Low-Latency Networking: Where Latency Lurks and How to Tame It

Low-latency Networking: Where Latency Lurks and How to Tame It

Xiaolin Jiang, Hossein S. Ghadikolaei, *Student Member, IEEE,* Gabor Fodor, *Senior Member, IEEE,* Eytan Modiano, *Fellow, IEEE,* Zhibo Pang, *Senior Member, IEEE,* Michele Zorzi, *Fellow, IEEE,* and Carlo Fischione *Member, IEEE*

***Abstract*—While the current generation of mobile and fixed communication networks has been standardized for mobile broadband services, the next generation is driven by the vision of the Internet of Things and mission critical communication services requiring latency in the order of milliseconds or sub-milliseconds. However, these new stringent requirements have a large technical impact on the design of all layers of the communication protocol stack. The cross layer interactions are complex due to the multiple design principles and technologies that contribute to the layers' design and fundamental performance limitations. We will be able to develop low-latency networks only if we address the problem of these complex interactions from the new point of view of sub-milliseconds latency. In this article, we propose a holistic analysis and classification of the main design principles and enabling technologies that will make it possible to deploy low-latency wireless communication networks. We argue that these design principles and enabling technologies must be carefully orchestrated to meet the stringent requirements and to manage the inherent trade-offs between low latency and traditional performance metrics. We also review currently ongoing standardization activities in prominent standards associations, and discuss open problems for future research.**

***Index Terms*—Internet of Things, low-latency communications, ultra-reliable communications, mission critical services.**

I. Introduction

A series of major technological revolutions have pushed the development of communication networks to the current state-of-the-art that includes Internet, pervasive broadband wireless access and low data rate Internet of Things (IoT). One of the first crucial steps of this series of revolutions is the global Public Switched Telephone Network (PSTN), the aggregate of the world's nationwide circuit switched telephone networks, which was designed to deliver arguably the most demanded and revenue generating services in the history of communication networks, namely circuit switched voice. As noted in [1], the PSTN not only ensures the latency requirements imposed by voice communication services, but also responds to randomly fluctuating demands and failures by rerouting traffic and reallocating communication resources. Due to its ability to reliably meet human-centered latency requirements and deliver the popular voice service over very long distances, even in the presence of fluctuating traffic demands and link failures, the PSTN is a technology to which Mark Weiser's observation truly applies:

> "The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it." [2], [3]

The second step in the communication networks revolutions has made PSTN indistinguishable from our everyday life. Such step is the Global System for Mobile (GSM) communication standards suite. In the beginning of the 2000, GSM has become the most widely spread mobile communications system, thanks to the support for users mobility, subscriber identity confidentiality, subscriber authentication as well as confidentiality of user traffic and signaling [4]. The PSTN and its extension via the GSM wireless access networks have been a tremendous success in terms of Weiser's vision, and also paved the way for new business models built around mobility, high reliability, and latency as required from the perspective of voice services.

The success of PSTN and GSM was ultimately due to the development of packet switching and wireless access methods. The circuit switching technology dedicates communication resources along the end-to-end path of pairs of transmitters and receivers on a coarse time scale, which leads to poor resource utilization. In contrast, with packet switching, a communication network is designed to share a single link between multiple pairs of transmitters and receivers. Although this idea boosts communication resource utilization, it may lead to packet congestion and consequent increase of latency. In fact, the Internet employs packet switching and Voice over IP (VoIP) technologies, and the interval between the delivery of voice and data packets at the receiving end is not deterministic: packets can arrive out of order, have variable latencies (jitter) or even get lost. To devise mechanisms that help VoIP systems to keep latency low and to compensate for jitter, VoIP endpoints use buffers to delay incoming packets so as to create a steady stream. For voice services delivered by VoIP technologies, the standard answer to the natural question "how much latency is too much?" is dictated by the ability of humans to adapt the cadence of a conversation, and is typically quantified as around 200-250 ms.

The boundary of 200-250 ms has been further pushed by the third and fourth steps of the communication network revolutions. As more powerful wireless access generations succeeded GSM, the 3rd Generation Partnership Project has designed the third and fourth generations of wireless cellular networks to meet 150 ms and subsequently an order of magnitude lower latency requirements in the wireless access part [5]. However, these third and fourth generations of wireless cellular networks were mostly driven by the need of high data rates and coverage. The requirements of low-latency for mission critical applications, such as tele-surgery, virtual reality over networks, autonomous control of vehicles and smart grids, were not the

concern of these generations.

The fifth step in the communication network revolutions has recently started. It arguably envisions a new form of proximity-aware networking or ubiquitous computing [3], [6]. The vision is founded on the implementation of a "wireless sense" using low-latency and highly reliable proximal communications via device-to-device and short-range communication technologies [7], thereby making the IoT a technological possibility. The "things", including sensors and actuators, smart meters, radio-frequency identification tags and other devices, are interconnected together either directly or through a gateway node and the global Internet. In such emerging IoT communication systems, low latency and high reliability in communicating messages will have a large impact on all elements of information and communication technology, ranging from mobile networks, consumer broadband, video, wireless sense-based proximal and cloud services. Ultimately, the fifth step of the communication network revolution is expected to ensure end-to-end communication latencies below 1 ms.

There are many use cases demanding below 1 ms low-latency communications. The development of information and communication technology for healthcare, industrial processes, transport services or entertainment applications, generates new business opportunities for network operators [8]. Part of this vision is grounded on the future availability of very low-latency communication networks to build the Tactile Internet that will extend touch and skills, and help realize real-time virtual and augmented reality experience [9]. The transmission of multi-sensorial signals, including the sense of touch (haptics), will contribute to the overall experience of real-time remote interactions. The healthcare industry is experimenting with remote diagnosis with haptic feedback, while remote robotic surgery with haptic feedback represents a potential future application of major impact for low latency communications. The transport sector is testing driver assistance applications and self-driving cars that will benefit from remote monitoring. The entertainment industry expects that immersive entertainment and online gaming incorporating Augmented reality (AR) will open new revenue streams, while the manufacturing industry expects significant productivity increase due to remote control with AR applications. Remote control applications can help improve the safety of personnel and reduce the cost of managing the on-site work force for hazardous environments such as mines or construction sites.

The use cases mentioned above, from the perspective of the wireless access network design, are expected to be addressed by the fifth step of the network revolutions, which is commonly called as 5^{th} Generation (5G). 5G is driven by the vision of the networked society [10], for which two generic communication modes of Machine-Type Communication will be supported [11]: Ultra Reliable Low-Latency Communication (URLLC) and massive Machine-Type Communication. URLLC is an innovative feature of 5G networks, as it will be used for a range of mission critical communication scenarios. Thereby, URLLC is expected to create near-term new business and service opportunities. Therefore, there is a great interest in technologies that will enable proximal communication links as well as global networks to operate with very low-latency (below milliseconds) while ensuring high reliability. As emphasized in [12], communication networks with latency below milliseconds will have a direct impact on network and proximal communication technology monetization, because latency performance is decisive for a number of revenue generating services, including high capacity cloud services, mission critical machine type communication services and high resolution video and streaming services. Indeed, the latency performance will be the determining factor between winning and losing business since meeting latency requirements directly impacts latency-sensitive, high capacity proximal as well as global Internet services.

Unfortunately, the vision of low-latency communication networks appears as a problem of formidable complexity. Ensuring that audio, visual and haptic feeds are sent with sufficiently low latency is a challenge, since the end-to-end route may incorporate multiple wireless access, local area and core network domains. All services delivered over proximal or long-distance communication networks are subject to latency, which is a function of several factors, such as link sharing, medium access control and networking technologies, competing service and traffic demands, or service-processing algorithms. Within a single network, there are several components that contribute to latency, such as at the physical, link and routing layers. If we try to ensure low latency only at one layer, we may have non-negligible latency components at other layers. Moreover, the optimization at a single layer may have undesired effects at other layers. To make the problem worse, complexity is not limited to individual networks. The many use cases demanding low latency as described above, use end-to-end connections that may be supported not only over a single network, but often over multi-domain networks. Here, "domain" refers to a part of the end-to-end communication network that is under the control of a network operator in terms of dimensioning and managing communication resources. Examples of multi-domain networks include the global PSTN, inter-networks consisting of multiple networks owned and operated by multiple Internet service providers, and cellular networks connecting cellular subscribers served by different mobile network operators. Such a complexity is largely unexplored.

In this paper we conduct an analysis and classification of the most prominent design principles and enabling technologies that are needed to meet the stringent requirements of low-latency networks. While it seems probable that such networks will be initially delivered in the proximity of communicating entities [3], [6], we also argue that there are strong business cases to deliver mission critical services even over long distances, such as for industrial remote operations, health care and intelligent transportation systems. Future low-latency services will be provided not only in limited geographical areas, but also over multi-domain networks. Therefore, it will be useful to analyze the latency and reliability requirements of the most important use cases for low latency and the sources of end-to-end latency for those cases. There is a need of substantial research, standardization and development of technology enablers that are applicable at the physical,

medium access control, network and transport layers in all segments of communication networks, including the access, core and service networks. Our analysis will greatly help in laying the foundation for developing technologies that can realize networks supporting services below 1 ms latency. Although all of these use cases require low latency, they may pose different level of reliability requirement. e.g., the reliability requirement of decentralized environmental notification messages for vehicular communications is more relaxed compared to that of Tactile Internet. More relaxed reliability requirements may render more options of the techniques to achieve the same level of latency performance.

The remainder of this article is structured as follows. Section II examines the latency requirements imposed by latency-critical applications and services. Then, based on this examination, the causes of latency components in single hop, multi-hop and multi-domain networks are examined and formally defined in Section III. Section IV surveys technology enablers applicable within a single domain, while Section V discusses inter-domain latency reduction and control techniques. Next, Section VI provides an overview of the most important related standardization activities. Section VII discusses open research questions, and Section VIII concludes the article.

II. Latency Requirements of Use Cases in Single Hop, Multi-Hop and Multi-Domain Networks

In this section, we analyze the latency requirements of future use cases of major societal impact, which will be enabled by the availability of networks capable to offer below 1 ms latencies. We argue that some of these use cases require a single domain network, whereas the rest should be supported by multi-domain networks.

The characterization of end-to-end latency and the identification of latency requirements associated with latency-critical applications and services are necessary steps to understand and compare promising technology enablers. We note that the definition of communication latency is not unique, but it depends on the use cases. Due to the stochastic nature of end-to-end latency, latency requirements may be typically specified in the form of stochastic measures, such as the cumulative distribution function or its moments [13], [14], and a probability of exceeding a predefined latency value. For mission critical and industrial process control systems, for example, the latency requirements can specify that a predefined latency value of a few milliseconds should be kept with probability 10^{-8} [13]. Alternatively, latency requirements may require that the expected value and the variance of the latency must remain under predefined thresholds [14]. The latency requirements of prominent latency-sensitive application scenarios are listed in Table I. In the following, we give a short description of each.

Smart Grid: Smart grid is proposed to solve the deficiencies in the old power system such poor controllability to the power generation utilities and slow response to the change of the power consumption [26]. Besides the power grid, the smart grid also has a control network with monitoring, communicating and computing abilities. This control network should support low-latency communications to enable mission critical applications such as substation automation and distributed energy resources [27]. Specifically, substation automation refers to monitoring, protection and control functions performed on substation and feeder equipments. The substation automations are extremely latency-sensitive as the time critical message will be irrelevant if not delivered within a specific time in the order of several millisecond. Distributed energy resources are small sources of power generation and/or storage that are located close to the load they serve and connected to the distribution grid. To integrate the distributed energy resources smoothly, automatic and remote monitoring, control, manipulation and coordination should be performed in real time. Protection and control traffics are the most latency-sensitive in smart grid applications, with latency requirements of 10 ms and 100 ms, respectively [22]. IEC 61850 is the international standard that ensures interoperability between the control networks of smart grids [28].

Industrial Automation: The forthcoming Industry 4.0 paradigm is expected to substantially boost interoperability in manufacturing by enabling machines, devices, sensors and people to connect and communicate with one another. Industrial manufacturing requires high reliability and stringent latency guarantees. By supervising the production activities, process automation aims to support more efficient and safe operation of industries such as paper, mining and cement [24], [25]. The latency requirements are related with the sampling times of different applications and are in the order of 100 ms to 1 s. Factory automation includes time-constrained operational applications, such as those used for motion control and certain power electronics applications, which requires the latency between 1-10 ms [23]. Ethernet-based solutions are gaining popularity in industrial communications, due to their capability of guaranteeing latency. [29]. The wireless solutions are also favored in many industrial scenarios with harsh environments.

Medical Applications: Tele-diagnosis and tele-surgery are promising trends for medical care. Not constrained by the geographical distances anymore, experienced surgeons will be able to diagnose or even perform surgeries along with audio-visual and haptic feedback by a robot [16]. Tele-monitoring is another emerging application that enables the experienced surgeon at the remote side to watch both the local surgeon to perform the surgery, and the patient's conditions, and provide real-time guidance and suggestions. These use cases require, in general, latencies of some milliseconds and at most 2% packet loss rates to realize real-time and reliable audio, visual, and haptic feedback [21].

Medical applications of low-latency networks go way beyond tele-surgery. The exoskeletons are new supportive protheses that enable to help aging people move independently, promote patient rehabilitation following an injury, and allow workers to carry heavier loads. Exoskeletons use sensors on the skin to detect voltage change of the signal at the muscles sent by the brain [19]. Prosthetic hand is another artificial device, designed for those that lost their own hand. Most of the existing prosthetic hands can only perform actions such as bracing and holding, which are control actions that can be performed with some hundreds of milliseconds' communication latency between the touching and the actuation. A prosthetic

TABLE I: THE REQUIREMENT FOR LOW-LATENCY APPLICATIONS (RELIABILITY MARKED WITH HIGH IS DUE TO NO AVAILABILITY OF PRECISE NUMBERS).

