time domain transmit vector for a particular multi-carrier
symbol of user k with filter length L and FFT length N is

$$\mathbf{X}_k = \sum_{i=1}^{B} \mathbf{F}_{ik} \mathbf{V}_{ik} \mathbf{S}_{ik},$$

where

| Cases | IDMA [81], [82] | SCMA [83] | GFDM [68], [84] |
|-------|----------------|-----------|----------------|
| Fundamental concept/features | • Specific interleaving  
 • User segregation  
 • Iterative multiuser identification | • Multiple dimensional code word  
 • QAM spreading combination | • Block frame consists of time slots and subcarriers  
 • Non-orthogonal  
 • FFT/IFFT implementation |
| Low complexity | ✓ | ✓ | | |
| Flexibility (in case of covering CP-OFDM and SC-FDE) | ✓ | | ✓ |
| Low latency | ✓ | ✓ |
TABLE VI: Waveform Contender for 5G.

| Cases                        | OFDM                                                                 | FBMC [7], [8] | UFMC [84], [85] |
|------------------------------|----------------------------------------------------------------------|---------------|------------------|
| Filtering                    | Generalized filtering to all subcarriers of entire band              | Filtering to each subcarriers | Generalized filtering to a group of consecutive subcarriers |
| Requirement of coordination  | Higher                                                                | Lower         | Lower            |
| Time-frequency efficiency    |                                                                         |               |                  |
| (due to CP and guard band)   | 0.84                                                                  | 1             | 1                |
| ICI (in case of lower degree of synchronization with UEs and eNBs) | Higher                                                                | Lower         | Lower            |
| Performance                  | Performs well for large packets with well coordination                | Performs well for large packets with less coordination | Performs well for short packets with less coordination |

- S is the complex QAM symbol vector;
- V is the transformed time domain vector by IDFT matrix;
  In this case, the relevant columns of the inverse Fourier matrix are incorporated in accordance with the respective subband position within entire available band;
- i is the index of each subband of B;
- F is a Toeplitz matrix. It is comprised of filter impulse response, and performs the linear convolution.

The main advantage of UMFC is the ability of using different subcarrier spacings or filter times for users in different subbands. If a user uses FFT size $N_1$ and filter length $L_1$, and another user uses filter length and FFT size of $N_2$ and $L_2$ respectively, then UMFC symbol duration can be designed such that $N_1 + L_1 - 1 = N_2 + L_2 - 1$. This makes UMFC a remarkable adaptive modulation scheme with capability to be tailored easily under various characteristics of communications, including delay/Doppler spread variations in the radio channel and user QoS needs. The comparative discussion among OFDM, FBMC and UMFC is presented in Table VII.

C. Modulation and Channel Coding

Although use of small packets is a potential approach for achieving low latency, appropriate modulation and coding is required for small packet transmission for acceptable reliability. In the literature, mainly three types of coding schemes are proposed for 5G. As presented in [50], low-density parity-check (LDPC) and polar codes outperform turbo codes in terms of small packets while for medium and large packets, the opposite is true. While small packet is a requirement for low latency, other aspects such as implementation complexity, performance in practical test, and flexibility need to be investigated. In [51], polar code has been tested in field for 5G considering various scenarios: air interface, frame structure, settings for large and small packets, OFDM, and filtered OFDM (f-OFDM) waveforms. In all cases, polar code performed better than turbo codes which makes it a candidate channel coding scheme for 5G. The comparison among the schemes are illustrated in Table VII.

In [52], a highly-parallel architecture for the latency sensitive turbo decoding is proposed by combining two parallel algorithms: the traditional sliding window algorithm and cross parallel window (CPW) algorithm. New IFFT design with butterfly operation is proposed in [53], which reduces IFFT output data delay through the reduction of IFFT memory size and butterfly operation (e.g. addition/subtraction). Input signal of the IFFT processor corresponding to guard band is assigned as zero (i.e. '0') revealing the existence of numerous zeros. If the sequence of OFDM symbol data which enter the IFFT is adjusted, the memory depth can be reduced from 1024 to 176.

A dynamic shrunken square searching (DSSS) algorithm is proposed in [54], which cuts off unnecessary communication control port (CCP) calculation by utilizing both the noise characteristic and state space structure. In this way, it can maintain close to optimal decoding performance in terms of the block error rate (BLER). This results in reduction of delay in communication. In [55], a latency reduction approach by introducing time division multiplexing (TDM) of higher priority ultra-low latency data over other less time critical services is proposed, which maps higher priority user data during the beginning of a subframe followed by the normal data. In [56], balanced truncation is applied for the model reduction in the linear systems that are being coupled over arbitrary graphs under communication latency constraints. In [57], recent advances in finite-block length information theory are utilized in order to demonstrate optimal design for wireless systems under strict constraints such as low latency and high reliability. For a given set of constraints such as bandwidth, latency, and reliability the bounds for the number of the bits that can be transmitted for an OFDM system is derived.

| Cases                          | Turbo coding [52] | LDPC-PEG [50] | Convolutional coding [50] | Polar codes [51] |
|-------------------------------|-------------------|---------------|---------------------------|------------------|
| Algorithm complexity for coding 1/3 of 40 bits with respect to turbo codes | 100%              | 98%           | 66.7%                     | 1.5%             |
| Algorithm complexity for coding 1/3 of 200 bits with respect to turbo codes | 100%              | 98%           | 66.7%                     | 110.7%           |
| Performance in short packets  | ✔                 |               | ✔                         |                  |
| Performance in medium packets | ✔                 | ✔             |                           |                  |

TABLE VII: Comparison Among Channel Coding Schemes for Low Latency [50].
D. Transmission

A representative set of approaches for reducing latency using transmission side processing are tabulated in Table VIII which will be overviewed in the rest of this subsection.

In [58], an asymmetric window is proposed instead of well-known symmetric windows for reduction of cyclic prefix by 30%, and hence reducing latency due to reduced overhead. This technique suppresses out of bound (OOB) emission but makes the system more susceptible to channel induced inter symbol interference (ISI) and inter carrier interference (ICI). Transmission power optimization by the steepest descent algorithm considering transmission delay, error probability and queuing delay is proposed in [59]. In [60], low latency packet transport system with a quick path-switching and a packet-recovery method is introduced for a multi-radio-access technology (multi-RAT) environment. In [35], use of diversity gain is proposed as a solution for capacity enhancement and latency reduction. Diversity could be achieved through various approaches such as spatial diversity, time diversity, and frequency diversity.

In [87], a mmWave based switched architecture system is proposed where control signals use low-resolution digital beamforming (to enable multiplexing of small control packets) with analog beamforming in the data plane (to enable higher order modulation). This reduces the overhead significantly due to the control signaling which results in more resources for data transmission. This technique leads to reduction of round trip latency in the physical layer.

Recent advancements in full duplex (FD) communication comes forward with feature of doubling the capacity, improving the feedback, and latency mechanism meanwhile upholding steady physical layer security [88]–[91]. Various proposed techniques of 5G networks such as massive MIMO and beamforming technology providing reduced spatial domain interference can be contributive for FD realization [89]. Besides that intelligent scheduling of throughput/delay critical packets along with proper rate adaption and power assignment can results in capacity gain and reduction of latency. However, this field needs to be extensively investigated for studying capacity and latency trade offs.