	Latency	Reliability	Other
Virtual reality [15]	1 ms	-	high data rate
Automated guided vehicle [16], [17]	few ms	99.99999%	high data rate
Financial market [18]	few ms	high	
Exoskeletons and Prosthetic hands [16], [19], [20]	few ms	high	-
Tele-surgery [16], [21]	1-10 ms	98%	high data rate
Protection traffic in smart grid [22]	1-10 ms	high	-
Factory automation [17], [23]	1-10 ms	99.9999999%	-
Control traffic in smart grid [22]	100 ms	high	-
Process automation [17], [24], [25]	100 ms-1s	99.9999999%	-

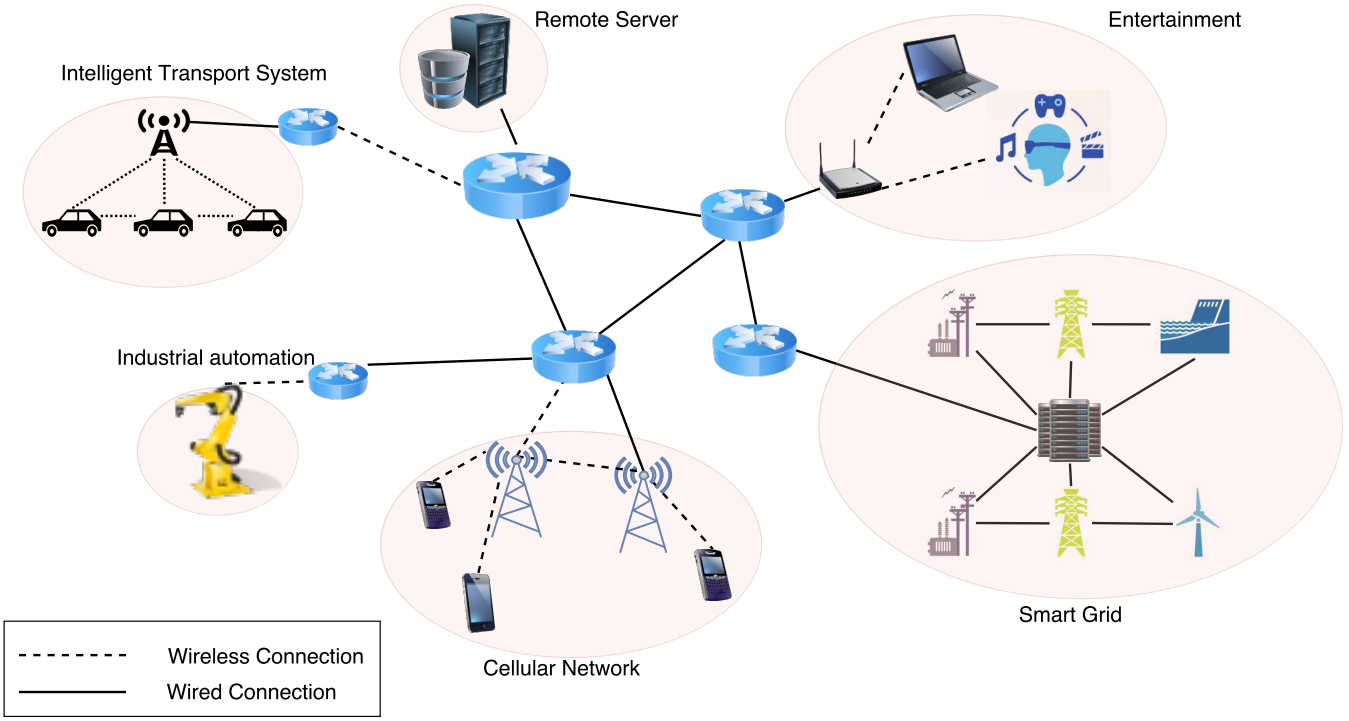


Fig. 1: General network architecture. The end devices are connected to base stations or access points, which further communicate with one another through router switches mostly by wired networks.

hand that enables the users to feel has been recently reported in [20]. The tactile and position sensors in such a prosthetic hand collect and send the data to a control unit that translates them into a neural code, which is then applied to the nerves in the user's arm. Meanwhile, the data is also processed to interpret the user's movement intention and send commands to the prosthetic hand. To provide the user with safety and a better experience wearing the exoskeletons and prosthetic hands, the latency should be no more than a few milliseconds.

Virtual Reality and Augmented Reality: Virtual reality (VR) and AR are revolutionary interfaces that provide unprecedented experiences and enable new applications. VR provides an immersive experience for a live concert, sports match or interactive game for users just sitting on the couch at home. AR augments reality by enabling better learning and working modes. For example, in a natural history museum, on top of the specimen of a dinosaur, some vivid three dimensional dinosaur projection provides a better conceptual learning environment. The real-time audio, visual, and haptic feedback by VR and AR requires latency from action to reaction below 1 ms to avoid nausea [15]. Within the world of musical instruments and music industry, the vision of Internet of Musical Instruments has recently been proposed. According to such a vision, any musical instrument will be connected in the future to Internet via wireless communications, provided that the end-to-end latency is around 5 ms [30]–[32].

Intelligent Transport Systems: Vehicular communications refer to the communication among vehicles that can improve driving safety, reduce traffic congestion and traffic accidents, improve fuel consumption efficiency (e.g., by platooning), support high quality entertainment, and ultimately enable driver-less cars [33]–[36]. AR can also be used in intelligent transport systems to create a bird view of the real-time traffic information. Vehicular cloud architectures are also proposed to support applications that require large computation and storage capacities [35]. To support all these functionalities, communication networks have to support latencies of only a few milliseconds [16]. To ensure safety, a reliability as high as 99.99999% is also needed [17]. Dedicated Short Range Communications based on IEEE 802.11p [37] and cellular vehicle-to-everything (C-V2X) communication are promising candidates to support vehicle-to-vehicle and vehicle-to-

roadside communication [38], [39].

Financial Markets: Time equals money in the financial market, and having low-latency between the placing of an order and its execution is essential to achieve the transaction at the desired price before the price changes. It has been stated that a 1-millisecond advantage in trading applications can be worth $100 million a year to a major brokerage firm [40]. The required latency in the financial trading environment is in the order of several milliseconds [18]. A high availability is also required to support a great number of online clients.

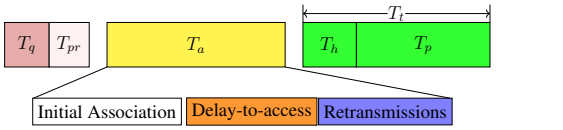
We conclude this section by noting that the major use cases mentioned above do not have a common definition of the latency requirement. Moreover, in some instances they require proximal communications, and in other instances they require remote communication services. In the next section, we will deepen the technical definition of latency and of delay components for proximal and remote communication services.

III. Causes and Definitions of Latency Components in Single Hop, Multi-Hop and Multi-Domain Networks

In the previous section we have seen that the use cases of major societal impact enabled by low-latency networks will have both proximal and remote communication sessions. As latency is a complex function of several factors, it is useful to examine the components of end-to-end latency in proximal communication scenarios, as well as over long distances involving wired and wireless access networks and multi-domain inter-networks, as illustrated by Fig. 1, which we explain shortly below. We will often refer to this figure in the rest of the paper, with the purpose of analyzing which design principle and which enabling technology are responsible for the various latency components.

In the generic multi-domain scenario depicted in Fig. 1, information messages are transported by several networks employing diverse technologies. For example, when streaming an online video on a cellular phone, the corresponding service request may first be sent through a single-hop wireless connection to the serving base station (BS). Subsequently, it may go through a series of wired/wireless connections in the backhaul network and a set of intermediate networks to an application server. As another example, communications among hundreds of sensors and actuators inside a vehicle and between vehicles and vulnerable road users and roadside infrastructure equipment are inherently local and mission critical. Although both examples impose latency requirements, it is clear that the requirements imposed by mission critical services can be order(s) of magnitude lower than quasi real-time entertainment services.

We can formally define latency as the time duration between the generation of a packet and its correct reception at the destination. This message may go through several networks as shown in Fig. 1, sometimes referred to as domains, which may not be handled by the same operator. It is difficult to guarantee a predefined total latency, as different networks may introduce random latencies whose exact values or even statistical distributions are typically not known a priori. In other words, even though the latency can be measured or estimated within a single network, meeting end-to-end latency guarantees remains a challenging task, since no single network operator or network entity has control over the end-to-end latency.



(a) Composition of single hop latency, e.g., from mobile phone to the BS.

(b) Latency components of multiple hops, e.g., from mobile phone to the remote server.

Fig. 2: The latency formation of single hop and multi-hop communication.

To gain a precise insight, consider Fig. 2, which illustrates the most important components of latency. Once a packet is generated, it is typically placed in a queue, and waits for transmission. Analogously, when a packet arrives at the receiver, it may be placed in a queue at a low layer of the protocol stack, waiting to be processed and delivered to higher layers. We define the sum of these queuing latency as T_q. To facilitate link sharing between multiple traffic classes, modern access and core networks provide differential treatments of packets, and implement Quality of Service (QoS) dependent packet handling, including, for example, QoS-aware scheduling and priority queueing mechanisms. When multiple QoS classes with different latency priorities are supported, the packets with higher priority have lower T_q. Similarly, the total processing time during the end-to-end communication can be conveniently characterized by the aggregate processing latency $T_{\rm pr}$. The processing latency $T_{\rm pr}$, different from T_q, is a function of physical, link layer and hardware technologies, node processing capacities and signal processing algorithms.

Apart from T_q and $T_{\rm pr}$, for a single hop, latency incorporates the time it takes for a packet to get access to the - typically shared - medium (denoted by T_a), including the time taken by technology-dependent control signaling. For example, successfully transmitting and receiving request-to-send and clear-to-send messages of the popular IEEE 802.11 protocol family or processing scheduling grant messages of the 3GPP Long Term Evolution protocol suite contribute to the latency over a single hop of an end-to-end path. Once the (wired or wireless) medium is accessed for the delivery of a packet, we need to account for the transmission of that packet, which typically includes a packet header and payload, amounting to a transmission time of $T_t = T_h + T_p$.

When the end-to-end path involves multiple hops – each of which is carried over dedicated or shared resources – the end-to-end latency is determined by the sum of the associated medium access (T_a) and packet transmission (T_t) times. In addition, for multihop communications, the time required for routing packets to the right outgoing interface (T_r) adds to the end-to-end latency, as illustrated in Fig. 2. The routing (T_r) latency may be zero when the packet is forwarded by an entity that does not perform routing, which is the case of

a wireless relay or other entity that operates at the Medium Access Control (MAC) layer.

Recall that the main sources of latency variations – in addition to unpredictable interference levels, random appearance of shadowing objects, and other factors appearing on the wireless interface – include the random traffic load along shared links and media as well as the variations of the fast fading wireless channels. These random factors make the end-to-end latency notoriously difficult to control and predict since the transmission of a tagged data stream is affected by the fluctuating load pattern of simultaneously delivered traffic streams [41]. In the following two sections, different techniques for intra networks and inter networks as shown in Fig. 3 will be investigated.

IV. Intra-network Techniques and Technologies

As illustrated in Fig. 2, the end-to-end latency is an aggregate of the latency components that are associated with queueing and processing ($T_q + T_{\rm pr}$), medium access (T_a), transmission ($T_t = T_h + T_p$) and routing (T_r). In low-latency networks, techniques at the physical, medium access control, network and transport layers will have to be designed to reduce these components (see Table II), also in a cross-layer collaboration. These techniques will have to be tailored to reduce a specific component or deal with a combination of them.

In this section, we discuss physical, medium access control, network and transport layer techniques that can be used to reduce latency, and at the same time we take into consideration constraints related to spectral efficiency, energy efficiency, peak data rate and capacity. Also, this section discusses technology enablers that are not specifically designed to deliver latency-critical services, but can be potentially used to meet latency targets.

A. Physical Layer Techniques

Physical layer (PHY) techniques are fundamental in striking a good engineering trade-off between latency, reliability, spectral and energy efficiency, and communication range. To quantify the effects of PHY techniques on latency, recall that, assuming a fixed-length packet, an higher transmission rate implies lower transmission time (T_t). In a single antenna system the maximum transmission rate between a transmitter and its intended receiver is upper bounded by

$$C = B \log_2(1 + \text{SINR}) \ \ [\text{bps}] \,, \tag{1}$$

where B is the bandwidth, and the instantaneous Signal-to-Interference-plus-Noise Ratio (SINR) is

$$\text{SINR} = \frac{|h|^2 P_t}{N_0 B + P_I} \,, \tag{2}$$

where h is the instantaneous channel gain, P_t is the transmit power, N_0 denotes the noise spectral density, and P_I is the received interference power.

Adaptive modulation and coding is a powerful technique to increase the spectral efficiency or decrease the Bit Error Rate (BER). Generally speaking, for sufficiently high SINR, higher modulation orders together with light coding schemes boosts the transmission rate and reduces T_t. However, when the channel becomes poor or the received interference is high, the transmitter should reduce the transmission rate (by adopting lower modulation orders or stronger coding schemes) to maintain a target BER [42].

From (1) it is straightforward to see that the bandwidth (B) acts as a multiplication factor in the transmission rate calculation, thus increasing the bandwidth is an effective way to increase the transmission rate and thereby to reduce latency. As B also affects SINR from (2), the capacity is not a linear function of the bandwidth. However, we focus on the range where capacity increment with bandwidth is not negligible, i.e., it has not yet saturated. Despite that this is the range in which the reliability is not very high, there could be an interesting trade-off among capacity, latency, and reliability. Taking Millimeter-Wave (mmWave) communication as an example, the abundant bandwidth at mmWave frequency bands enables it to achieve multi-gigabit transmission rates – even with a very low order of modulation – at the expense of lowering spectral efficiency.

1) Adaptive Modulation and Coding Schemes: The randomness of the fading wireless channel, $|h|$ in (2), causes the SINR at the receiver to fluctuate. However, when the transmitter is able to acquire channel state information (CSI), it can adapt its transmission parameters, which helps to achieve high data rates, which in turn helps to reduce T_t by adaptively setting the transmit power, modulation scheme, coding rate, or the combinations of these parameters. Indeed, as it was shown in [43], adaptive transmission schemes provide higher average transmission rates compared to non-adaptive transmission schemes. This is because in order for non-adaptive schemes to operate with acceptable BER, they need to be designed for the worst case channel conditions, and therefore operate with low spectral efficiency even when the channel condition is favorable. The seminal work in [43] presented optimal and suboptimal adaptation policies, including the total channel inversion scheme that adjusts the transmission power to maintain a constant received power, and truncated channel inversion scheme that only compensates for fading above a certain cutoff fade depth (see [44] for details). The CSI feedback of adaptive modulation and coding takes resources and increases the latency. However, the overhead by the CSI feedback may be compensated by the gain of the average transmission rate. For bi-directional communication with a packet exchange period shorter than the required CSI feedback interval, the CSI feedback can always be piggybacked in the packets sent back to the transmitter, further reducing the CSI overhead.

2) Waveform Design: Once a message (and the corresponding data packet) is mapped onto symbols with an appropriate modulation and coding scheme, selecting a proper waveform modulation scheme has a great impact on the transmission time. Indeed, recognizing the importance of waveform selection, the research and standardization community spent a large effort on analyzing the advantages and disadvantages of waveform candidates for the next generation of wireless systems [45].

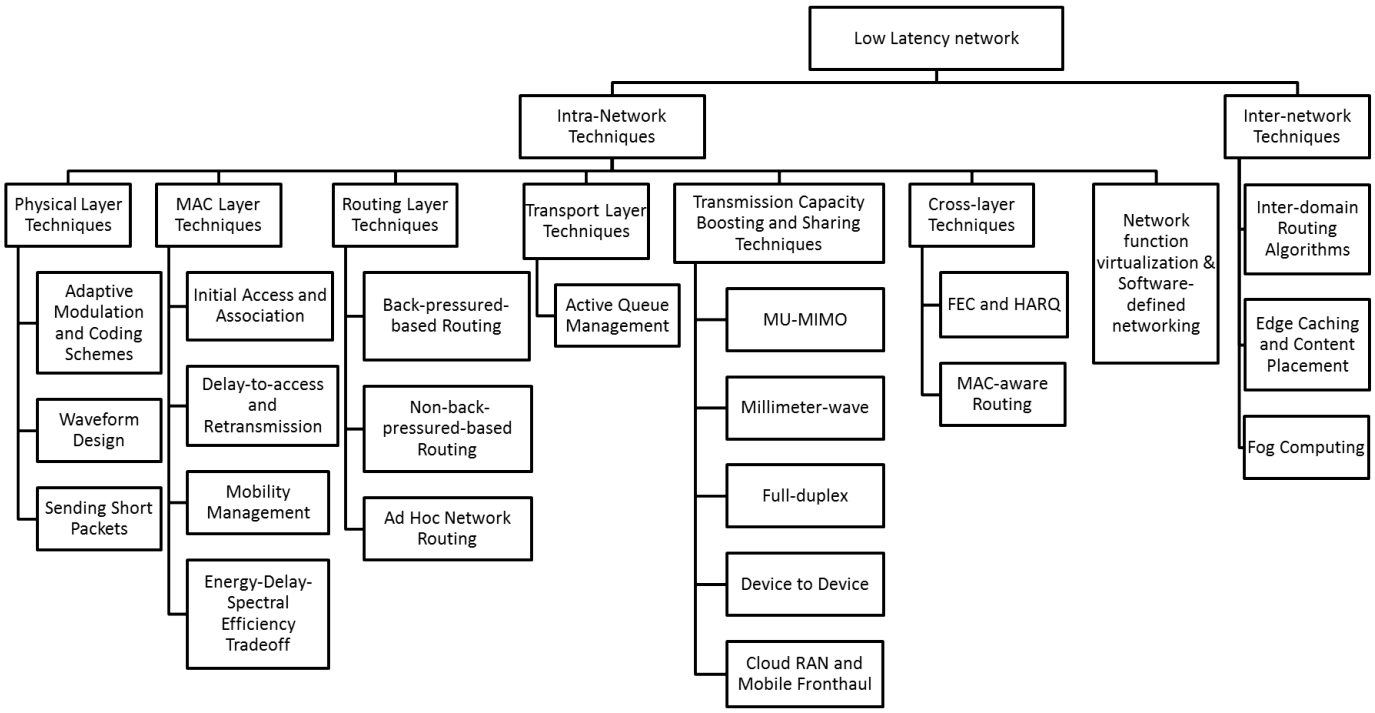


Fig. 3: Enabling technologies for low-latency communication, each of them may affect single or multiple delay components.

TABLE II: POTENTIAL METHODS TO ACHIEVE LOW-LATENCY IN A SINGLE NETWORK.

Methods	Transmission time T_t	Medium access latency T_a	Routing latency T_r	Queueing latency T_q	Processing latency T_{pr}
Adaptive modulation & coding	√				
Waveform design	√				√
Sending short packets	√				
MAC algorithms		√			
Routing algorithms			√		
Transport Layer algorithms				√	
Capacity boosting & sharing	√	√			
Device to device communication	√	√			
Cloud RAN & Mobile Fronthaul			√		√
FEC & HARQ	√	√			
MAC-aware Routing Algorithms		√	√		
NFV & SDN	√	√	√	√	√

Single-carrier (SC) and multi-carrier (MC) modulations are two categories of waveform modulations. With SC, each signal is spread over the whole available bandwidth. The transmission time for each signal is short, but without time guard protection, the delay spread in wireless channels may cause Inter Symbol Interference (ISI) to subsequent symbols. To combat ISI, MC modulation divides the available bandwidth into multiple narrow sub-carriers, and maps one modulated signal to each sub-carrier. The transmission time of each signal increases, making the delay spread relatively short and presenting a higher SINR at the receiver side. However, the overall transmission time of the payload of MC is comparable to that of SC.

Orthogonal Frequency Division Multiplexing (OFDM) is a widely used MC modulation technique employed by IEEE and 3GPP systems. It employs orthogonal sub-carriers to achieve bandwidth efficiency. A brief comparison between OFDM and SC in terms of latency is given below. First, the transmission rate of each sub-carrier in OFDM is smaller than that provided by SC modulation. However, the aggregate transmission rate over all sub-carriers is comparable to that of SC. In OFDM a Cyclic Prefix (CP) is used to increase robustness against ISI in multipath propagation environments. Although the overhead introduced by the CP, which must be greater than the maximum delay spread of the channel, is not negligible, it is still more economical than SC modulation where one guard time interval is needed for each symbol [46]. Apart from the high achievable spectral efficiency in multi-path propagation environments, OFDM has the flexibility of adaptively assigning different transmission powers and symbol constellations to each sub-carrier, which helps strike a good

balance between transmission rate and reliability in frequency selective environments. In terms of processing latency, OFDM leads to somewhat longer latency than block processing at both the transmitter and the receiver, but its equalization can be done efficiently on a sub-carrier basis using a simple one-tap equalizer. A more detailed analysis can be found in [47] and the references therein.

The orthogonality of the sub-carriers is essential for OFDM to avoid inter-carrier interference, which requires high precision synchronization between the transmitter and the receiver. Thus traditional OFDM with long synchronization is not ideal for meeting stringent latency requirements in a spectral efficient manner when transmitting short packets. The recently proposed Wireless Communication for High Performance aims to support latency in the order of μs for industrial control applications using short packets [48]. Wireless Communication for High Performance reduces the physical layer header of OFDM by taking advantage of the predictive and periodic traffic pattern of industrial applications.

The research community has developed new waveforms that help relax the strict synchronization requirement, e.g., Universal Filtered Multi-Carrier, Filter Bank Multi-Carrier, Generalized Frequency Division Multiplexing (GFDM) [45], and Filter-OFDM [49]. These waveforms are non-orthogonal and thus inter-carrier interference must be kept under control by using filters to suppress the out-of-band emission. Among the above asynchronous waveforms, GFDM is regarded as the most suitable for low-latency communication [50], [51]. Compared with OFDM's division only in the frequency dimension, GFDM has a block frame structure composed by M sub-symbols and K sub-carriers, and each sub-carrier is modulated and filtered individually.

For low-latency communications, GFDM acts as a comparable candidate to OFDM for several reasons. First, using GFDM may achieve lower T_h, as it requires less synchronization accuracy compared to OFDM. Second, the duration of the GFDM symbol is more flexible than in OFDM, and may achieve lower T_p. To combat ISI, a CP is appended to each OFDM symbol, while non-orthogonal GFDM needs a single CP to protect the information contained in M sub-symbols. To compare the T_p's of OFDM and GFDM, set the number of sub-carriers of OFDM N as the product of the numbers of sub-carriers and of sub-symbols ($N = M \cdot K$), which suggests that the sub-carrier spacing of GFDM is M times that of OFDM, and the sub-symbol duration of GFDM is $1/M$ that of OFDM. Given a CP duration, when mapping N bits of data, the T_p's of OFDM and GFDM are the same. However, if $N+1$ bits of data are sent, GFDM can easily be adapted to $(M+1)$ sub-symbols while still using a single CP. On the other hand, OFDM can change the sub-carrier size to K resulting in $(M+1)$ OFDM symbols, each having one CP. This is equivalent to M extra CP durations compared to GFDM. Third, distortion accumulation can grow without bounds in OFDM when orthogonality is not perfectly maintained, so OFDM must have proportional sub-carrier spacing to guarantee orthogonality [45]. GFDM inherently deals with non-orthogonality, and can therefore use non-proportional sub-carrier spacing, and non-continuous spectrum aggregation. Thus, GFDM can also be used to improve the transmission rate. On the negative side, GFDM requires high transmitter filter order to achieve sharp filter edge, which increases both complexity and processing latency. To summarize, the ideal waveform should be determined by the design criterion and operational environment, including the trade-off between reducing the transmission time and increasing complexity and processing latency.

3) Sending Short Packets: It is intuitive to think that short packets are needed for short transmission times in latency sensitive communications. The use of short packets brings a difference to the maximum coding rate and the packet error probability [52]. Specifically, for a given packet error probability and for a finite packet length n, the maximum coding rate reduces by a factor proportional to $1/\sqrt{n}$ [53]. The maximum coding rate for finite packet length with multiple antennas when considering the trade-off of diversity, multiplexing, and channel estimation is investigated in [54], [55]. The maximum coding rate affects the minimum packet transmission time T_t, and further affects the end-to-end latency.

With traditional long packets, the payload data is much longer than the metadata (control information), and thus the metadata is coded with a low rate to be robust. As the meta data accounts for only a small fraction of the whole packet, the overhead caused by the low coding rate can be neglected since it virtually does not affect the latency performance. With short packets, however, the size of meta data T_h is comparable to the size of payload data T_p. In this case, the overhead introduced by the meta data cannot be neglected, rendering coding metadata with a low rate inefficient. Therefore, to increase the spectral efficiency, metadata and payload should be coded jointly, the details of which is an open research issue [52].

B. MAC Layer Techniques

The MAC layer is responsible for synchronization, initial access, interference management, scheduling, and rate adaptation. While (1) governs the maximum achievable rate, inefficient initial access, queue management, and channel access strategies may substantially reduce the effective transmission rate of individual devices.

Define the one-hop access latency for every packet as the time from the instant the node starts sending that packet for the first time until the beginning of its successful transmission. It includes the processing and queuing latencies, which we do not cover in this paper, the association latency (only in the initial access phase), delay-to-access (latency until the start of the scheduled time slot in contention-free manner or that before the first transmission attempt in the contention-based manner), and the retransmission delay in the case of decoding failure. In the following, we review main techniques used to reduce these latency components. As we will see, this reduction comes (usually) at the price of a higher computational and signaling overheads, and also a penalty in energy and spectral efficiencies.

1) Initial Access and Association: Initial access and association are amongst the most important MAC layers functions that specify how a new device should connect to the network.

This is usually handled by a synchronization process and then a random access phase, by which the network registers the device as active. The latency caused by the association procedure may be tolerable when the devices have to be connected all the time (e.g., mobile phones), the data size is large (e.g., camera sensor networks), or the handover time is negligible; see also Section IV-B3. In many IoT applications with massive number of wake-up radios each having just a few bits of data, however, the association latency may become much longer than the data transmission time. Unfortunately, most of the currently used standards are not capable of supporting a low-latency initial access procedure: the initial access deadline of 10 ms for 3GPP [56] or 20 ms for ITU [57]. That is why the existing standards only consider "connected users" for low-latency services and assume "normal" initial access. Designing more efficient synchronization and initial access procedures for low-latency networks seems to be widely open areas.

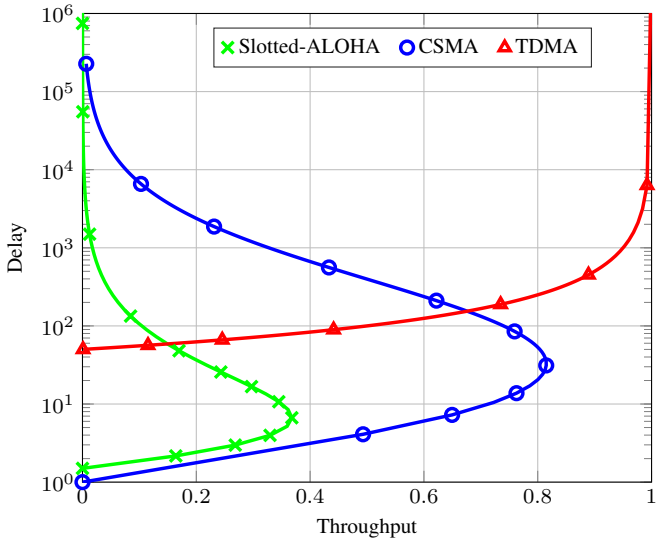
2) Delay-to-access and Retransmission: The MAC layer scheduling is responsible for the delay-to-access component. A MAC protocol is contention-free if messages do not collide during its execution, which is usually guaranteed by orthogonal communication resource allocation to different devices. The orthogonality of contention-free can be realized in the time domain (e.g., Time Division Multiple Access (TDMA)), in the frequency domain (e.g., Frequency Division Multiple Access (FDMA)), in the code domain (e.g., Code-Division Multiple Access (CDMA)), in the spatial domain (e.g., Multi-User Multiple Input Multiple Output (MU-MIMO)) or any combination of those domains. In contention-based MAC protocols, devices contend to access the channel, and as a result, some messages are lost due to inevitable strong interference (also called collision).

Contention-free MAC protocols can guarantee certain latency and jitter at the price of signaling and computational overheads, which may not be tolerable or efficient in the use cases of low-latency networks with a large number of devices and each just has a few bits of data. Consider the example of massive IoT devices with wake-up radios. After being associated to the network, the radios should send a channel access request to a local coordinator (e.g., AP or BS), wait for the channel access notification and follow the instruction to transmit their few bits of information. For the TDMA strategy applied to N devices, the average latency is $N/2 \times T_t$, where T_t is the time slot duration for each device. Clearly, this delay is not scalable with the number of devices. FDMA, CDMA, and MU-MIMO do not have this problem but the devices should still register their channel access requests and wait for the instruction (carrier frequencies in FDMA, codes in CDMA, and beamforming vectors in MU-MIMO), which could be problematic in massive wireless access scenarios [58]; see also Section IV-E.

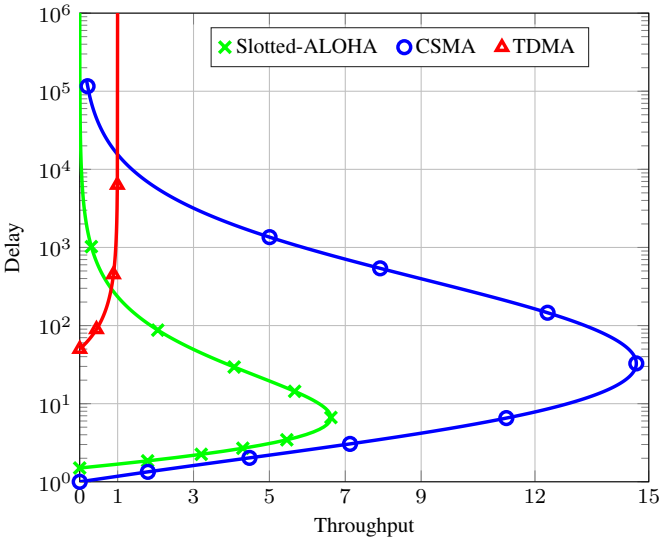
In contention-based MAC protocols, there is no controller that governs the scheduling. This class of protocols have low signaling and computational overheads, but they impose a random latency for channel establishment. Slotted-ALOHA is a very simple contention-based scheduling, in which a device tries to access the channel at the start of the next slot as soon as it gets a new packet [59]. In the case of collision, it retransmits the packet after a random backoff whose average increases after every new collision event. It has a good delay and throughput performance when the offered load (traffic arrived per unit time) is low. The devices do not need to have any signaling to reserve the channel, unlike the contention-free protocols. When the offered load increases, however, the delay and throughput performance of slotted-ALOHA degrades dramatically due to continuous backoff caused by repeated collisions.

Carrier-Sense Multiple Access (CSMA) protocol introduces channel assessment that allows each device to check whether the channel is idle before transmission. This technique, together with the random backoff procedure, substantially reduces the collision probability, leading to a better performance in high-load scenarios. Yang *et. al.* [60] studied the delay distribution of slotted-ALOHA and CSMA and showed good delay performance of both schemes in low traffic regimes. CSMA is known to have hidden and exposed node problems, which may cause collision and unnecessary deferred transmission [61]. Carrier-Sense Multiple Access with Collision Avoidance uses Request to Send and Clear to Send signals to reserve the channel in a distributed fashion so as to reduce collision probability. However, the extra signaling may reduce the throughput by as much as 30% [62] and increase the delay. This inefficiency becomes more prominent when we go to high data rate technologies like mmWave networks where the transmission rate of data signals are 100–1000x higher than that of control signals. In that case, the overhead of the collision avoidance mechanisms can drop the link performance to only 10% [63]. Fig. 4(a) shows the latency (due to delay-to-access and retransmission) performance of slotted-ALOHA, CSMA, and TDMA for a network of single antenna devices containing 100 transmitters. In all scenarios, increasing the input traffics increases both network throughput and average delay, almost linearly. Once the network operates close to its capacity, increasing the input traffic would not improve the network throughput but substantially increases the delay. In contention-based algorithms there is a critical value of the input traffic after which the network becomes congested, and adding more traffic decreases the network throughput and increases the delay exponentially. From this figure, slotted-ALOHA and CSMA are preferable for light traffics, Carrier-Sense Multiple Access with Collision Avoidance (not shown in this figure) for medium traffics, and TDMA for intensive traffics. Note that the overhead of TDMA channel access request is not shown in this figure.