E. Control Channel

When the packet size is reduced as envisioned in 5G systems, control overhead takes the major portion of the packet. Addressing this, various approaches are proposed in order to reduce the control channel overhead. The potential solutions targeting the control channel enhancements to achieve low latency are illustrated in Table IX.

In [61], control channel sparse encoding (CCSE) is introduced with vision to transmit the control information by means of non-orthogonal spreading sequences. A scaled-LTE frame structure is proposed in [62], assuming the scaling factor to be 5 with dedicated UL control channels (CCHs) for all sporadic-traffic users in each TTI with possible smallest scheduling request (SR) size. In [92], short TTI based uplink frame has been proposed for achieving E2E latency no longer than 1 ms. In the proposed scheme, subslot consisting of 2 symbols has been proposed for uplink data and control channel. A sequence-based sPUCCH (SS-PUCCH) incorporating two single carrier-frequency division multiple access (SC-FDMA) symbols is introduced in [63] in order to meet a strict latency requirement. Symbol-level frequency hopping technique is employed to achieve frequency diversity gain and reliability enhancement.

In the proposed procedure of [64], establishment of radio bearer and S1 bearer in parallel are proposed where eNB and mobility management element (MME) manages and controls radio bearer and S1 bearer, respectively. The eNB sends only single control signal in order to configure radio bearers such as SRB1, SRB2 and DRBs, that decreases the signaling interaction rounds between the UE and the eNBs. In [65], a new outer-loop link adaptation (OLLA) scheme is proposed. The scheme controls the size of the compensation in the estimated SINR based on the time elapsed after a UE transits from an idle state to an active state, which helps to reduce latency for small packet applications. The study [93] proposed a slotted TTI based radio resource management for LTE-A and 5G in order to achieve low latency. The approach can serve low latency services utilizing short TTI and enhance download control channel (ePDCCH).

The study [94] proposed a novel mechanism that introduces an adaptive radio link control (RLC) mode which dynamically alternates between unacknowledgment mode (UM) and acknowledgment mode (AM) according to the real-time analysis of radio conditions. This technique reduces system latency and processing power, and improves throughput using UM. On the other hand, it improves data reliability by activating AM during the degraded radio conditions. In [95], SDN based control plane optimizing strategy is presented to balance the

### Table VIII: Overview of Solutions in Transmission Side for Low Latency.

| Focus            | Reference | Techniques                                      |
|------------------|-----------|-------------------------------------------------|
| Transmission     | [58]      | Asymmetric window                               |
|                  | [59]      | Transmission power optimization                 |
|                  | [61]      | Path-switching and packet-recovery method       |
|                  | [62]      | Diversity gain                                  |
|                  | [63]      | Beam forming using mmWave                       |
|                  | [87]–[91]| Full duplex communication in same channel       |

### Table IX: Overview of Solutions in Control Channel for Low Latency.

| Focus            | Reference | Techniques                                              |
|------------------|-----------|---------------------------------------------------------|
| Control channel  | [61]      | Control channel sparse encoding (CCSE)                 |
|                  | [62]      | Dedicated UL CCHs                                       |
|                  | [63]      | Sub slotted data and control channel                   |
|                  | [64]      | SS-PUCCH consists of SC-FDMA symbol                     |
|                  | [65]      | Radio bearer and S1 bearer management                  |
|                  | [66]      | Outer-loop link adaptation                              |
|                  | [67]      | Slotted TTI based radio resource management            |
|                  | [68]      | Adaptive radio link control (RLC)                      |
|                  | [69]      | SDN based control plane optimization                  |
latency requirement of vehicular ad hoc network (VANET), and the cost on radio networks. The interaction between vehicles and controller is formulated and analyzed as a two-stage Stackelberg game followed by optimal rebating strategy, which provides reduced latency compared to other control plane structures.

F. Symbol Detection

As illustrated in Fig. 8, symbol detection encompasses various processes such as channel estimation and decoding, which can all contribute into the overall latency. The related literature in the symbol detection side for latency reduction are tabulated in Table X.

In [66], a low-complexity and low-latency massive SM-MIMO detection scheme is introduced and validated using SDR platforms. The low complexity detection scheme is proposed with a combination of zero forcing (ZF) and maximum-ratio-combining-zero-forcing (MRC-ZF). In [67], a linear minimum mean square error (MMSE) receiver is presented for low latency wireless communications using ultra-small packets. The estimation receiver filter using the received samples is proposed during the data transmission period in lieu of interference training period. Additionally, soft decision-directed channel estimation is argued using the data symbols for re-estimation of the channels. In [68], space-time encoding is introduced within a GFDM block in order to achieve transmit diversity for overall low latency in the system. On the other hand, a widely linear estimator is used to decode the GFDM block at the receiver end, which yields significant improvements in terms of symbol error rate and latency over earlier works.

In [69]–[71], compressed sensing is proposed for latency reduction in networked control systems if the state vector can be modeled as sparse in some representation domain. In [72], a low complexity receiver design is proposed and the superiority of an SCMA system is verified via simulations. In addition, it is demonstrated with a real-time prototype that the whole system throughput triples while maintaining low latency similar to flexible orthogonal transmissions.

G. mmWave Communications

Carrier aggregation using the mmWave spectrum (as illustrated in Fig. 9) is widely considered to be a promising candidate technology for 5G, capable of providing massive bandwidth and ultra low latency. The mmWave technology is especially critical for VR/AR type of applications which require high throughput and low latency. The works in mmWave spectrum for achieving low latency is summarized in Table XI.

In [46], a new frame design for mmWave MAC layer is introduced which provides several improvements including adaptable and smaller transmission intervals, dynamic locations for control signals, and the capability of directional multiplexing for control signals (dynamic HARQ placement). It addresses ultra low latency along with the multiple users, short bursty traffic and beam forming architecture constraints. The study [73] focuses on three critical higher-layer design areas: low latency core network architecture, flexible MAC layer, and congestion control. Possible solutions to achieve improvements in these critical design areas are short symbol periods, flexible TTI, low-power digital beam forming for control, and low latency mmWave MAC, which can all be

---

TABLE X: Overview of Solutions in Detection Side for Low Latency.

| Focus                | Reference | Technique/Approach                      |
|----------------------|-----------|----------------------------------------|
| Symbol detection     | [66]      | SM-MIMO detection with ZF and MRC-ZF    |
|                      | [67]      | Linear MMSE receiver                    |
|                      | [68]      | Space-time encoding and widely linear estimator |
|                      | [69]–[71] | Compressed sensing                      |
|                      | [72]      | Low complexity receiver in SCMA system  |

TABLE XI: Overview of Solutions in mmWave Communications for Low Latency.

| Focus                  | References | Techniques                                           |
|------------------------|------------|------------------------------------------------------|
| mmWave                 | [46]       | mmWave based MAC layer frame structure               |
|                        | [72]       | Low latency core network architecture, flexible MAC layer, and congestion control |
|                        | [74]       | mmWave based physical layer air interface with basic numerology and logical channel arrangement |
|                        | [75]       | Low latency frame structure with beam tracking       |
considered for data channel, downlink control channel, and uplink control channel.