When we deal with devices with multiple antennas, the traditional trade-offs and design constraints may change dramatically [64]–[68]. Directional communication alleviates the hidden and exposed node problems and reduces the interference footprint. The collision avoidance mechanism are less important and CSMA and slotted-ALOHA substantially outperform TDMA in terms of throughput and delay [64]. The analysis in [64] shows that directional communication is key to support low latency in massive wireless access scenarios. In particular, higher directionality levels reduce the need for complicated scheduling approaches with high signaling and computational overheads. Fig. 4(b) shows the delay performance of TDMA,



(a) Omnidirctional antenna



(b) Directional antenna

Fig. 4: Delay versus throughput of slotted-ALOHA, CSMA, and TDMA. Two numbers of users, 10 and 100, are set for TDMA. The unit of delay is the packet transmission time. (b) is based on [64]/

slotted-ALOHA, and CSMA for a network of multiple antenna devices. Now, slotted-ALOHA and CSMA are preferable for almost any traffic, given that obtaining beamforming vectors is not time consuming. Nitsche *et al.* [69] uses dual radio wherein one radio is responsible for obtaining directional information all the time to be fed to the other radio that is responsible for the data transmission. This technique decreases the delay to find proper beamforming vectors. Olfat *et al.* [70] relaxes the dual-radio requirement and develops a reinforcement-learning approach to design the optimal beamforming vectors with a minimal number of pilots. However, the research area of low-overhead beamforming is widely open and further research is required.

An advanced technique to utilize correlatable symbol sequences in place of the traditional control messages is proposed in 802.11ec [62]. Correlatable symbol sequences are predefined pseudo-noise binary codewords that retain the statistical properties of a sampled white noise, the cross correlation of which obtains spike value only with a matching copy. These sequences can be correctly decoded at very high transmission rates due to the robustness of the receiver. [63] extends the use of these sequences to the mmWave networks to address the imbalance transmission rates of the control and data planes, and improves the efficiency of distributed channel reservation for ad hoc mmWave networks. Interestingly, the extended approach can substantially alleviate the need for retransmission of collided channel reservation packets as different packets (of different types like Request to Send-Clear to Send or of the same type transmitted by different devices) can be recognized at the receiver even when they arrive at the same time. This technique proves to be suitable for low-latency communications in ad hoc networks. Grant-free communication schemes for short packets are a new approach that is proposed in [71], [72]. These schemes tolerate collisions by sending replica packets at the transmitter side and applying successive interference cancellation at the receiver side. In [73], the authors propose a frame-based contention-free protocol for star networks that exploits the cooperative multi-user diversity. Specifically, in each frame, the devices that have successfully sent their data, help the remaining unsuccessful devices by relaying their data toward the access point (AP). This multiuser diversity approach is also combined with network coding to further improve the reliability and delay performance [74].

3) Mobility Management: Mobility management and handover are essential procedures of a mobile network to maintain a connection. During the handover, the mobile terminal switches the set of its serving APs/BSs based on some performance measures (e.g., received signal strength indicator). When a terminal is capable of communicating to more than one AP at a time, soft handover eliminates the need for breaking radio links. For other technologies where no such capability is supported, the mobile terminal can be served in parallel by up to one AP and therefore has to break its communication with its current AP before establishing a connection with a new one. Traditionally, reducing the signaling overhead and algorithmic complexity of handover were the primary design concerns of almost all existing commercial systems.

References [75]–[77] proposed various algorithms for fast hard handover algorithms in cellular networks. The general idea of all these schemes is to anticipate handover events and perform some operations prior to the break of the radio link. More precisely, these references focused on mobile broadband wireless access (in particular IEEE 802.16e) and introduced different mobility management messages that enable receiving data packets during the handover process. Li *et al.* [78] generalized the idea of fast handover to mobile IPv6 to maintain the connectivity of any mobile node to the Internet with low latency. In this approach, previous AP and new AP will coordinate to speed up the re-association process of the mobile device to the new AP. Reference [79] provides a through review of the existing fast handover strategies for Wireless Local Area Network networks.

A common drawback of these approaches is that the prediction of the handover event is only based on the location of the mobile terminal. In a wireless network with directional communications, however, even fixed terminals may need to execute the handover procedure if an obstacle appears on the communication link. When handover becomes frequent, e.g., in mmWave networks [80], soft handover strategies seem to be a better option to support low-latency services. Semiari *et al.* [81] considered a dense network of dual-mode APs that integrate both mmWave and microwave frequencies. Each mobile terminal can cache their requested contents by exploiting high capacity of mmWave connectivity whenever available, and consequently avoid handover while passing cells with relatively small size. Xu *et al.* [82] introduced a distributed solution approach that reserves some portions of the capacity of each AP to ensure fast handover in mmWave networks. In particular, the proposed algorithm maintains two lists for every AP: "active" and "potential" terminals. In the case of handover event for potential terminals, the old and new AP coordinate to eliminate the registration latency and support seamless handover.

4) Energy-Delay-Spectral Efficiency Trade-offs: There are applications that impose less stringent latency requirements than URLLC use cases, but require long life-time. Sensors packaged in certain products and sensors implanted in the human body, for example, typically tolerate a couple of hundreds of milliseconds latency, but require a battery lieftime of years or decades. For such applications, the energy-delay-spectral efficiency trade-off arises as an important issue. Modern wireless devices are equipped with rate-adaptive capabilities which allow the transmitter to adjust the transmission rate over time. Associated with a rate, there is a corresponding power expenditure that is governed by the power-rate function. Specifically, such a function is a relationship that gives the amount of transmission power that would be required to transmit at a certain rate. For most encoding schemes, the required power is a convex function of the rate, which implies that transmitting data at a low rate and over a longer duration has less energy cost compared to a fast rate transmission. This fact, first observed in [83], induces a trade-off between transmission energy and delay. In particular, in [83] the problem of rate adaptation subject to deadline constraints on packets was studied, and it was shown that a "lazy" packet scheduling scheme that transmits at the lowest possible rate to meet the deadline is energy optimal. This idea was extended in [84] to general quality-of-service (QoS) constraints, where packets have different deadlines, and using techniques from network calculus [85], an energy-optimal rate adaptation policy was devised. Moreover, algorithms for minimum energy transmission over time-varying channels were developed in [86]. A comprehensive overview of energy-delay trade-offs in wireless networks can be found in [87].

C. Routing Layer Techniques

To ensure a desired end-to-end latency in a multihop network, the routing policy plays an important role in end-to-end latency. In this subsection, we first focus on the routing algorithms that are proposed for fixed topologies and then discuss those developed for dynamic and ad hoc networks.

1) Routing Algorithms Based on Back-pressure Scheduling: Throughput optimality is at the heart of most existing routing algorithms. Back-pressure (BP) routing algorithm [88] is amongst the most celebrated ones that is provably throughput optimal, but has quadratic end-to-end queuing delay in terms of the number of hops. When implementing the traditional back-pressure scheduling, each node keeps a separate queue for each flow. Then, in each time slot, each link checks the queues for all the flows passing through it, finds the flow with the maximum differential backlog, and uses this value as the weight for that link. At last, the transmission scheme with the highest weight under the interference constraints will be scheduled.

Bui *et al.* [89] modified the BP algorithm to make the end-to-end delay grow linearly, instead of quadratically, with the number of hops. The price is a minor throughput loss, which might be of secondary importance for low-latency networks. Their proposed algorithm uses the concept of shadow queues to allocate service rates to each flow on each link with a substantially lower number of real queues.

Another limitation of the BP routing algorithm is that every device in the network should be able to apply the BP algorithm. In a real network, however, we may be able to upgrade just a few routers. To address this problem, Jones *et al.* [90] considered an overlay architecture in which a subset of the nodes act as overlay nodes and perform the BP routing, while the other nodes can perform legacy routing algorithms such as shortest-path [91], [92]. Simulation results show that with a small portion of overlay nodes, the overlay network can achieve the full throughput, and outperforms the BP algorithm in terms of delay. A decentralized overlay routing algorithm was proposed to satisfy the average end-to-end delay for delay-sensitive traffic [93].

In a wireless network, the delay performance of any scheduling policy is largely affected by the mutual interference, caused by the concurrent transmissions. Gupta *et al.* [94] derived a lower bound for the average delay of a multi-hop wireless network with a fixed route between each source-destination pair. The authors proposed a variant of BP to optimize the energy-delay-spectral efficiency trade-offs and to perform close to the delay lower bound.

2) Non-back-pressure-based Routing: The BP routing algorithm and its variants utilize the queue length as the metric for the delay [88]–[90], [94]. To keep the queues stable and achieve throughput optimality, BP algorithms prioritize highly loaded queues. When a flow has light traffic, this queue length metric can lead to a huge delay as a short queue gets a small chance to be served. To address this problem, head-of-line delays can be used to act as the link weights instead of queue lengths [95], [96]. For a multihop network with a contention-based MAC protocol, [97] develops a new metric, called Q-metric, to jointly optimize the routing and MAC layer parameters with the objective of optimizing the energy-delay trade-off.

Recently, a new throughput optimal routing and scheduling algorithm, known as Universal Max-Weight (UMW), was

proposed in [98]. Unlike BP, UMW uses source routing, where each packet is routed along a shortest path and the cost of each edge is given by the queue backlog in a corresponding virtual network. Since UMW uses source routing, it can choose a "shortest-path" for each packet and avoid the looping problem that is inherit to BP. Consequently, UMW results in a significant delay reduction as compared to BP and its variants [98].

The surveyed routing algorithms need a fixed and known topology a priori. In an ad hoc network, however, the topology may be unknown and dynamic, and the cost of topology discovery may not be negligible when we consider low-latency services. In the following, we review some new classes of routing algorithms that suit ad hoc networks.

3) Ad Hoc Network Routing: The ad-hoc network is representative for scenarios without static structure, like vehicular networks. The seminal work of [99] shows how the delay scales with the number of devices in the network. For static random networks with n nodes, the optimal throughput per node scales as $T(n) = \Theta(1/\sqrt{n \log n})$, and the delay scales as $D(n) = \Theta(n/\sqrt{\log n})$[1]. The results suggest that there is a trade-off between the throughput and the delay by varying the transmission power, i.e., $T(n) = \Theta(D(n)/n)$. Higher transmission power increases the transmission range and can decrease the delay, due to the possibly reduced number of hops. However, it also increases the interference, leading to a drop in the throughput performance. Yi *et al.* [100] extended these results to the case of directional antennas and showed that, under ideal beamforming assumptions, the throughput gain scales with $\theta_t^{-1}\theta_r^{-1}$ where θ_t and θ_r are the antenna beamwidths of the transmitter and the receiver, receptively. Grossglauser and Tse [101] showed that mobility improves transmission range and achievable throughput at the price of higher delay. Specifically, the achievable throughput can be improved to $T(n) = \Theta(1)$, and delay scales as $D(n) = \Theta(n \log n)$. Note that these works do not consider the overhead of channel estimation, nor signaling and computational overheads, so that the actual delay can be larger.

The delay and capacity trade-off for multicast routing with mobility in ad hoc networks is studied in [102]. The results are based on 2-hop relay algorithms without and with redundancy (redundancy is used to denote transmitting redundant packets through multiple paths with different relay nodes). When a packet is sent to k destinations, the fundamental trade-off ratio is $D(n)/T(n) \geq \mathcal{O}(n \log k)$. The delay of the 2-hop relay algorithm with redundancy is less than that without redundancy, but will cause decrease to the capacity. More recently, in [103] the scaling of capacity and delay for broadcast transmission in a highly mobile cell partitioned network was studied. Using an independent mobility model as in [104] it was shown that, in a dense wireless network (number of nodes per cell increases with n), the broadcast capacity scales as $1/n$, while the delay scales as $\log \log n$. Surprisingly, it is also shown that both throughput and delay have worse scaling when the network is sparse.

[1]Notation: i) $f(n) = \mathcal{O}(g(n))$ means that there exists a constant c and an integer N such that $f(n) \leq cg(n)$ for $n > N$. ii) $f(n) = \Theta(g(n))$ means that $f(n) = \mathcal{O}(g(n))$ and $g(n) = \mathcal{O}(f(n))$.

SPEED [105] proposed a routing algorithm to support real-time communications via providing per-hop delay guarantees and bounding the number of hops from source to destination. SPEED is extended in [106] to satisfy multiple QoS levels measured in terms of reliability and latency. To this end, this approach employees a localized geographic packet forwarding augmented with dynamic compensation for local decision inaccuracies as a packet travels towards its destination. Therefore, there is no need for any global information or a centralized entity to design the routing, making this algorithm suitable for large and dynamic wireless sensor networks.

Zhang *et al.* [107] proposes a variation of the celebrated proactive distance vector routing algorithm [108] to guarantee end-to-end latency. In particular, each node establishes and maintains routing tables containing the distance (number of hops) and next hop along the shortest path (measured by the minimal number of hops) to every destination. To support a delay-sensitive service, the transmitter probes the destination along the shortest path to test its suitability. If this path meets the delay constraint, the destination returns an ACK packet to the source, which reserves link-layer resources along the path. Otherwise, the destination initiates a flood search by broadcasting a route-request packet. The flooding is controlled by the delay constraint, namely an intermediate node forwards the route-request packet only if the total delay to the destination is less than the threshold. When a copy of this packet reaches the source with a path that meets the delay constraint, the route discovery process is complete. [109] proposed a similar idea for mobile networks where the latency requirement will be met by reserving sufficient link-layer resources along the shortest path.

There are several survey papers on the QoS routing protocols for general ad hoc networks [110] and for vehicular networks [111], [112] where the topology is dynamic.

D. Transport Layer Techniques

The transport layer is responsible for the flow and congestion controls, and affects the queueing latency T_q. In general, the communication ends may belong to different network operators. However, when they both belong to the same network, we can optimize the transport layer for low-latency services. Transmission Control Protocol (TCP) and User Datagram Protocol are the two main protocols in this layer. TCP provides reliable end-to-end communications, independent of the underlying physical layer[2], while User Datagram Protocol does not have such a reliability guarantee [113]. When observing a packet loss at the receiver, TCP assumes that the underlying network is congested and intermediate queues are dropping packets. Therefore, it reduces the transmission rate at the transmitter to a small baseline rate to alleviate the congestion in a distributed fashion, of course at the expense of a higher end-to-end delay. This congestion control approach is problematic in the presence of some faulty physical layers (e.g.,

[2]Note that packets may traverse several networks and therefore different Physical Layer technologies. For example, to watch a YouTube video on our mobile phone, the packets will be transmitted over some fibers (YouTube server to our serving BS/AP) and also wireless links (BS/AP to our phone).

wireless links) in the end-to-end connection, because by every packet loss, which may happen frequently, TCP drops the transmission rate dramatically. Moreover, the measure of TCP to detect congestion cannot help early congestion avoidance and leads to the well-known "full buffer problem" [114]. In the last decade, there have been tremendous efforts to address these problems with active queue management, and many of them are surveyed in [115].

To reduce the end-to-end delay, the existing solutions essentially change the congestion indicators, feedback types, and control functions at every intermediate node. For instance, Controlled Delay (CoDel) [116] changes the congestion indicator to a target queue latency experienced by packets in an interval, which is on the order of a worst-case round trip time. CoDel starts dropping packets when the expected queue latency exceeds the threshold. The authors showed that CoDel absorbs packet bursts in a short term manner, while keeping buffers far from fully occupied in the long term, to guarantee low-latency performance.

To reduce the latency in a mesh network, Alizadeh *et al.* [117] propose the High-bandwidth Ultra-low Latency (HULL) architecture. The main idea of HULL is similar to that of CoDel, i.e., keeping the buffers of intermediate nodes largely unoccupied. HULL trades bandwidth for buffer space, as low-latency packets require essentially no buffering in the network. HULL uses a counter, called a phantom queue in [117], to simulate queue buildup for a virtual egress link with a slower rate than the actual physical link (e.g., 95% of the link rate). The counter is put in series with the switch egress port, and is incremented by every new received packet, and decremented according to the virtual drain rate (e.g., 95% of the link rate). When the counter exceeds the threshold, it will send explicit congestion notifications to adjust the contention window size adaptively. Essentially, HULL sends early congestion signals before the saturation of the queues and reserves some portion of the link capacity for latency-sensitive traffics to avoid buffering, and the associated large delays.

E. Transmission Capacity Boosting and Sharing Techniques

In the following, we review some recent techniques that were originally proposed for boosting the transmission capacity, but have the potential to reduce the end-to-end latency.

1) MU-MIMO: MU-MIMO and beamforming are essential elements of almost all modern wireless systems including LTE-A and IEEE 802.11ac. Notably, these techniques can help steer the radiated/received energy beams toward the intended locations while minimizing interference, thereby improving the capacity region of the system [118]–[120]. We can now serve multiple users in the same time-frequency channel, which can substantially reduce the delay-to-access component of the MAC layer latency. However, MU-MIMO requires the knowledge of CSI at the transmitter and at the receiver.

As investigated in [121], [122], increasing the number of antennas simplifies the design of beamforming. In the asymptotic regime where the number of transmit antennas goes to infinity, very simple beamforming schemes, like matched filters [123], become optimal, as the wireless channels among the transmitter (e.g., AP or BS) and different devices become quasi-orthogonal [122]. However, the price of increasing the number of antennas is a higher CSI acquisition delay, which may limit the applicability of this technique for low-latency services. This problem is exacerbated in mmWave communications, as we discuss in the next subsection. Beamforming design for Multiple Input Multiple Output networks has a very rich literature with focus on spectral efficiency [124]–[126], energy efficiency [125], [127], [128], and interference cancellation [129]–[131], among others. However, surprisingly, designing beamforming for low-latency MIMO networks is a largely open problem.

2) Millimeter-wave: mmWave systems operate on a large bandwidth and employ large antenna arrays to support extremely high-data rate services, including the 8 Gbps peak data rate of IEEE 802.11ad and 100 Gbps of IEEE 802.11ay[3] for a single link [132]. In a multiuser mmWave network, the use of large antenna arrays drastically reduces the interference footprint and boosts the throughput, as shown in [64], [67], [133]–[138]. As a result, mmWave networks experience almost negligible transmission latency, but they may suffer from delay-to-access which includes pilot transmission and beamforming design. [139] overviews existing approaches for beamforming design in mmWave networks. Most of the existing approaches are based on some iterations (or equivalently a huge number of pilot signals) among transmitters and receivers, which could be very time-consuming given the low transmission rate of the control signals [63]. To reduce the beamforming setup delay, [140] and [141] augmented the beamforming part by a tracking algorithm based on extended Kalman filters, which track the second-order statistics of the channel. However, those approaches need a mobility model. Olfat *et al.* [70] alleviates that assumption by proposing a model-free (data-driven) approach based on reinforcement learning to drastically reduce the number of pilots. Still, the area of low-overhead beamforming design for large antenna arrays is in its infancy and needs further research.

Ford *et al.* [142] studied the feasibility of supporting low-latency services in mmWave cellular networks, particularly 20 Gbps data rate and 1 ms delay as specified by IMT 2020 [143]. The authors focused on beamforming aspects, MAC layer, congestion control at the transport layer, and core network architecture[4], and proposed a set of solution approaches to meet the delay and throughput requirements. In particular, the authors concluded that digital beamforming using low resolution A/D converters is a good choice for reducing beamforming delay and control channel overhead. Moreover, a flexible transmission time interval exhibits a better latency performance at the MAC layer than the conventional fixed transmission time interval.

3) Full-duplex: Full-duplex techniques enable transmission and reception at the same time-frequency resources, leading

[3]Detailed information about this project can be found at http://www.ieee802.org/11/Reports/ng60_update.htm.

[4]Due to the strong connection to the inter-network latency components, we discuss low-latency core network architectures and edge computing in the next section.

to a substantial improvement in the transmission capacity and therefore a reduction in the end-to-end latency [144]. The challenge is the strong self-interference, which can be alleviated by passive suppression (mainly related to antenna design) and active suppression (signal processing and beamforming). The latter increases the processing delay.

The low-latency merits of full duplex are in three domains. First, it decreases the transmission time (T_t) by increasing the spectral efficiency. Second, it decreases the MAC layer latency, as it reduces the potential contention, especially for the star topology where the central coordinator can send a message to one node while receiving an uplink message from another one at the same time. Last, it can decrease the routing latency (T_r). With full duplex, the route selection algorithms may activate adjacent hops simultaneously where an intermediate (i.e., relay) node operates in both downlink and uplink directions, which can substantially reduce the routing delay.

To harvest the gains of full duplex, good cancellation techniques as well as a good MAC layer design are necessary and other research problems need to be addressed. The interested readers can refer to [145] and the references therein.

4) D2D: Device-to-Device (D2D) communication enables direct communication between devices without going through the core of a cellular network [146], [147]. D2D communication is a promising solution for the increasing number of connected devices and data rate. It helps to reduce the latency in two perspectives. First, D2D enables the devices to communicate with each other in single hop or fewer hops than communication via a BS, thus reducing the routing delay T_r [148]. Secondly, the local traffic is separated from the global network (local traffic offloading). Thus, the D2D mode reduces medium access delay T_a, as the devices in D2D mode retrieving from the local source devices are fewer than those communicating with BS, and will access the channel faster. Non-D2D mode devices retrieving contents from the core network also benefits with lower T_a, because D2D mode offloads part of the contending devices.

If the D2D mode shares the same resources with the core network communication, resource allocation should be used to mitigate interference and guarantee low latency to both D2D and non-D2D users. The integration of mmWave and D2D can substantially alleviate the resource allocation problem [149], thanks to the small interference footprint of mmWave networks [64], [67], [133]–[138]. Niu *et al.* [150] showed that we can add many D2D links to a mmWave cellular network so as to substantially boost the network capacity. The authors have also developed a simple scheduling scheme to activate concurrent transmissions to support low-latency content downloading. The main challenges are the device discovery and the exchange of control signals, which are usually much harder in mmWave networks [80].

5) Cloud RAN and Mobile Fronthaul: A radio BS consists of a Baseband Unit (BBU) and a radio frequency unit. The concept of cloud Radio Access Network consists in breaking the fixed topology between BBUs and Remote Radio Headss (RRHs), and to form a virtual BBU pool for centralized control and processing [151]. Cloud Radio Access Network supports inter-cell communication and joint processing, which can reduce the routing latency (T_r) as well as the processing latency (T_{pr}). Mobile fronthaul is a novel optical access method that connects a centralized BBU to a number of RRHs with fiber links in mobile networks [152]. A SDN-controlled optical topology-reconfigurable fronthaul architecture is proposed for 5G mobile networks [153]. The SDN-controlled fronthaul architecture is responsible for the dynamic configuration of the BBUs and RRHs connections to support coordinated multipoint and low-latency inter-cell D2D connectivity. The experimental results show that with 10 Gbps peak data rate, sub-millisecond end-to-end D2D connectivity is achievable [153].

In [154], a combination of fiber and mmWave is proposed for efficient fronthauling to lower the cost and support mobility in small cells and moving cells. mmWave is used for transmission between the large number of RRHs or moving RRHs and a remote antenna unit, and the fiber optic is used for the connection between the remote antenna unit and BBU pool. To reduce the latency caused by the conversion between optic signal and mmWave signal, analog waveform transmission is used for eliminating the digital processing, and a uni-travelling carrier photodiode optical-to-electrical converter is used to provide fast conversion.

F. Cross-layer Techniques

1) Forward Error Correction and Hybrid Automatic Repeat Request Techniques: Automatic Repeat Request is a simple error-control method for data transmission that uses acknowledgements and timeouts to achieve reliable data transmission over an unreliable physical layer [155]. Acknowledgments are short messages sent by the receiver indicating that it has correctly received a packet, and timeout is a predefined latency allowed to receive an acknowledgment. If the sender does not receive an acknowledgment before the timeout, it usually re-transmits the frame/packet until the sender receives an acknowledgment or exceeds a predefined number of re-transmissions. Block acknowledgement –initially defined in IEEE 802.11e– is a simple way to reduce the MAC layer latency, especially for high-throughput devices, by sending one acknowledge packet for multiple data packets [156].

Hybrid Automatic Repeat Request combines high-rate forward error-correcting coding and Automatic Repeat Request error-control to flexibly perform retransmission of incremental redundancy or a complete new retransmission according to different channel states. If the received packet can not be decoded correctly due to high noise, incremental redundancy can be retransmitted for joint decoding together with the previously received packets. If strong interference is the reason, a new start of transmission is needed. There is a trade-off between the transmission time T_t and medium access delay T_a. With lower coding rate, T_t gets longer, while T_a is shorter as the message can be received successfully with a higher probability and possibly fewer retransmissions.

2) MAC-aware Routing Algorithms: The delay and reliability performance interaction between the MAC and routing of the protocol IEEE 802.15.4 is analytically studied in [97].

Therein, it has been shown that the MAC parameters will influence the performance of different routing paths, and, in turn, the traffic distribution determined by the routing will also affects the MAC parameters. For a given topology, the MAC parameters (affecting T_a) can be tuned to satisfy a certain reliability and latency requirement by using the Q-metric proposed in [97]. The Q-metric measures the contention level without measuring the queues, and adapt the routing patterns (affecting T_r). While the back-pressure is proved to be throughput optimal, it is efficient only when the forwarded traffic is high, and it can not capture the contention of low traffic. So Q-metric is more efficient in latency-sensitive wireless sensor networks and other cases where the traffic is low.

A TDMA-based MAC for wireless sensor networks that have latency requirements is Delay Guaranteed Routing and MAC [157]. However, unlike traditional TDMA MACs that require a separate routing mechanism, the routes of Delay Guaranteed Routing and MAC as well as the medium access slot schedule are determined and fixed according to the position of each node. The algorithm is TDMA-based and no retransmissions are permitted. As long as the transmission interval of two successive packets of each node is larger than the TDMA superframe duration, a deterministic upper delay bound and minimal packet loss can be guaranteed.

G. Network Function Virtualization and Software-defined Networking

Network Function Virtualization (NFV) and Software-Defined Networking (SDN) cannot reduce the latency by themselves alone, but they are the premise of some algorithms that help achieve low latency, so we also study them together as an enabling technique.

NFV can virtualize the network node functions into building blocks that may connect, or chain together, to create communication services. With the necessary hardware support, the system can change flexibly by employing different NFV blocks. For instance, different asynchronous waveforms can be generated by combining different filters and modulation schemes [158], which affects the transmission time (T_t).

Though NFV can act alone, it is generally working together with SDN. SDN decouples the control plane and data plane, promoting centralized network control and the ability to program the network [159]. SDN can be used to implement and manage the NFV infrastructure by combining and tuning parameters of multiple NFV blocks, such as to enable flexible configuration of the virtual network slices for different latency priorities [142]. By tuning the parameters of different NFV blocks, which further virtualize different techniques and affects different latency components, SDN and NFV may have an ample effect on various latency components. On the other hand, SDN will introduce overheads from the control plane and flow setup latency as it is flow based. Software-based NFV also tends to increase the processing latency (T_{pr}) compared to the pure hardware operation. The effect of these on the delay performance should be carefully controlled and minimized.

The ETSI NFV architecture [160] acts as a reference standard, whose performance and security have been well studied [161], [162]. [163] proposes a 5G architecture for low-latency and secure applications based on ETSI NFV architecture. To achieve low latency with shared resources, the architecture employs the scalability enabled by SDN and NFV to perform on-demand caching and switching, which can guarantee low medium access latency (T_a), routing latency (T_r), and queuing latency (T_q) by flexibly allocating different resources based on different traffic load and requirements including latency. Moreover, it also proposes to use a smart network interface card with NFV acceleration capability to mitigate the processing latency (T_{pr}) caused by NFV. A backbone network that provides high-performance connectivity in Japan is equipped with the capabilities of NFV and SDN [164]. The NFV orchestrator creates virtual network appliances, such as virtual routers, virtual firewalls, and virtual load balancers. SDN enables users to establish connections with specified bandwidth and demand, which flexibly scale up or down the virtual network appliances. In this way, SDN affects all the latency components whose parameters has been dynamically tuned, such as transmission time (T_t), medium access latency (T_a) and routing latency (T_r). The latency is reduced by 20% compared to the previous version network, and the auto-healing time of a virtual network functions also drops to 30 s from 6 min from the previous version.

V. Inter-network Techniques and Technologies

Many modern end-to-end services require packets to traverse multiple networks, usually handled by different operators. Unfortunately, in most practical scenarios, the exact latencies of the intermediate networks are not known, and the service provider may have access only to the delay probability distributions. As a result, controlling the latency in a multi-domain scenario is significantly more complicated than managing latency in single domain scenarios.

In this section, we address techniques that can be used to control inter-network (also called multi-domain) latency. We first review some multi-domain routing algorithms that guarantee end-to-end latency, though they may not meet the tight requirements of low-latency services. We then highlight the importance of content placement and edge caching to facilitate multi-domain routing and substantially reduce latency.

A. Inter-domain Routing

In Section IV-C, we have discussed various routing techniques that can reduce the latency within a domain. Border Gateway Protocol (BGP) is a celebrated inter-domain routing protocol used throughout the Internet to exchange routing information among different domains [165]. In BGP, the edge routers frequently send path vector messages to advertise the reachability of networks. Each edge router that receives a path vector update message should verify the advertised path according to the policy of its domain. If it complies with its policy, the router modifies its routing table, adds itself to the path vector message, and sends the new message to the neighbor domain. As a result, edge routers of every domain maintain only one path per destination.

Although BGP has evolved for many years, its current implementation is based on the number of hops. Reference [166] proposes a method to measure the latency, share it with the edge routers of the neighboring domains, and modify the BGP routing decisions throughout the entire network. [167] and [168] generalize this idea to incorporate a logical SDN in a multi-domain network. [167] formulated a constrained shortest path problem to minimize a convex combination of packet loss and jitter subject to an end-to-end latency constraint. [169] demonstrated a multi-domain SDN orchestrator to select the shortest routes based on end-to-end latency, where the delay statistics are captured by the proposed segment routing monitoring system.