In [74], in order to decrease the latency of the system, two different physical layer numerologies are proposed. The first approach is applicable for indoor or line of sight (LOS) communications, and the second one is suitable for non line of sight (NLOS) communications. This is justified by some channel measurements experiments in 28 – 73 GHz range. In [75], a 5G mmWPoC system is employed to evaluate the throughput functionality in field tests at up to 20 km/h mobile speed in an outdoor LOS environment. Additionally, some improvements for a frame design is obtained which decrease the latency in the field tests. In the experiments, it is observed that the new slotted frame design can decrease the RTT to 3 ms for 70% – 80% of the cases in experiments, alongside the observed throughput up to 1 Gbps.

H. Location-Aware Communications for 5G Networks

Location knowledge (in particular, the communication link distance) can be considered as a criterion of received power, interference level, and link quality in a wireless network. Therefore, overhead and delays can be reduced with location-aware resource allocation techniques because of the possibility of channel quality prediction beyond traditional time scales. The literature on location-aware communications regarding low latency are tabulated in Table XII.

In [77], several approaches are presented for monolithic location aware 5G devices in order to identify corresponding signal processing challenges, and describe how location data should be employed across the protocol stack from a big picture perspective. Moreover, this work also presents several open challenges and research directions that should be solved before 5G technologies employ mmWave to achieve the performance gains in terms of latency, connectivity and throughput. In [50], 5G flexible TDD is proposed for local area (5GETLA) radio interface with FFT size of 256 and 512, and short frame structure to achieve latency lower than 1 ms. The packets of size of less than 50 kbits can be transmitted with end-to-end latency of 0.25 ms. The main focus in designing physical layer parameters is on FFT, frame duration and physical channel (LA).

In [40], a novel numerology and radio interface architecture is presented for local area system by flexible TDD, and frame design. The proposed framework ensures coexistence with overlay LTE-A network, sleeping modes, contention based data channel, and channel quality indicator and interference statistics. Here, the channel quality and traffic statistics are accumulated from the small cells which can help to gain high throughput and low latency. Especially, in order to reduce the latency, the delay due to packets containing critical data for the higher layer protocols, for instance transmission control protocol (TCP) acknowledgment (ACK) packets, must be optimized. To do so, one possible approach is to carry out the retransmissions as quick as possible compared to the higher layer timers. Moreover, capability of data transmission to a contention based data channel (CBDCH) can play a key role here. As a result, by introducing CBDCH in small cells that are not highly loaded, the average latency of small packets transmission can be decreased considerably.

I. Reinforcing QoS and QoE

As shown in Fig. 10(a) and Fig. 10(b), respectively, reinforcing constraints on QoS and QoE can maintain low latency in 5G services including ultra high definition and 3D video content, real time gaming, and neurosurgery. The related literature on QoS and QoE control for low latency services are tabulated in Table XIII.

![Fig. 10: (a) Attributes of QoS, and (b) Attributes of QoE.](image-url)

*TABLE XII: LOCATION-AWARE COMMUNICATIONS FOR LOW LATENCY.*

| Focus | References | Techniques |
|-------|------------|------------|
| Location aware communications | [40] | Location information utilization in protocol stack |
| | [50] | Physical layer parameters design using FFT, frame duration and local area (LA) physical channel |
| | [77] | Utilization of channel quality and traffic statistics from small cell |
technologies can ensure to achieve the best QoS and QoE, as discussed in [97]. In [98], a QoS-aware multimedia scheduling approach is proposed using propagation analysis and proper countermeasure methods to meet the QoS requirements in the mmWave communications. Mean opinion score (MOS) which is a criterion for user satisfaction can be employed for functionality evaluation of the newly presented QoS approach and well-known distortion driven scheduling in different frequency ranges.

Client based QoS monitoring architecture is proposed in [99] to address the issue of QoS monitoring from server point of view. Different criteria such as bandwidth, error rate and signal strength are proposed with the well-known RTT delay for maintaining desirable QoS. A colored conflict graph is introduced in [100] to capture multiple interference and QoS aware approaches in order to take the advantage of beamforming antennas. In this case, reduction in call blocking and handoff failure helps to have a better QoS for multi class traffic. Each device can be sensitive to time based on its application. This can be considered as an issue for QoS provisioning. To address this, a novel QoS architecture is presented in [101] with heterogeneous statistical delay bound over a wireless coupling channel. The authors presented the dynamic energy efficient bandwidth allocation schemes in [102], which improve system quality significantly and maintain QoS.

Previous QoS criteria which consist of packet loss rate, network latency, peak signal-to-noise ratio (SNR) and RTT are not sufficient for streaming media on Internet, and therefore, users’ perceived satisfaction (i.e. QoE) needs to be addressed [103], [104]. Higher QoS may not ensure the satisfactory QoE. Different routing approaches of video streams in the mobile network operators’ scenario is discussed in [105] for substantial refinement in QoE considering bit rate streams, low jitter, reduced startup delay and smoother playback. A predictive model from empirical observations is presented in [106] to address interdependency formulated as a machine learning problem. Apart from that, a predictive model of user QoE for Internet video is proposed in [107].

### TABLE XIII: LITERATURE OVERVIEW RELATED TO QoS/QoE REINFORCEMENT.

| Aspect       | Reference | Techniques/Approaches                                          |
|--------------|-----------|----------------------------------------------------------------|
| QoS reinforcing | [96]      | mmWave utilization with beam tracking                           |
|              | [97]      | SDN and cloud technology                                        |
|              | [98]      | QoS-aware multimedia scheduling                                 |
|              | [99]      | Client based QoS monitoring architecture                        |
|              | [100]     | Colored conflict graph                                          |
|              | [101]     | QoS architecture with heterogeneous statistical delay bound     |
|              | [102]     | Dynamic energy efficient bandwidth allocation scheme            |
| QoE reinforcing | [103], [104] | Emphasis on users’ perceived satisfaction (i.e. QoE)             |
|              | [105]     | Routing using proximity information                              |
|              | [106]     | Predictive model based on empirical observations                |
|              | [107]     | Predictive scheme of user QoE for Internet streaming             |

### J. Other Aspects

Cloud radio access network (CRAN) (as illustrated in Fig. 11) is introduced for 5G in order to reduce the capital expenditure (CAPEX) and simplify the network management. CRANs combine baseband processing units of a group of base stations into a central server retaining radio front end at the cell sides. However, this requires connection links with delay of 250 µs to support 5G low latency services. In order to meet strict latency requirements in CRAN, two optimization techniques including (i) fine-tuned real-time kernel for processing latency and (ii) docker with data plane development kit (DPDK) for networking latency have been proposed in [117]. The experimental results clearly demonstrate the effectiveness of the approaches for latency optimization. In [118], split of PHY and MAC layer in a CRAN with Ethernet fronthaul is proposed, and verified through experimental test followed by latency interpretation. It is found that latency for packets of size 70 and 982 bytes is 107.32 µs and 128.18 µs confirming 20% latency increase from small to large packet. The promising results affirm that latency critical services in 5G can be supported by CRAN.