Reference [170] extends BGP to allow for path-diversity in the sense that multiple paths are advertised by any edge router, multiple QoS metrics including latency are propagated in the BGP update message, and multiple routes for any source-destination pair are selected. The authors showed that path-diversity substantially improves load balancing throughout the multi-domain network and can reduce the end-to-end latency. [171] developed a dynamic routing policy for delay-sensitive traffics in an overlay network. In particular, the authors replaced the average end-to-end latency requirement by an upper bound on the average queue length of every flow on every link. Then, considering an underlying legacy network whose latency characteristics are unknown to the overlay network, they formulated a constrained Markov decision process that keeps the average queue lengths bounded, and proposed a distributed algorithm to solve that problem. Although this paper targets a single domain, the main idea can be extended to the case of multi-domain networks. Classical stochastic shortest path [172] and its online variation [173] are highly relevant to the problem of routing design in multi-domain networks when only imperfect knowledge of latency within each domain is available. Multi-domain routing design with limited domain-specific information is a very interesting and wide open area for future research.

B. Edge Caching and Content Placement

Define the feasible latency region as the set of all possible latency values that can be achieved by some routing policy. In the previous subsection, we observed that optimizing inter-domain routing reduces the end-to-end latency. However, when the content is located far from the terminal, the feasible latency region may not include the desired value of the low-latency service; namely, there is no routing algorithm that can lead to the target latency value. As a simple example, consider a line network of 10 domains (including source and destination ones), and assume that each adds at least 10 ms latency. Therefore, the round-trip latency is lower-bounded by 200 ms, which may be way beyond the tolerable latency. Caching popular contents at the edge routers of local domains is a promising approach to bring the contents closer to the devices so as to improve the feasible latency region.

Edge caching brings the content closer to the end terminals, which substantially reduces the inter-domain routing delay. From the local server perspective, fewer terminals (only local ones) contend to access that content, which improves the service rate to those terminals. Moreover, this technique offloads parts of the traffics of the main server, freeing the capacity for other terminals and further reducing the latency. Altogether, edge caching is especially beneficial for services with stringent latency requirements.

The design of efficient caching strategies involves a broad range of problems, such as accurate prediction of demands, intelligent content placement, and optimal dimensioning of caches. Moreover, as the caches are physically scattered across the network and the user requests are generated almost everywhere in the network, caching policy favors low-complexity distributed algorithms. Borst *et al.* [174] propose a light-weight cooperative content placement algorithm that maximizes the traffic volume served from caches and thus minimizes the bandwidth cost. Maddah-Ali and Niesen [175], [176] focus on the problem of caching within a network where popular content are pre-fetched into the end-user memories to bypass a shared link to the server and showed that there is a trade-off among the local cache size (i.e., the memory available at each individual user), aggregated global cache size (i.e., the cumulative memory available at all users), and the achievable rate (and latency) of individual terminals. The authors develop a simple coded caching scheme in [175] and its decentralized variant in [176] that substantially improves the memory-rate trade-off. The analysis and proposed algorithms, however, are limited to the single-domain single-server case. Doan *et al.* propose a novel popularity predictingcaching procedure for backhaul offloading in cellular network, and an optimal cache placement policy to minimize the backhaul load is computed by taking both published and unpublished videos as input [177].

Reference [178] focuses on a video-on-demand application over a network with two modes of operations: peer-to-peer and data center. In particular, video requests are first submitted to the peer-to-peer system; if they are accepted, uplink bandwidth is used to serve them at the video streaming rate (potentially via parallel substreams from different peers). They are rejected if their acceptance would require the disruption of an on-going request service. Rejected requests are then handled by the data center. The authors developed a probabilistic content caching strategy that enables downloaders to maximally use the peers uplink bandwidth, and hence maximally offload the servers in the data centers. Golrezaei *et al.* [179] extended the model of [178] to the scenario of video-on-demand streaming to mobile terminals from Internet-based servers and proposed a distributed caching network to reduce the download latency. In this setting, local caches with low-rate backhaul but high storage capacity store popular video files. If the file is not available in the local cache, it will be transmitted by the cellular network. The authors analyzed the optimal assignment of files to the caches in order to minimize the expected downloading time for files. They showed that caching the coded data can substantially improve the performance in terms of both computational complexity and aggregated storage requirement. Bastug *et al.* [180] proposed a caching strategy that predicts the demand pattern by the users and caches them, in a proactive manner, in local BS during the off-peak hours.

When the users actually make the demands, the contents can be retrieved with high probability directly from the cache instead of waiting for the backhaul network and inter-domain routing latency.

Bhattacharjee *et al.* [181] proposed a self-organizing cache scheme in which every router of the network maintains a small cache and applies an active caching management strategy to organize the cache contents. [182] focused on the optimal placement of M web proxies in N potential sites with the objective of minimizing the overall latency of searching a target web server for a given network topology. The authors formulated this problem as a dynamic program, and obtained the optimal solution in polynomial time.

C. Fog Computing

With IoT, billions of previously unconnected devices are generating more than two exabytes of data each day, and 50 billion devices are estimated to connected to the Internet [183]. Such large amount of data cannot be processed fast enough by the cloud. Fog computing is proposed to extend cloud RAN further to the edge, such that any device with computing, storage, and network connectivity can be a fog node [184], [185]. The cloud and fog nodes merge into a new entity, referred to as cloud+fog [186]. In a cloud+fog architecture, critical IoT data with stringent latency requirement can be processed at the closest fog node to minimize latency, while less delay-sensitive data can be passed to the aggregation node or the cloud. Fog computing also offloads gigabytes of network traffic from the core network. Similar to edge caching, fog computing reduces the routing latency and channel establishment latency to both data processed by the fog nodes and by the core network.

Two main challenges in realizing fog computing's full potential are to balance load distribution between fog and cloud, and to integrate heterogeneous devices into a common computing platform [187]. To evaluate resource-management and scheduling policies across fog and cloud resources, an open source simulator called iFogSim is developed, which can model and simulate the performance in terms of latency, energy consumption, network congestion and operational costs [188]. An architecture of Smart Gateway with Fog Computing is presented in [189], where the Smart Gateway can collect, preprocess, filter, reconstruct, and only upload necessary data to the cloud. To handle heterogeneous data collected from heterogeneous devices, transcoding and interoperability are either achieved by equipping the Smart Gateway with more intelligence or through the fog computing resources.

VI. Standards for Low-latency and Ultra-reliable Communications

In this section, we analyze how intra networks and inter network can be supported by current and emerging communication standards. We will review the low-latency characteristics of standards in cellular networks, industrial communication, and WLAN group. One or a combination of these standards are used for the uses cases discussed in Section II.

A. The 3GPP New Radio and 5G Initiative

5G of mobile communication aims to support 1 ms latency, 10 Gbps peak speed, and, at a global level, 100 billion connections. 5G has three main classes of use cases: URLLC, enhanced mobile broadband, and massive machine type communication. URLLC has the most stringent requirement for very low latency and high reliability, which suits applications such as factory automation, smart grid, and intelligent transportation. With a vision to provide a unified infrastructure for different use cases, 5G will include both the evolution of 4G and a new radio access technology.

5G is designed to operate in a wide range of spectrum including frequency bands below 1 GHz up to 100 GHz. The wide bandwidth at higher frequencies including mmWave band can effectively boost the transmission rate and reduce the transmission time (T_t). Moreover, flexible deployment of more micro/pico sites at traffic dense spots eases the medium access pressure, and reduces medium access time (T_a). At the physical layer, 5G supports various modulations from QPSK, 16 QAM to 1024 QAM to support different transmission rates and transmission time (T_t) for different use cases. The current candidate waveform is OFDM with scalable numerology and adjustable sub-carrier spacing, CP duration and OFDM symbol duration, which supports adjustable transmission time (T_t). A detailed assessment of OFDM in 5G can be found in [190]. MAC scheduling in 5G follows the time-slotted framework of 4G, and the transmission can only start at the beginning of the scheduled slot. To improve the access efficiency of short packets for URLLC, the time slot in 4G can be divided into multiple mini-slots (the length can be as short as one OFDM symbol) to enable lower delay to access, which can greatly reduce the medium access time (T_a). Multiple antennas will also be supported in 5G for MU-MIMO, which also helps reduce the medium access time (T_a). With the increase of the carrier frequency, the number of antennas and the multiplexing order will increase. Meanwhile, the complexity to obtain CSI for beamforming also increases. A highly flexible but unified CSI framework is supported by 5G, which enables different antenna deployments corresponding to different CSI settings.

B. Industrial Communication Networks

Industry 4.0 is currently seen as the most advanced industrial automation trend, where one of the important aspects is real-time communication between industrial modules. Wireless networks offer simple deployment, mobility and low cost, and are gaining popularity in the industrial sites [191]. The requirements in industrial applications to support high reliability, high data rates and low latency pose difficulties to wireless networks deployment, where the bottleneck mostly lies in the latency. To address these challenges, there have been some proposals, such as WirelessHP and IEEE 802.15.4e.

The recently proposed Wireless High Performance (Wireless HP) aims to provide a physical layer solution to support multi-Gbps aggregate data rate, very high reliability level ranging from 10^{-6} to 10^{-9}, and packet transmission time (T_t) lower than 1 μs [48]. Taking the advantage of deterministic and periodic traffic in the latency-sensitive industrial applications,

WirelessHP reduces the PHY layer preamble length (part of T_h) (while still ensuring reliable packet decoding) and optimizes OFDM parameters to reduce the inefficiencies that affect short packet transmission.

IEEE 802.15.4 is a successful protocol that also forms the basis for the first open standard WirelessHart for process automation applications in the industrial field. However, the drawbacks of low reliability and unbounded packet latency limit its deployment in the industrial applications that have stringent requirements for latency and reliability [192]. To overcome these limitations, the recently released IEEE 802.15.4e amendment introduces MAC layer enhancements in three different MAC modes. Despite the individual features among the different modes, here we focus on the modifications in terms of latency performance improvement. To guarantee bounded medium access time (T_a), channel access is time slotted and included both contention-free and contention-based modes for periodic and aperiodic traffic respectively. The number of time slots can be flexibly tuned according to the traffic load, which avoids the inefficiency of idle slots incurred in fixed framing structures. Channel hopping is applied to combat fading and improve reliability, which in turn will lower the retransmission latency, thus reducing medium access time (T_a).

C. IEEE WLAN Group Standardization

IEEE 802.11 is a set of MAC and PHY specifications to implement Wireless Local Area Network in the 0.9, 2.4, 3.6, 5, and 60 GHz frequency bands. The key parameter index of the IEEE 802.11 family focus on data rate, coverage range, connectivity. Though latency performance is not specified in the protocols, the protocol IEEE 802.11ak is designed to support industrial control equipment, and latency should be bounded in this scenario. Moreover, the ability and infrastructure of IEEE 802.11 to support high data rate make it an indispensable part in the ecosystem to achieve low latency. IEEE 802.11ax, which is due to be publicly released in 2019, is designed to improve spectral efficiency at 2.4 GHz and 5 GHz. The technical highlights are the modulation support of up to 1024 QAM and multiuser support in both frequency and spatial domains by the combination of OFDMA and MU-MIMO, which are effective to decrease the transmission time T_t and medium access latency T_a respectively. IEEE 802.11ay is the follow-up of 802.11ad working at 60 GHz. Compared to 802.11ad with 2.16 GHz bandwidth, 802.11ay has four times the bandwidth by channel bonding. Moreover, MIMO is added with a maximum of 4 streams with a per-stream link rate of 44 Gbps, which can substantially decrease the transmission time T_t for heavy traffic by using the large bandwidth at higher frequency band.

D. Other Standardization Activities

Among the whole wide spectrum, only a small portion is regulated as licensed spectrum. While the licensed spectrum has better performance due to less interference, the increasing number of connected devices drives the necessity to use unlicensed spectrum. License Assisted Access, LTE in Unlicensed Spectrum and Multefire [193] are three representatives to explore LTE services in the unlicensed 5 GHz band [194]. License Assisted Access and LTE in Unlicensed Spectrum use the unlicensed band by offloading traffic to boost data rate and reduce transmission time (T_t), while the control signals stay in the licensed band. Multefire takes a step even further, with no anchor at the licensed band, and both the control signal and the data traffic are transmitted in the unlicensed band. Thus Multefire not only helps to reduce the transmission time (T_t) thanks to the boosted data rate, but it also reduces the access latency (T_a) with more access resources.

IEEE 802 Time-Sensitive Networking aims to deliver deterministic latency over Ethernet networks [195]. Possible applications include converged networks with real-time Audio/Video Streaming and real-time control streams which are used in automotive or industrial control facilities. The objective latency for short messages per hop is set to be 4 μs or less with 1 Gbps transmission rate. To guarantee deterministic delay for the time-sensitive data, the switches are used to schedule the data transmission. The switches should be aware of the cycle time of these latency-sensitive data, and during the window expected for the arrival of time-sensitive data, the switch will block non-time-sensitive interfering traffic to eliminate queueing latency T_q. To enhance the reliability, more than one path is used simultaneously.

VII. Further Discussions

A. Short Packets

In Section IV-A3, we described research topic about sending short packets when the sizes of payload and header are comparable, how to jointly code the payload and meta data in an efficient way is of great importance to improve the spectral efficiency and to reduce the transmission time T_t. Moreover, for short packets, channel establishment delay T_a can be much longer than the transmission time, which is very inefficient, so the MAC control overhead should be modified. Another open issue is to get the optimal value of the maximum coding rate for finite packet length and finite packet error probability, which is an NP-hard problem with exponential complexity [52].

B. The Trade-off for Low latency

In Subsection IV-B and IV-C, we described the research activities around the combinations of different techniques and different parameter settings have different latency performance and other performance indicators. Expectedly, there are trade-offs between the latency performance and other performance metric such as throughput, reliability and energy consumption. There is a trade-off between latency and throughput for different MAC scheduling schemes [196], the trade-off between latency and reliability of slotted ALOHA and CSMA is shown in [60], and the trade-off between delay and energy consumption in [83]. The trade-off between latency and throughput in ad hoc network routing is given in [99], [101]. The trade-offs may be different for different regions (i.e., different throughput or reliability). When the performance other than latency does not exceed a boundary, the latency grows mildly, while the

latency increases sharply when exceeding that boundary. For different techniques, such trade-offs are expected to exist, and the related boundaries should be well determined.

C. mmWave

In Section IV-E2, we discussed about mmWave as an enabling technology for low-latency communication, and it is also a research direction that attracts much attention with many open problems to be investigated.

Firstly, beamforming is the premise to support the directional transmission and the very high data rate to help reduce latency, however, the delay caused by beamforming may also prohibit mmWave techniques to work in low-latency communications. Moreover, as the coherence time in mmWave band is shorter than that in microwave band, efficient beamforming algorithms in mmWave prove to be very prominent research direction in mmWave and are essential for mmWave techniques to be used in low-latency scenarios. Giordani *et al.* surveyed some existing beamforming and beam-tracking techniques for 3GPP New Radio at the mmWave frequencies [197].