The CRAN can utilize backhaul information to redistribute users for QoE maximization and adaptation of temporal backhaul constraints. Corresponding to this, authors in [119] proposed a centralized optimization scheme to control the cell range extension offset so as to minimize the average network packet delay. In [76], a CRAN on the basis of the optical network (PON) architecture is presented which is called virtualized-CRAN (V-CRAN). The proposed scheme can dynamically affiliate each radio unit (RU) to a digital unit (DU) which results in coordination of multiple RUs with their corresponding DU. Moreover, definition of virtualized BS (V-
BS) is brought up which is able to mutually send shared signals from several RUs to a user. V-CRAN can reduce latency for joint transmission due to the following reasons. First, it can provide more than enough bandwidth for data transmission between RUs and DU. Furthermore, for each DU, a dedicated hardware/software is assigned that can be utilized by joint transmission controller in order to provide data and signaling for RUs. The last but not least, in order to handle the load distribution between DUs, the virtualized PON can connect a DU directly via a linecard (LC).

Before starting the packet transmission in data applications, a tolerable initial delay can be considered. Thus, this short delay can be employed to decrease the energy required to operate the small cell base stations (SBSs). With the goal of average power consumption reduction of SBSs, when the load is low, under-utilized SBSs can be are switched off. As a result, the users and the network can be able to save energy by postponing transmissions. By doing so, the users have time to wait for an SBS with better link quality. It can be observed that sleeping mode operation for energy efficiency improvements in SBSs will also introduce a source of latency. In [122], the energy-efficiency versus delay trade-off is investigated and optimality conditions for UE’s transmit power is derived. By postponing the access of the users, energy efficiency of the system can be improved. The optimal threshold distance is derived in order to minimize the average distance between a user and an SBS. Simulation results demonstrate that the energy consumption of SBS can be reduced by about 35% if some of the SBSs are switched off.

In [123], authors introduced an energy efficient, low-complexity technique for load-based sleep mode optimization in densely deployed 5G small cell networks. By defining a new analytic model, the distribution of the traffic load of a small cell is characterized using Gamma distribution. It is shown that the network throughput can be improved significantly while some amount of energy is saved by taking the benefit of the initial delay. In [124], impact of average sleeping time of BS and association radius on the mean delay in an UDN is investigated using a M/G/1/N queuing model. An explicit equation of delay is derived and the effect of average sleeping time and association radius on the mean delay is analyzed.

VII. CORE NETWORK SOLUTIONS FOR LOW LATENCY

To meet the vision of 5G encompassing ultra low latency in addition to enhancements in the RAN, drastic changes are also proposed in the core network. The new core network includes some new entities such as SDN, MEC, and NFV as well as new backhaul techniques [131], [132]. These enhancements

| Case/Area | Reference | Approach | Summary |
|-----------|-----------|----------|---------|
| Core Network Architecture | [29], [108], [110], [111], [112], [113], [115], [116] | SDN based architecture | The architecture of 5G is proposed based on SDN with 5G vision to meet large throughput, massive connectivity and low latency. |
| | | NFV based architecture | NFV invalidates the dependency on hardware platform and makes easy deployment of EPC functions as well as the sharing of resources in the RAN. This can reduce the end to end latency with improved throughput performance. |
| | | MEC/fog based network | MEC/fog provides computation and storage near user end and also separates the data plan from control plan. These reduces the latency. |
aim to reduce the processing time, bypass several protocol layers, and ensure seamless operation. The core network solutions for low latency are reviewed in further detail in the rest of this section.

A. Core Network Entities

The existing literature on the core network entities that can facilitate to achieve low latency are summarized in Table XIV. Exemplary architectures of SDN and NFV of the 5G core network are illustrated in Fig. [12].

In [29], authors presented an LTE compliant architecture to decrease delay for combination of fog networks, MEC and SDN in which the architecture is supposed to take the advantage of NFV in the evolved packet core (EPC) functions. Following this, optimization of general packet radio service (GPRS) tunneling protocol (GTP) is introduced for supporting low latency services. GTP tunnels management is accomplished by a novel element between the eNB and the mobile network interface with the Internet. In [108], [109], SDN is proposed along with some changes in the existing 4G architecture for moving forward towards 5G. The changes include reduction of number of serving gateways (S-GW) and elimination of some protocol layers. In SDN based system, virtualization is possible, and routes can be optimized. This will allow handling of QoS by setting specific rules in the switches along the path. Network coding integrated with SDN is proposed in [110] for low latency and reduced packet-retransmission. Network coding can work as network router SDN is proposed in [110] for low latency and reduced packet-retransmission. Network coding integrated with SDN can be integrated with SDN, which provides seamless network operation and reduction in latency.

The NFV is proposed as a major entity of 5G core network in [112]-[116], [133]. NFV removes the dependency on the hardware platform and makes flexible deployment of EPC functions as well as sharing of resources in RAN. This can reduce the E2E latency with improved throughput performance. SDN and NFV based 5G architecture with enhanced programmability of the network fabric, decoupled network functionalities from hardware, separated control plane from data plane, and centralized network intelligence in the network controller is presented in [112]. In [113], an information centric scheme is presented in order to integrate the wireless network virtualization with information centric network (ICN). In this architecture, key components such as wireless network infrastructure, radio spectrum resource, virtual resources (including content-level slicing, network-level slicing, and flow-level slicing), and information centric wireless virtualization controller have been introduced. In [114], deployment of EPC support solutions as a Service (EPCaaS) based on a uniform edge cloud structure is urged using the concepts of NFV. Furthermore, the authors proposed an approach to share states over several data centers with availability of network encapsulation across different data centers even in presence of firewall/network encapsulation with dynamically selectable state portions and edge nodes. In [115], authors presented the optimization problem of composing, computing and networking virtual functions to select those nodes along the path that minimizes the overall latency (i.e., network and processing latency). The optimization problem is formulated as a resource constrained shortest path problem on an auxiliary layered graph followed by initial evaluation.

The proposed core network entities for 5G vision are summarized in Table XVI.

B. Backhaul Solutions

Backhaul between base stations and the core network carries the signaling and data from the core and the Internet. Due to the enormous number of small cells and macro cell base stations supporting 1000x capacity, massive connectivity and latency critical services in 5G, the capacity of backhaul is a bottleneck for achieving low latency. At current scenario, microwave, copper and optical fiber links are used for backhaul connections based on availability and requirements. 5G backhaul requires higher capacity, lower latency, synchronization, security, and resiliency [13]. An exemplary 5G backhaul is illustrated in Fig. [13] in which the small cell base stations are connected by fronthaul whereas the macro cell base stations are connected by backhaul.

Referring to Table XV, we divide existing backhaul solutions into 2 parts: (1) General backhaul and (2) mmWave backhaul. The solutions are described as follows.

1) General backhaul: General backhaul includes a dynamic GPRS tunneling protocol (GTP) termination mechanism that combines cloud based GTP with a quick GTP tunnel proposed in [111]. Based on the user request or other factors, the system can change its mode from a cloud-based GTP tunnel to the quick GTP tunnel. In [29], a 5G vision compliant architecture is presented to reduce latency combining with the fog network, MEC and SDN. The optimization of GTP tunnels is accomplished by a novel element acting as an intermediate node between eNB and the mobile network interface accompanied with the Internet. In [125], modified

| Category            | Reference | Approach/Techniques                                                                                                                                 |
|--------------------|-----------|---------------------------------------------------------------------------------------------------------------------------------------------------|
| General backhaul   | [111]     | A dynamic GTP termination scheme combining cloud based GTP with a quick GTP tunnel with a dedicated hardware.                                      |
|                    | [29]      | GTP tunnel optimization by a new component in 5G compliant network consists of fog networks, MEC, and SDN.                                        |
|                    | [125]     | Modified VLC technology to set up an OW link for low-cost back haulng of small cells.                                                              |
|                    | [126]     | PON-based architecture with a tailored dynamic bandwidth allocation algorithm.                                                                        |
| mmWave backhaul    | [127]     | mmWave for fronthaul and backhaul, and split of control and user plane.                                                                          |
|                    | [128]     | mmWave based architecture, with a dynamic phase-shifter network based hybrid precoding/combining scheme for mmWave massive MIMO.                     |
|                    | [129]     | A framework supporting of in-band, point-to-multipoint, non-Line-of-sight and mmWave backhaul.                                                        |
|                    | [130]     | A mmWave based backhaul frame structure in 3 - 10 GHz carrier frequencies.                                                                          |
TABLE XVI: PROPOSED CORE NETWORK ENTITIES FOR 5G VISIONS.

| Reference | SDN | NFV | MEC | Fog networks | CRAN |
|-----------|-----|-----|-----|--------------|------|
| [29]      | ✔   |     | ✔   | ✔            |      |
| [108–111] | ✔   |     | ✔   | ✔            |      |
| [112, 114, 116, 133] | ✔   |     | ✔   | ✔            |      |
| [115]     | ✔   | ✔   |     | ✔            |      |
| [116]     | ✔   | ✔   |     | ✔            |      |
| [117]     | ✔   | ✔   | ✔   | ✔            |      |
| [76]      | ✔   | ✔   | ✔   | ✔            | ✔   |

VLC technology is used to set up an optical window (OW) link for low-cost backhaul of small cells to achieve a latency of 10 ms. Moreover, using a next generation baseband chipset, end-to-end latency below 2 ms can be achieved. An efficient PON-based architecture is proposed in [126] that offers ultra-short latency for handovers by enhancing connectivity between neighboring cells. Additionally, the authors propose a tailored dynamic bandwidth allocation algorithm for a fast handover between eNBs, which are associated to the same or diverse optical network units.

2) mmWave Backhaul: In addition to the presented solutions for backhaul, mmWave employment in backhaul can be considered as a promising solution for latency reduction. In order to have the enhanced user experience, the BSs should be in touch with core network and all other BSs via a low latency backhaul [159]. In [127], the authors proposed a scheme that employed mmWave links as backhaul, fronthaul and access in which a new separation method between control and user plane is proposed for 5G cellular network. A reasonable split among control and user plane improve the user QoS by providing ubiquitous high data rates in mmWave SBS coverage.

In [128], to implement an ultra-dense network (UDN) for the future 5G network and providing high data rates, the need of a reliable, gigahertz bandwidth, and economical backhaul is emphasized. Since mmWave can be easily integrated with massive MIMO to improve link reliability, and can provide sufficient data rate for wireless backhaul, it is a promising candidate for such a scenario. Considering a massive MIMO scenario, a hybrid precoding approach is considered, in which
each BS can cover several SBSs with multiple streams for each SBS at the same time. In [129], the authors proposed a solution framework for supporting an in-band, point to multi-point, NLOS, mmWave backhaul in order to provide a cost-effective and low latency solution for wireless backhaul. It is shown that an in-band wireless backhaul for inter BS coordination is feasible while the cell access capacities are not affected considerably.

In [130], a frame design for mmWave communications is proposed for 5G SBS network radio interface in 3-10 GHz. For both of LOS and NLOS scenarios different frame designs are proposed, which have a frame duration of 0.1 ms and 0.05 ms, respectively, to achieve low latency. The proposed LOS structure can be assumed as a suitable solution for short distance indoor wireless access or in-band backhaul.

VIII. CACHING SOLUTIONS FOR LOW LATENCY

In addition to the shortage of the radio spectrum, the insufficient capacity of backhaul links can be considered as a bottleneck for low latency communication. The long delay can be due to the requests of too many users in peak-traffic hours. Thus, latency reduction is crucial for users’ QoS and QoE in the 5G networks. Caching and in a more general category, information centric networking, can be assumed as one of the promising candidate technologies to design a paradigm shift for latency reduction in next generation communication systems [11], [12].

In this section, referring to Table XVII we present a detailed overview of caching concepts for cellular network followed by fundamental limits and existing solutions.

A. Caching for cellular network

Let us consider, a scenario that a user requests content from a content library $F = \{f_1, f_2, \ldots, f_k\}$, where $k$ is the number of files. The files are sorted with popularity where $f_1$ and $f_k$ are the most and least popular files, respectively. The popularity of a requested file $l$ can be written [160] as

$$\phi_l = \frac{l^{-\gamma}}{\sum_{i=1}^{k} l^{-\gamma}}, \quad (7)$$

where $l \in \{1, 2, \ldots, k\}$ and $\gamma$ is the parameter for uneven distribution of popularity in $F$ which follows Zipf distribution. For $N$ eNBs $B = \{BS_1, BS_2, \ldots, BS_N\}$ with each eNB having capacity $C$, the probability of caching of file $f_i$ by an eNB can be obtained as

$$P_{i} = 1 - e^{-\rho \sigma \pi R^2}, \quad (8)$$

where $\rho$ is the spatial density of eNBs following a Poisson point process [161], [162], and $\sigma$ is the probability that file $f_i$ is cached within $B$. Then, the total probability of getting content from the eNB can be written as

$$P_{i} = \sum_{i=1}^{N} \phi_i P_{i}. \quad (9)$$

The probability of getting the content as in (9) is directly associated with the latency of downloading it, and hence,

**TABLE XVII: OVERVIEW OF LITERATURE IN CACHING.**

| Aspect                  | Reference | Summary                                                                 |
|-------------------------|-----------|-------------------------------------------------------------------------|
| Content caching         | [134]–[155] | Filling of appropriate data is investigated by diverse techniques employing time intervals in which the network is not congested. |
| Content delivery        | [135], [138]–[156] | Content delivery to requested users is presented by different approaches for reduction of latency. |
| Centralized caching     | [139], [142]–[149], [151]–[155] | Various centralized caching is investigated with assumption that a coordinator with access to almost all the information about the storage capacities of different BSs, the connectivity of the users and BSs, and etc. |
| Distributed caching     | [137]–[139], [150]–[153], [140]–[142], [154], [156] | Various aspects of distributed caching has been investigated in order to minimize the communication overhead among SBSs and the central scheduler. |
| Latency-Storage trade-off | [143]–[148], [157]–[158] | Fundamental trade off between storage and latency is investigated in radio networks complemented with cache-enabled nodes. |
TABLE XVIII: DIFFERENT TYPES OF CACHING SCHEMES FOR THE CELLULAR NETWORK.