Secondly, although mmWave communication is mostly noise limited, the SINR at the receiver side might probably degrade greatly when the beam at the receiver side is also aligned to undesired transmitting beams besides the desired one. Thus, multilevel HARQ is needed to indicate retransmission for certain frames corrupted by random strong noise, or a new transmission when experiencing high interference [198]. When the high interference lasts for a long time, retransmission may not be useful, so effective and efficient solutions are needed.

Another problem occurs in the case of blockage in the main beam aligned link. As the PHY layer frame duration is much smaller compared to the time for the vehicle or people to move away, under this case the device should fall back to search other mmWave BSs in the local area instead of waiting. And if no suitable mmWave BSs are available, UE may also fall back to use microwave. Then whether to reassociate in the mmWave band or to change to microwave band during blockage is an open problem. An efficient way to determine the reason for performance degradation is of practical importance to reduce the delay in mmWave.

D. Combination with PHY Parameters

In Subsection IV-B and IV-C, we discussed the effect of PHY parameters on the corresponding latency components. For contention-based MACs, the probability of a successful transmission is often calculated by the probability when there are no more than one node transmitting at the same time. This practice holds for the omnidirectional transmission. When we use directional transmission with a large number of antenna elements, e.g., in mmWave systems, some simultaneous transmissions may not cause interference to each other. Due to this reason, the contention-based MACs need to be revised for certain scenarios. Moreover, given a reliability requirement, retransmissions due to poor SINR may be needed besides collision during the access phase, so the PHY parameters should be combined to determine the MAC parameters, and more effective and efficient cross layer design to achieve low-latency need to be investigated.

At the routing layer, as the topologies are often unknown and may change in time, it is difficult to determine the efficient interference model. However, in directional communications, multiple links can be activated without causing interference, and this suggests new research topics at the routing level. The trade-off between the time to establish directional communication and the improvement of the routing delay should be investigated.

E. Unified Communication Network for Low-Latency Applications

In Sections VI and IV-G, we described the standardization activities, NFV and SDN separately, here we will discuss the open research that arise under the vision of 5G. 5G aims at providing a unified infrastructure for a wide variety of use cases from media sharing (requiring high transmission date), massive machine type communication (requiring access availability), to applications that requires low latency. Moreover, for the latency-sensitive applications, different requirements in terms of latency, reliability, transmission rate and energy consumption are all different. When considering different applications as verticals, they operate on top of horizontal communication infrastructures, resource, and techniques [29]. A unified communication ecosystem should figure out how to share horizontal infrastructures and resource tailored to different verticals. NFV and SDN are necessary to enable flexible configuration of the communication services. When designing protocols, interfaces between different communities should also be taken into consideration to ensure easy and efficient combination of the horizontal techniques.

Traffic offloading is another key technique to promote a unified communication network. By 2020, 5G is expected to increase the area capacity 1000-fold, and to be able to connect 100 billion devices. Despite the remarkable growth in capacity, it is still difficult to support such a large number of connections. Even with different priorities, latency-sensitive transmission may not be guaranteed. Traffic offloading may greatly relieve the burden of the mobile network by using other communication forms to share the traffic. For instance, local traffic can be directly exchanged by D2D communication mode without going through the infrastructure. An architecture unifying the coverage-centric 4G mobile networks and data-centric fiber-wireless net broadband networks was proposed in [154]. This architecture uses fiber as backhaul sharing and WiFi to offload the mobile user traffic from the 4G mobile networks, and as a result the end-to-end delay decreases dramatically on the order of 1 ms.

F. Age-of-Information

Age of Information (AoI) is a new performance metric that measures the amount of time that elapsed since the most recently received packet was generated at its source. As such, AoI measures the ”freshness of information” from the perspective of the destination. AoI has been receiving increasing attention in the literature, particularly for applications that

generate time-sensitive data such as position, command and control, or sensor data. AoI is a function of packet delay and packet inter-delivery time. Thus, low delay alone many not be sufficient to achieve good AoI performance. For example, an M/M/1 queue with a low arrival rate and a high service rate may have low queueing delay but high AoI because the packet inter-arrival times are large. Thus, achieving good AoI performance involves a balance between maintaining low delays and small packet inter-arrival times. To better understand these phenomena, reference [199] modeled the network between the source and destination as a single first-in-first-out (FIFO) queue, and proved that there is indeed an optimal rate at which AoI is minimized. Since then, most of the work on AoI has focused on single queue models. Age for FIFO M/M/1, M/D/1, and D/M/1 queues was analyzed in [199], multiclass FIFO M/G/1 and G/G/1 queues were studied in [200]. However, the problem of minimizing AoI in a wireless network with interference constraints has received limited attention. In [201], the authors develop scheduling algorithms for minimizing AoI over a wireless base-station, where only one node can transmit at a time, and in [202] the authors study the AoI minimization problem in a wireless network subject to general interference constraints. The approach is generalized to multi-hop wireless networks in [203].

G. Hardware and Smart Devices

Due to the limitation of our expertise, we mainly discuss from technique respective, there are also many open research problems about hardware and smart devices which enable the functioning of different techniques. When we seek to exploit the wide bandwidth at higher frequencies such as mmWave, the according hardware that operates in the high bands are needed, and technologies in the higher frequencies can be widely used to achieve good performance when the price of the hardware becomes lower.

Nowadays the devices are becoming more and more smart in terms of better processing capability and larger caching size. This smartness not only enables installing and running applications smoothly, but also help to achieve low-latency. With larger caching capability, the devices can proactively cache popular contents by an analysis of the user preference from the previous contents the user has browsed [180]. Then when the user actually makes the request to these cached contents, the device can display them with local cache instead of accessing through the network.

Another direction is to enable the devices to support D2D communication from both the hardware and protocol perspective [204]. Then the devices can communicate with others in proximity using lower power and probably in a single hop, instead of communicating through the BS going through multiple hops or using strong transmitting power. However, the trade-off between the extra overhead for control and channel estimation and the delay saved by communicating in the D2D mode still needs further research.

H. Security

Currently, the packet overheads due to security or privacy methods substantially contribute to the delay both for the communication of information itself and for the decoding/processing. The classic approach to privacy and security is cryptography [205]. However, cryptography introduces formidable overheads and heavy coordination over the transmitters/receivers that are involved, which substantially makes it impossible to achieve very low latencies. Existing alternative privacy methods for low latency networking demand substantial investigation. Alternative methods to cryptography namely, information-theoretic secrecy [206], differential privacy [207], k-anonymity [208], and signal processing security methods [209] present shortcomings when it comes to their use for low latency networking. Like cryptography, information-theoretic secrecy prevents an eavesdropper acquiring information from two communicating nodes. However, the channel acquisition phase that is needed for these methods includes substantial delays that are in contrast with low latencies. Differential privacy, k-anonymity and signal processing methods perturb the original data to make data analysis, and as such it is not clear how they can be used to ensure private or secure low latency communications.

VIII. Conclusion

Low-latency communications are arguably the most important direction in the next generation of communication networks. In this paper, we showed that the most prominent use cases demanding low latency are supported by complex network interactions, including inter-network and intra-network interactions. To realize low-latency networks, it is important to determine where and how latency occurs and what methods can help to reduce it. We investigated how the delay accumulates from physical layer to transport layer, and we showed how to characterize the end-to-end delay into several components. Then we discussed how different techniques may influence one or multiple delay components. We argued that these techniques should be optimized together to reduce the delay while satisfying other requirements such as reliability and throughput. MU-MIMO, mmWave, and full-duplex, which can greatly improve the data rate, were considered as three enabling technologies to support low-latency communication, but each of the three also poses challenges, e.g., MU-MIMO and mmWave will introduce beamforming delay before the start of the transmission. Finally, sending short packets, the trade-off between latency and other network performance indexes, using mmWave bands, design with physical parameters, hardware and smart devices design, and traffic offloading are some of the most promising research areas that will need substantial future research developments.

References

[1] F. P. Kelly, "Network routing," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 337, no. 1647, pp. 343–367, 1991.

[2] M. Weiser, "The computer for the 21st century," *Scientific american*, vol. 265, no. 3, pp. 94–104, 1991.

[3] M. S. Corson, R. Laroia, J. Li, V. Park, T. Richardson, and G. Tsirtsis, "Toward proximity-aware internetworking," *IEEE Wireless Communications*, vol. 17, no. 6, pp. 26–33, 2010.

[4] N. Katugampala, K. T. Al-Naimi, S. Villette, and A. M. Kondoz, "Real-time end-to-end secure voice communications over GSM voice channel," in *Signal Processing Conference, 2005 13th European*. IEEE, 2005, pp. 1–4.

[5] T. Chen, G. Charbit, K. Ranta-aho, O. Fresan, and T. Ristaniemi, "VoIP end-to-end performance in HSPA with packet age aided HSDPA scheduling," in *Personal, Indoor and Mobile Radio Communications, 2008. IEEE 19th International Symposium on*. IEEE, 2008, pp. 1–5.

[6] M. S. Corson, R. Laroia, J. Li, V. D. Park, T. Richardson, G. Tsirtsis, and S. Uppala, "Flashlinq: Enabling a mobile proximal internet," *IEEE Wireless Communications*, vol. 20, no. 5, pp. 110–117, 2013.

[7] N. Brahmi, O. N. Yilmaz, K. W. Helmersson, S. A. Ashraf, and J. Torsner, "Deployment strategies for ultra-reliable and low-latency communication in factory automation," in *Globecom Workshops*. IEEE, 2015, pp. 1–6.

[8] M. A. Lema, A. Laya, T. Mahmoodi, M. Cuevas, J. Sachs, J. Markendahl, and M. Dohler, "Business Case and Technology Analysis for 5G Low Latency Applications," *IEEE Access*, vol. 5, pp. 5917–5935, 2017.

[9] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-enabled tactile internet," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 460–473, 2016.

[10] A. A. Zaidi, "Designing for the Future: The 5G NR Physical Layer," Ericsson Review, 2017.

[11] J. Torsner, "Industrial Remote Operation: 5G Rises to the Challenge," Ericsson Review, 2015.

[12] I. Redpath, "Monetizing High-Performance, Low-Latency Networks," Ovum, 2017.

[13] M. Weiner, M. Jorgovanovic, A. Sahai, and B. Nikolić, "Design of a low-latency, high-reliability wireless communication system for control applications," in *Communications IEEE International Conference on*. IEEE, 2014, pp. 3829–3835.

[14] K. Yamamoto, F. Ichihara, K. Hasegawa, M. Tukuda, and I. Omura, "60 GHz wireless signal transmitting gate driver for IGBT," in *Power Semiconductor Devices & IC's, IEEE 27th International Symposium on*. IEEE, 2015, pp. 133–136.

[15] M. Maier, M. Chowdhury, B. P. Rimal, and D. P. Van, "The tactile internet: vision, recent progress, and open challenges," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 138–145, 2016.

[16] G. P. Fettweis, "The tactile internet: applications and challenges," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 64–70, 2014.

[17] A. Osseiran, J. Sachs, and M. Puleri, "Manufacturing reengineered: robots, 5G and the Industrial IoT," *Ericsson Business Review*, 2015.

[18] J. Hasbrouck and G. Saar, "Low-latency trading," *Journal of Financial Markets*, vol. 16, no. 4, pp. 646–679, 2013.

[19] R. Bogue, "Exoskeletons and robotic prosthetics: a review of recent developments," *Industrial Robot: An International Journal*, vol. 36, no. 5, pp. 421–427, 2009.

[20] D. J. Tyler, "Restoring the human touch: Prosthetics imbued with haptics give their wearers fine motor control and a sense of connection," *IEEE Spectrum*, vol. 53, no. 5, pp. 28–33, 2016.

[21] M. E. Tozal, Y. Wang, E. Al-Shaer, K. Sarac, B. Thuraisingham, and B.-T. Chu, "Adaptive information coding for secure and reliable wireless telesurgery communications," *Mobile Networks and Applications*, vol. 18, no. 5, pp. 697–711, 2013.

[22] ETRI TR 102 935 V2.1.1, "Machine-to-Machine communications (M2M); Applicability of M2M architecture to Smart Grid Networks;Impact of Smart Grids on M2M platform," ETRI TR 102 935 V2.1.1, Tech. Rep., September 2012.

[23] D. Orfanus, R. Indergaard, G. Prytz, and T. Wien, "Ethercat-based platform for distributed control in high-performance industrial applications," in *Emerging Technologies & Factory Automation IEEE 18th Conference on*, 2013, pp. 1–8.

[24] Y. Liu, R. Candell, and N. Moayeri, "Effects of wireless packet loss in industrial process control systems," *ISA transactions*, vol. 68, pp. 412–424, 2017.

[25] Z. Pang, M. Luvisotto, and D. Dzung, "High performance wireless communications for critical control applications," *IEEE Industrial Electronics Magazine*, 2017, to be published.

[26] V. C. Güngör, D. Sahin, T. Kocak, S. Ergüt, C. Buccella, C. Cecati, and G. P. Hancke, "Smart grid technologies: Communication technologies and standards," *IEEE transactions on Industrial informatics*, vol. 7, no. 4, pp. 529–539, 2011.

[27] R. H. Khan and J. Y. Khan, "A comprehensive review of the application characteristics and traffic requirements of a smart grid communications network," *Computer Networks*, vol. 57, no. 3, pp. 825–845, 2013.

[28] IEC, *IEC 61850 Standard*, http://http://www.iec.ch/, Std.

[29] M. Wollschlaeger, T. Sauter, and J. Jasperneite, "The future of industrial communication: Automation networks in the era of the internet of things and industry 4.0," *IEEE Industrial Electronics Magazine*, vol. 11, no. 1, pp. 17–27, 2017.

[30] L. Turchet, A. McPherson, and C. Fischione, "Smart instruments: Towards an ecosystem of interoperable devices connecting performers and audiences," in *Proceedings of the Sound and Music Computing Conference*, 2016.

[31] L. Turchet, M. Benincaso, and C. Fischione, "Examples of use cases with smart instruments," in *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*. ACM, 2017, p. 47.

[32] L. Turchet, C. Fischione, and M. Barthet, "Towards the internet of musical things," in *Proceedings of the Sound and Music Computing Conference*, 2017, pp. 13–20.

[33] J. Santa and A. F. Gomez-Skarmeta, "Sharing context-aware road and safety information," *IEEE Pervasive Computing*, vol. 8, no. 3, pp. 58–65, 2009.

[34] M. Alsabaan, K. Naik, and T. Khalifa, "Optimization of fuel cost and emissions using V2V communications," *IEEE Transactions on intelligent transportation systems*, vol. 14, no. 3, pp. 1449–1461, 2013.

[35] R. Yu, Y. Zhang, S. Gjessing, W. Xia, and K. Yang, "Toward cloud-based vehicular networks with efficient resource management," *IEEE Network*, vol. 27, no. 5, pp. 48–55, 2013.

[36] M. Gerla, E.-K. Lee, G. Pau, and U. Lee, "Internet of vehicles: From intelligent grid to autonomous cars and vehicular clouds," in *Internet of Things, IEEE World Forum on*. IEEE, 2014, pp. 241–246.

[37] R. A. Uzcátegui, A. J. De Sucre, and G. Acosta-Marum, "Wave: A tutorial," *IEEE Communications Magazine*, vol. 47, no. 5, pp. 126–133, 2009.

[38] 5G PPP, "5G Automotive Vision," White Paper, Oct 2015. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-Automotive-Vertical-Sectors.pdf

[39] A. Festag, "Standards for vehicular communicationfrom IEEE 802.11p to 5G," *e & i Elektrotechnik und Informationstechnik*, vol. 132, no. 7, pp. 409–416, 2015.