| Cachings      | Reference | Description                                                                 |
|---------------|-----------|-----------------------------------------------------------------------------|
| Local caching | [163]     | When a UE wants to access a content, it first checks in itself. Once the    |
|               |           | content is confirmed in the local caching storage, it is accessed by the    |
|               |           | UE without any delay.                                                       |
| D2D caching   | [160]     | If the requested content is not found locally, user will seek it within the  |
|               |           | range of it’s D2D communication. If it is found nearby devices, it is       |
|               |           | delivered to the requester UE by D2D communication.                         |
| SBS caching   | [164]     | If the requested content is available in the local SBS, the content is      |
|               |           | delivered to the UE by the local SBS.                                      |
| MBS caching   | [160]     | If the content is not found in local caching storage, nearby devices or SBS  |
|               |           | caching, the content is delivered by MBS caching.                           |

TABLE XIX: DEFINITION OF THE METRICS USED FOR LATENCY EVALUATION IN CACHING SCHEMES.

| Cases     | Reference | Definition                                                                 |
|-----------|-----------|-----------------------------------------------------------------------------|
| NDT       | [143]     | Defined as asymptotic delivery delay per bit in the high-power, long-blocklength case. |
| DTB       | [148]     | Defined as the ratio between the duration of transmission in channel and the file size in bits for the very large file size regime. |
| FDT       | [147]     | Defined as the worst-case delivery latency for the real load at a rate described by the DoF of the channel. |

Effective caching strategies can help in significantly reducing latency in 5G networks.

The proposed caching schemes for mobile networks can be divided into 4 categories: (1) Local caching, (2) Device to device (D2D) caching, (3) SBS caching, and (4) Macro base station (MBS) caching. Each of these caching types can reduce the latency by providing the requested content for the users using a way other than bringing it from the core network using backhaul links. In fact, each user starts from the nearest source to look for its desired content and proceed until finding it in any of the proposed sources. The different types of caching for cellular network are illustrated in Fig. [14] followed by the summarized descriptions presented in Table XVIII.

B. Fundamental Latency-storage trade-off in Caching

There are several fundamental limits for caching in mobile networks including latency versus storage, memory versus rate [149], memory versus CSIT [150], storage versus maximum link load [151], and caching capacity versus delivery rate [152]. As defined in Table XIX from an information theoretic point of view, authors employed the metrics such as normalized delivery time (NDT), fractional delivery time (FDT), and delivery time per bit (DTB) for investigation of the latency storage trade-off in caching networks. In most of these works [137], [138], [140], [142], [150], [153], [156], for a given scenario, an upper bound or lower bound for the considered metric is derived in order to get useful insights of this trade-off. The summary of latency storage trade-off works is presented in Table XX.

The authors in [143] investigated the storage latency trade-off using a new metric called NDT. This metric measures the worst-case latency that can happen in a cache-aided wireless network divided by that of an ideal system with unlimited caching capability. Considering a general cache-aided wireless network, the lower bound for NDT is presented in terms of the ratio of the existing file memory at the edge node and the total size of files for both perfect channel state information (CSI) and imperfect CSI.

Authors in [144] employed NDT as well in order to characterize the trade-off between NDT and fronthaul/caching resources. Using this information-theoretic analysis of fog radio access networks, optimal caching front-haul transmission is obtained. In [145], the latency storage trade-off in a 3×3 wireless interference network is investigated while all transmitters and receivers are equipped with caches. Another metric called (FDT) is proposed in order to characterize the trade-off between latency and storage. This information theoretic performance metric is actually a refined version of the metric originally proposed in [143]. The FDT can reflect the load reduction as well. In a similar work [146] the well-known DoF metric is used which does not reflect the load reduction. Moreover, the proposed approach in [146] just considers the one-shot linear processing, but interference alignment scheme in [143] may require infinite symbol extension.

In [146], an optimization problem is designed to minimize the number of required communication blocks for content delivery. Then, a lower bound is proposed on the value of the objective function. Using the same metric, the authors in [147] investigated the fundamental trade-off for a cache-enabled MIMO system. Considering a scenario with a 3×3 MIMO system in which the nodes are enabled with several antennas, the trade-off between storage and latency is investigated. In addition to FDT and showing its optimality for some ranges of cache size, the model can consider the effect of real traffic load at a rate specified by the DoF of the channel.

In [148], a cellular network is considered with multiple SBSs with limited cache capacity in which there is interference among them. Here, another information theoretic metric based on delivery latency is defined as well in order to characterize the system performance as a function of SBS cache memory and capacity of backhaul links connected to SBS. Using this metric which is called DTB, the trade-off between latency and system resources is investigated. In [157], using NDT trade-off between storage and latency a distributed caching scenario in fog radio access networks is characterized. In the presented approach, a coded delivery scheme is proposed to minimize the latency for delivering user demands for two edge-nodes and arbitrary number of users. It is shown that using decentralized placement, the presented delivery approach can obtain a considerable performance improvement in comparison to the derived lower bound.

In [158], again NDT is employed to characterize the fundamental trade off between delivery latency and system architecture. Considering NDT as the criterion for latency evaluation of the system, some bounds on its value are proposed. In the light of such bounds, useful insights on the latency and storage trade off are obtained. It is demonstrated that in order to obtain...
the lowest delivery latency, cloud-based compressed precoding and edge-based interference management should be considered as two major techniques for optimal performance of cloud and caching resources in different cases. In [165], for a fog radio access network, the heterogeneous timeliness requests depending on application is considered while in the existing works the assumption is that all requests have identical latency for all files in the content library. The fundamental trade-off between the delivery latencies of different users’ requests is characterized using NDT. The minimization of the average delivery latency as a function of cache and fronthaul resources is investigated. There are some efforts in the literature on investigation of different challenges raised up in centralized cache placement problems. In [134], the cache placement problem was investigated in a scenario including SBSs, called helpers with weak backhaul links but large memory size. In experimental evaluation, it is shown that the proposed scheme can achieve a considerable performance improvement for the users at reasonable QoS levels. In [136], authors aim at minimizing the average download delay of wireless caching networks with respect to caching placement matrix. It is demonstrated that the backhaul propagation delay can affect the caching placement.

In some scenarios, it is more desirable to design the caching problems in a distributed approach. In [157], a distributed cache placement approach is proposed in order to minimize the average download delay while some constraint for BSs storage capacities are met. The formulated optimization problem which is NP-hard is solved using a belief propagation based distributed algorithm with low complexity. This optimization problem makes sense because there is a trade-off between latency and storage capacity in caching networks. The comparison between the performance of the proposed distributed scheme with that of the centralized algorithm in [134] is presented as well. In [138], a decentralized content placement caching scheme is presented. Although there is no coordination, the proposed scheme can attain a rate as good as the optimal centralized scheme proposed in [134]. In [170], two caching and delivery schemes are considered. The first one operates in a centralized manner, while the second one is based on decentralized caching. For both cases, the trade-off between coded multi-casting and spatial reuse is reflected by the code length.

In addition to the aforementioned literature, in [180], content caching, and content delivery schemes are proposed considering cooperation to address the explosive enhancements
of demand for mobile network applications. Defining the objective as minimizing the average downloading latency, it is demonstrated that the proposed content assignment and delivery policy scheme has a better performance in comparison to the previous known content caching schemes in terms of average downloading latency. A weighted optimization problem is formulated in [139] to minimize the traffic of backhaul and downlink while the constraints for cache memory size and bandwidth limitation for D2D communication is considered. It is shown in [156] that if latency awareness is considered in caching management, it is an effective approach to reduce the delivery time of latency sensitive applications, and the global delivery time of users. In the proposed model, two main advantages is claimed. First, it not only has a better performance in term of delivery time at the end-user, but also affects the link load reduction. Second, a faster convergence with respect to probabilistic caching is achieved. In [140], the effect of joint latency awareness and forwarding is investigated in a cache-enabled network. Authors proposed a scheme which is based on caching and forwarding strategies in order to improve E2E experienced latency by the UEs while there is no coordination among them.

In [141], a cooperative content caching approach between BSs in cache-enabled multi-cell network is considered. Due to the trade-off between storage and latency, cooperative caching optimization problem is designed in order to minimize the average delay while a constraint on the finite cache size at BSs is met. It is shown that cooperation among cells can considerably reduce delay in comparison to that of non-cooperative case. Moreover, the gains of the proposed scheme will be increased in more diverse and heavier load traffics. In [142], the aim of the work is to minimize the data transmission delay for the P2P caching system while considering the effect of cache size, all mobiles in a cell are considered as several P2P caching groups. Then, the problem is formulated as a stochastic optimization problem and solved using Markov decision process (MDP) to obtain the optimal solution.

In [181], a cooperative multicast-aware caching strategy is proposed for the BSs to decrease the average latency of content delivery in 5G cellular networks. The proposed scheme is carefully designed in order to take into account the benefit of multicast and cooperation while in the existing caching schemes, the popular content simply is brought close to the users. The optimization problem is formulated in order to minimize average latency for all the content requests. It is demonstrated that via various trace-driven simulations that the proposed cooperative multicast-aware caching scheme can provide up to 13% decrease in the average content-access latency in comparison to multicast-aware caching scheme with the same total cache capacity.

In [182], the authors presented a cooperative caching architecture in which multiple locally cache-enabled nodes of cloudlets interact cooperatively in a decentralized cloud service networks. By proposing a content distribution strategy, the problem is formulated so that the mean total content delivery delay for all users in the proposed scheme is minimized. It is confirmed that the approach can enhance the cache hit rate, and also reduce the content delivery latency in comparison to existing solutions. In [183], the end-to-end packet transmission in a cache-enabled network is modeled in which both the wired backhaul and the RAN are jointly considered. The performances of both the on-peak and the off-peak network are investigated while both the wired backhaul and the RAN are considered. The E2E average packet latency is elaborated with the change of the request rate. It is shown that the average packet latency reduces in comparison to that of the system without caching ability due to the traffic offloading of the wired backhaul via caching.

IX. FIELD TESTS, TRIALS AND EXPERIMENTS

In this section, we present some representative field tests, trials and experiments for 5G low latency. The related literature is summarized in Table XXI, where each of the individual references will be described in further detail below.

The study [171] presents SDR based hardware platform to verify the concept of 5G. This facilitates initial proof-of-concepts (PoC) of novel 5G air interface and other concepts by extending hardware-in-the-loop (HIL) experiments to small laboratory experiments and finally trials of outdoor tests.

| Reference | Evaluation methodology | SDR | DSP | mmWave | Conventional/Proprietary LTE | Remarks |
|-----------|------------------------|-----|-----|--------|-----------------------------|---------|
| [171]     | ✓                      | ✓   | ✓   | ✓      | Trials of 5G concepts along with a novel air interface |
| [172]     | ✓                      | ✓   | ✓   | ✓      | 4 test cases and 15 KPIs    |
| [173]     | ✓                      | ✓   | ✓   | ✓      | mmWave aggregation          |
| [174]     | ✓                      | ✓   | ✓   | ✓      | Minimum latency µs          |
| [175]     | ✓                      | ✓   | ✓   | ✓      | Low latency VANET           |
| [176]     | ✓                      | ✓   | ✓   | ✓      | Round-trip latency ≤ 2µs    |
| [177]     | ✓                      | ✓   | ✓   | ✓      | 20 times latency reduction in comparison to existing works |
| [178]     | ✓                      | ✓   | ✓   | ✓      | Minimum latency 3 ms        |
| [179]     | ✓                      | ✓   | ✓   | ✓      | Latency ≤ 1 or 2 ms         |
| [172]     | ✓                      | ✓   | ✓   | ✓      | Latency ≤ 17 ms             |
| [178]     | ✓                      | ✓   | ✓   | ✓      | HARQ RTT ≤ 1.5 ms           |
| [42]      | ✓                      | ✓   | ✓   | ✓      | RTT latency ≤ 1 ms          |
Such an SDR based hardware can demonstrate high-capacity, low latency and coverage capabilities of LTE-A solutions. In [172], evaluation methodology including some novel test environments and certain key performance indicators are discussed in order to evaluate 5G network. Here, four candidate test environments such as indoor isolated environment and high speed train environment, and fifteen key performance indicators such as latency, throughput, network energy efficiency and device connection density are emphasized for performance evaluation.

In [173], 5G system operating at 15 GHz is presented followed by some experimental results. Here 0.2 ms subframe (14 OFDM symbols) is used for throughput, latency and other performance evaluation. The hardware implementation results of digital signal processing (DSP) and SDR based 5G system for low latency is presented in [174]. In this study, both the short TTI (sTTI) frame structures and wider subcarrier spacings are implemented in DSP platform. Based on the configurations of the system, RTT latency as low as 1 ms can be achieved. However, for achieving latency on the order of a few µs, optimization in between controllers and processing machines needs to be performed by cross-layer fashion. Additionally, the tail latency is argued to be considered in strict latency requirements assessment while maintaining required reliability.

An SDR based test bed is presented in [175] for cooperative automated driving with some experimental results from lab measurements. It implements flexible air interface consisting of re-configurable frame structure with fast-feedback, new pulse shaped OFDM (P-OFDM) waveform, low latency multiple-access scheme and robust hybrid synchronization, which ensure low latency high reliable communication. Results of the experimental trials are presented in [176], which utilizes DSP techniques for channel aggregation and de-aggregation, adjacent channel leakage ratio reduction, frequency-domain windowing, and synchronous transmission of I/Q waveforms and code words used in control and management function. In the proposed experiment, transmission of 48 chunks of 20 MHz LTE signals using a common public radio interface of capacity 59 Gb/s can achieve RTT DSP latency of less than 2 µs and mean error-vector magnitude of about 2.5% after fronthaul fiber communication. This mobile fronthaul technique shows the path towards ultra low latency integrated fiber/wireless access networks.

In [66], a multi-terminal massive SM-MIMO system is evaluated considering realistic scenarios. The authors developed a massive SM-MIMO OFDM system prototype utilizing multiple off-the-shelf SDR modules which serve as IoT terminals. Two linear detection schemes with diverse complexity levels were tested for detection instead of maximum likelihood detection (MLD) schemes. It demonstrates the similar real-time SINR performance of the MLD techniques along with 20 times latency reduction over existing works. The promising results urge the utilization of massive SM-MIMO systems for latency reduction and reliability enhancement in IoT transmissions. In [25], the performance of a lower latency frame structure was evaluated in field tests using a 5G mmWave proof-of-concept (PoC) system. It was found that the slot interleaving frame structure can achieve RTT latency of 3 ms in the 70 − 80% of the trial course. Additionally, beam tracking and throughput performance were evaluated in field tests at a speed up to 20 km/h on LOS outdoor environment. It was confirmed that mmWave system can obtain throughput of 1 Gbps in the 38% of the trial time at 20 km/h speed.

In [72], a low complexity receiver design was introduced followed by verification of superiority of an SCMA system via simulations and real-time prototyping. It can provide up to 300% overloading that triples the whole system throughput while still enjoying the link performance close to orthogonal transmissions. In [178], the concept of MEC was introduced first time for 5G followed by promising field tests. The MEC was tested and analyzed on various cases including local breakout and network end-to-end latency. It was concluded that MEC can support low latency services of not lower than 17 ms. It also urged that stricter requirement of latency needs to be investigated from the new radio technologies or D2D communication. In [179], a lab trial is presented to study the feasibility of ultra-low latency for 5G. It is shown that 1.5 ms HARQ RTT for TDD downlink in a lab trial is achievable when using the available test equipment in the literature while a novel frame structure and the associated signaling procedure is employed. The proposed scheme has 5 times better latency performance in comparison to the existing LTE-Advanced standard.

In [22], a novel frame structure is tested using a proprietary quasi-static system simulator for ultra-dense 5G outdoor RANs. In this regard, a frame structure is designed in order to facilitate low latency and multiuser spatial multiplexing on radio interface along with small-scale packet transmission and mobility support. It is found that performance of the introduced 5G network is better than that of LTE in case of air interface latency. In particular, considering UL scheduling requests in the RTT latency, the proposed frame structure can achieve latency as low as 0.8365 ms which is reduced by a factor of 5 in comparison to that of LTE. This satisfies 5G latency requirement (i.e.,1 ms latency).

X. Conclusion

Along with very large capacity, massive connection density, and ultra high reliability, 5G networks will need to support ultra low latency. The low latency will enable new services such as VR/AR, tele-medicine and tele-surgery; in some cases, latency not more than 1 ms is critical. To achieve this low latency, drastic changes in multiple network domains need to be addressed. In this paper, an extensive survey on different approaches in order to achieve low latency in 5G networks is presented. Different approaches are reviewed in the domain of RAN, core network and caching for achieving low latency. In the domain of RAN techniques, we have studied short frame/packets, new waveform designs, multiple access techniques, modulation and coding schemes, control channel approaches, symbol detection methods, transmission techniques, mmWave aggregation, cloud RAN, reinforcing
QoS and QoE, and location awareness communication as different aspects of facilitating low latency.

On the other hand, SDN, NFV and MEC/fog network architectures along with high speed backhaul are reviewed in the literature for core network with vision to meet the low latency requirements of 5G. The new core network will provide diverse advantages such as distributed network functionality, independence of software platform from hardware platform, and separation of data plane from software plane, which will all help in latency reduction. In caching, distributed and centralized caching with various trade-offs, cache placement and content delivery have been proposed for latency reduction in content download. Following this, promising results from field tests, trials and experiments have also been presented here. However, more practical and efficient techniques in the presence of existing solutions need to be investigated before the standardization of 5G. The authors believe that this survey will serve as a valuable resource for latency reduction for the emerging 5G cellular networks and beyond.

APPENDIX

Definitions of the acronyms used in this survey article are listed in Table XXII.

ACKNOWLEDGMENT

The work was supported by the National Science Foundation entitled “Towards Secure Networked Cyber-Physical Systems: A Theoretic Framework with Bounded Rationality” under the grant number 1446570.

REFERENCES

[1] S. Zhang, X. Xu, Y. Wu, and L. Lu, “5G: Towards energy-efficient, low-latency and high-reliable communications networks,” in Proc. IEEE Int. Conf. on Comm. Syst. (ICCS), Nov 2014, pp. 197–201.
[2] 5 Things Worth Knowing About 5G. [Online]. Available: [http://wi360.blogspot.com/2015/05/5-things-worth-knowing-about-5g.html]
[3] M. Agiwal, A. Roy, and N. Saxena, “Next Generation 5G Wireless Networks: A Comprehensive Survey,” IEEE Commun. Surv. Tutor., vol. 18, no. 3, pp. 1617–1655, thirdquarter 2016.
[4] A. Gupta and R. K. Jha, “A Survey of 5G Network: Architecture and Emerging Technologies,” IEEE Access, vol. 3, pp. 1206–1232, 2015.
[5] ITU-R, “Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond,” Feb. 2015.
[6] A. Kumbhar, F. Koohifar, I. Guvenc, and B. Mueller, “A Survey on Legacy and Emerging Technologies for Public Safety Communications,” IEEE Commun. Surv. Tutor., vol. 19, no. 1, pp. 97–124, Firstquarter 2017.
[7] K. I. Pedersen, F. Frederiksen, G. Berardinelli, and P. E. Mogensen, “The Coverage-Latency-Capacity Dilemma for TDD Wide Area Operation and Related 5G Solutions,” in Proc. IEEE Veh. Technol. Conf. (VTC Spring), May 2016, pp. 1–5.
[8] P. K. Agaypong, M. Iwamura, D. Staehle, W. Kiess, and A. Benjebbour, “Design considerations for a 5G network architecture,” IEEE Commun. Mag., vol. 52, no. 11, pp. 65–75, Nov. 2014.
[9] V. G. Nguyen, A. Brunstrom, K. J. Grinnemo, and J. Taheri, “SDN/NFV-based Mobile Packet Core Network Architectures: A Survey,” IEEE Commun. Surv. Tutor., vol. pp. 00, no. 1, pp. 1–1, 2017.
[10] T. Taleb, K. Samdanis, B. Mada, H. Flinkc, S. Dutta, and D. Sabella, “On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Architecture Orchestration,” IEEE Commun. Surv. Tutor., vol. PP, no. 99, pp. 1–1, 2017.
[11] A. Ioannou and S. Weber, “A Survey of Caching Policies and Forwarding Mechanisms in Information-Centric Networking,” IEEE Commun. Surv. Tutor., vol. 18, no. 4, pp. 2847–2886, Fourthquarter 2016.
[12] M. Zhang, H. Luo, and H. Zhang, “A survey of caching mechanisms in information-centric networking,” IEEE Commun. Surv. Tutor., vol. 17, no. 3, pp. 1473–1499, thirdquarter 2015.
[13] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, “5G Backhaul Challenges and Emerging Research Directions: A Survey,” IEEE