[40] R. Martin, "Wall streets quest to process data at the speed of light," *Information Week*, vol. 4, no. 21, p. 07, 2007.

[41] L. Massoulié and J. Roberts, "Bandwidth sharing: objectives and algorithms," in *Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 3. IEEE, 1999, pp. 1395–1403.

[42] O. N. Yilmaz, Y.-P. E. Wang, N. A. Johansson, N. Brahmi, S. A. Ashraf, and J. Sachs, "Analysis of ultra-reliable and low-latency 5G communication for a factory automation use case," in *Communication Workshop, IEEE International Conference on*. IEEE, 2015, pp. 1190–1195.

[43] A. J. Goldsmith and S.-G. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE transactions on communications*, vol. 45, no. 10, pp. 1218–1230, 1997.

[44] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE transactions on communications*, vol. 43, no. 6, pp. 1986–1992, 1997.

[45] G. Wunder, P. Jung, M. Kasparick, T. Wild, F. Schaich, Y. Chen, S. Brink, I. Gaspar, N. Michailow, A. Festag *et al.*, "5GNOW: non-orthogonal, asynchronous waveforms for future mobile applications," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 97–105, 2014.

[46] P. Banelli, S. Buzzi, G. Colavolpe, A. Modenini, F. Rusek, and A. Ugolini, "Modulation formats and waveforms for 5G networks: Who will be the heir of OFDM?: An overview of alternative modulation schemes for improved spectral efficiency," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 80–93, 2014.

[47] F. Pancaldi, G. M. Vitetta, R. Kalbasi, N. Al-Dhahir, M. Uysal, and H. Mheidat, "Single-carrier frequency domain equalization," *IEEE Signal Processing Magazine*, vol. 25, no. 5, pp. 37–56, 2008.

[48] M. Luvisotto, Z. Pang, D. Dzung, M. Zhan, and X. Jiang, "Physical Layer Design of High Performance Wireless Transmission for Critical Control Applications," *IEEE Transactions on Industrial Informatics*, 2017, to be published.

[49] J. Abdoli, M. Jia, and J. Ma, "Filtered OFDM: A new waveform for future wireless systems," in *Signal Processing Advances in Wireless Communications, IEEE 16th International Workshop on*. IEEE, 2015, pp. 66–70.

[50] A. Farhang, N. Marchetti, F. Figueiredo, and J. P. Miranda, "Massive MIMO and waveform design for 5th generation wireless communication systems," in *5g for Ubiquitous Connectivity, 1st International Conference on*. IEEE, 2014, pp. 70–75.

[51] G. Fettweis, M. Krondorf, and S. Bittner, "GFDM-generalized frequency division multiplexing," in *Vehicular Technology Conference, IEEE 69th*. IEEE, 2009, pp. 1–4.

[52] G. Durisi, T. Koch, and P. Popovski, "Towards massive, ultra-reliable, and low-latency wireless: The art of sending short packets," 2015, arXiv preprint arXiv:1504.06526.

[53] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[54] J. Ostman, W. Yang, G. Durisi, and T. Koch, "Diversity versus multiplexing at finite blocklength," in *Wireless Communications Systems, 11th International Symposium on*. IEEE, 2014, pp. 702–706.

[55] G. Durisi, T. Koch, J. Ostman, Y. Polyanskiy, and W. Yang, "Short-packet communications with multiple antennas: Transmit diversity, spatial multiplexing, and channel estimation overhead," *CoRR*, 2014.

[56] D. Soldani, Y. J. Guo, B. Barani, P. Mogensen, I. Chih-Lin, and S. K. Das, "5g for ultra-reliable low-latency communications," *IEEE Network*, vol. 32, no. 2, pp. 6–7, 2018.

[57] S. Parkvall, E. Dahlman, A. Furuskar, and M. Frenne, "Nr: The new 5g radio access technology," *IEEE Communications Standards Magazine*, vol. 1, no. 4, pp. 24–30, 2017.

[58] Y. Gu, T. He, M. Lin, and J. Xu, "Spatiotemporal delay control for low-duty-cycle sensor networks," in *Real-Time Systems Symposium, 30th IEEE*. IEEE, 2009, pp. 127–137.

[59] L. G. Roberts, "ALOHA packet system with and without slots and capture," *ACM SIGCOMM Computer Communication Review*, vol. 5, no. 2, pp. 28–42, 1975.

[60] Y. Yang and T.-S. P. Yum, "Delay distributions of slotted ALOHA and CSMA," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1846–1857, 2003.

[61] K. Sohraby, D. Minoli, and T. Znati, *Wireless sensor networks: technology, protocols, and applications*. John Wiley & Sons, 2007.

[62] E. Magistretti, O. Gurewitz, and E. W. Knightly, "802.11ec: collision avoidance without control messages," *IEEE/ACM Transactions on Networking*, vol. 22, no. 6, pp. 1845–1858, 2014.

[63] H. Shokri-Ghadikolaei, C. Fischione, P. Popovski, and M. Zorzi, "Design aspects of short-range millimeter-wave networks: A MAC layer perspective," *IEEE Network*, vol. 30, no. 3, pp. 88–96, 2016.

[64] H. Shokri-Ghadikolaei and C. Fischione, "The transitional behavior of interference in millimeter wave networks and its impact on medium access control," *IEEE Transactions on Communications*, vol. 62, no. 2, pp. 723–740, Feb. 2016.

[65] R. Ramanathan, J. Redi, C. Santivanez, D. Wiggins, and S. Polit, "Ad hoc networking with directional antennas: A complete system solution," *Journal on Selected Areas in Communications*, vol. 23, no. 3, pp. 496–506, Mar. 2005.

[66] I. K. Son, S. Mao, M. X. Gong, and Y. Li, "On frame-based scheduling for directional mmWave WPANs," in *Proc. IEEE International Conference on Computer Communications*, 2012, pp. 2149–2157.

[67] S. Singh, R. Mudumbai, and U. Madhow, "Interference analysis for highly directional 60-GHz mesh networks: The case for rethinking medium access control," *IEEE/ACM Transactions on Networking*, vol. 19, no. 5, pp. 1513–1527, Oct. 2011.

[68] T. Stahlbuhk, B. Shrader, and E. Modiano, "Topology Control for Wireless Networks with Highly-Directional Antennas," in *Proc. IEEE International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, 2016.

[69] T. Nitsche, A. B. Flores, E. W. Knightly, and J. Widmer, "Steering with eyes closed: mm-wave beam steering without in-band measurement," in *Proc. IEEE Conference on Computer Communications*, 2015, pp. 2416–2424.

[70] E. Olfat, H. Shokri-Ghadikolaei, N. N. Moghadam, M. Bengtsson, and C. Fischione, "Learning-based pilot precoding and combining for wideband millimeter-wave networks," in *Proc. IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, 2017.

[71] A. Azari, P. Popovski, G. Miao, and C. Stefanovic, "Grant-Free Radio Access for Short-Packet Communications over 5G Networks," 2017, arXiv preprint arXiv:1709.02179.

[72] F. Clazzer, C. Kissling, and M. Marchese, "Exploiting Combination Techniques in Random Access MAC Protocols: Enhanced Contention Resolution ALOHA," 2016, arXiv preprint arXiv:1602.07636.

[73] V. N. Swamy, S. Suri, P. Rigge, M. Weiner, G. Ranade, A. Sahai, and B. Nikoli, "Cooperative communication for high-reliability low-latency wireless control," in *Communications, IEEE International Conference on*, June 2015, pp. 4380–4386.

[74] V. N. Swamy, P. Rigge, G. Ranade, A. Sahai, and B. Nikoli, "Network coding for high-reliability low-latency wireless control," in *Wireless Communications and Networking Conference, IEEE*, April 2016, pp. 1–7.

[75] S. Choi, G.-H. Hwang, T. Kwon, A.-R. Lim, and D.-H. Cho, "Fast handover scheme for real-time downlink services in IEEE 802.16e BWA system," in *Vehicular Technology Conference, IEEE 61st*, vol. 3. IEEE, 2005, pp. 2028–2032.

[76] D. H. Lee, K. Kyamakya, and J. P. Umondi, "Fast handover algorithm for IEEE 802.16 e broadband wireless access system," in *Wireless pervasive computing, 2006 1st international symposium on*. IEEE, 2006, pp. 6–pp.

[77] W. Jiao, P. Jiang, and Y. Ma, "Fast handover scheme for real-time applications in mobile wimax," in *Communications, IEEE International Conference on*. IEEE, 2007, pp. 6038–6042.

[78] R. Li, J. Li, K. Wu, Y. Xiao, and J. Xie, "An Enhanced Fast Handover with Low Latency for Mobile IPv6," *IEEE Transactions on Wireless Communications*, vol. 7, no. 1, pp. 334–342, 2008.

[79] L. Dimopoulou, G. Leoleis, and I. S. Venieris, "Fast handover support in a wlan environment: Challenges and perspectives," *IEEE Netw.*, vol. 19, no. 3, pp. 14–20, May 2005.

[80] H. Shokri-Ghadikolaei, C. Fischione, G. Fodor, P. Popovski, and M. Zorzi, "Millimeter wave cellular networks: A MAC layer perspective," *IEEE Transactions on Communications*, vol. 63, no. 10, pp. 3437–3458, Oct. 2015.

[81] O. Semiari, W. Saad, M. Bennis, and B. Maham, "Caching meets millimeter wave communications for enhanced mobility management in 5G networks," *IEEE Transactions on Wireless Communications*, to be published.

[82] Y. Xu, H. Shokri-Ghadikolaei, and C. Fischione, "Distributed association and relaying with fairness in millimeter wave networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 7955–7970, 2016.

[83] B. Prabhakar, E. Uysal-Biyikoglu, and A. El Gamal, "Energy-efficient transmission over a wireless link via lazy packet scheduling," in *Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*. IEEE, 2001.

[84] M. A. Zafer and E. Modiano, "A calculus approach to energy-efficient data transmission with quality-of-service constraints," *IEEE/ACM Transactions on Networking*, vol. 17, no. 3, pp. 898–911, 2009.

[85] R. L. Cruz, "A calculus for network delay-part i: Network elements in isolation," *IEEE Trans. Inform. Theory*, vol. 37, no. 1, pp. 114–131, Jan 1991.

[86] M. Zafer and E. Modiano, "Optimal rate control for delay-constrained data transmission over a wireless channel," *IEEE Transactions on Information Theory*, vol. 54, no. 9, pp. 4020–4039, 2008.

[87] R. Berry, M. Zafer, and E. Modiano, *Energy Efficient Wireless Transmission with Delay Constraints*. Morgan & Claypool, 2012.

[88] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Autom. Control*, vol. 37, no. 12, pp. 1936–1948, Dec 1992.

[89] L. X. Bui, R. Srikant, and A. Stolyar, "A novel architecture for reduction of delay and queueing structure complexity in the back-pressure algorithm," *IEEE/ACM Transactions on Networking*, vol. 19, no. 6, pp. 1597–1609, 2011.

[90] N. M. Jones, G. S. Paschos, B. Shrader, and E. Modiano, "An overlay architecture for throughput optimal multipath routing," *IEEE/ACM Transactions on Networking*, vol. 25, no. 5, pp. 2615–2628, Oct. 2017.

[91] R. Bellman, "On a routing problem," *Quart. Appl. Math.*, vol. 16, pp. 87–90, 1958.

[92] E. Dijkstra, "A note on two problems in connection with graphs," *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.

[93] R. Singh and E. Modiano, "Optimal routing for delay-sensitive traffic in overlay networks," 2017.

[94] G. R. Gupta and N. Shroff, "Delay analysis for multi-hop wireless networks," in *Proc. IEEE International Conference on Computer Communications*. IEEE, 2009, pp. 2356–2364.

[95] N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," *IEEE Transactions on Communications*, vol. 47, no. 8, pp. 1260–1267, 1999.

[96] M. J. Neely, "Delay-based network utility maximization," *IEEE/ACM Transactions on Networking*, vol. 21, no. 1, pp. 41–54, 2013.

[97] P. Di Marco, G. Athanasiou, P.-V. Mekikis, and C. Fischione, "MAC-aware Routing Metrics for the Internet of Things," *Computer Communications*, vol. 74, pp. 77–86, 2016.

[98] A. Sinha and E. Modiano, "Optimal control for generalized network-flow problems," in *Proc. IEEE International Conference on Computer Communications*. IEEE, 2017.

[99] A. El Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Optimal throughput-delay scaling in wireless networks-part i: The fluid model," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2568–2592, Jun 2006.

[100] S. Yi, Y. Pei, and S. Kalyanaraman, "On the capacity improvement of ad hoc wireless networks using directional antennas," in *Proceedings of the 4th ACM international symposium on Mobile ad hoc networking & computing*. ACM, 2003, pp. 108–116.

[101] M. Grossglauser and D. Tse, "Mobility increases the capacity of ad-hoc wireless networks," in *Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 3. IEEE, 2001, pp. 1360–1369.

[102] X. Wang, W. Huang, S. Wang, J. Zhang, and C. Hu, "Delay and capacity tradeoff analysis for motioncast," *IEEE/ACM Transactions on Networking*, vol. 19, no. 5, pp. 1354–1367, 2011.

[103] R. Talak, S. Karaman, and E. Modiano, "Capacity and delay scaling for broadcast transmission in highly mobile wireless networks," in *ACM MobiHoc*. ACM, 2017.

[104] M. J. Neely and E. Modiano, "Capacity and delay tradeoffs for ad-hoc mobile networks," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1917–1937, Jun 2005.

[105] T. He, J. A. Stankovic, C. Lu, and T. Abdelzaher, "SPEED: A real-time routing protocol for sensor networks," Virginia Univ Charlottesville Dept of Computer Science, Tech. Rep., 2002.

[106] E. Felemban, C.-G. Lee, and E. Ekici, "Mmspeed: Multipath multi-speed protocol for qos guarantee of reliability and timeliness in wireless sensor networks," *IEEE Trans. Mobile Comput.*, vol. 5, no. 6, pp. 738–754, Jun 2006.

[107] B. Zhang and H. T. Mouftah, "QoS routing for wireless ad hoc networks: Problems, algorithms, and protocols," *IEEE Communications Magazine*, vol. 43, no. 10, pp. 110–117, 2005.

[108] C. E. Perkins and E. M. Royer, "Ad hoc on-demand distance vector routing," in *Proc. Workshop Mobile Computing Systems Applications (WMCSA)*, Feb 1999, pp. 90–100.

[109] C. R. Lin, "On-demand QoS routing in multihop mobile networks," in *Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 3. IEEE, 2001, pp. 1735–1744.

[110] L. Hanzo and R. Tafazolli, "A survey of QoS routing solutions for mobile ad hoc networks," *IEEE Communications Surveys & Tutorials*, vol. 9, no. 2, pp. 50–70, 2007.

[111] B. Paul and M. J. Islam, "Survey over VANET routing protocols for vehicle to vehicle communication," *IOSR Journal of Computer Engineering, ISSN*, pp. 2278–0661, 2012.

[112] M. Altayeb and I. Mahgoub, "A survey of vehicular ad hoc networks routing protocols," *International Journal of Innovation and Applied Studies*, vol. 3, no. 3, pp. 829–846, 2013.

[113] D. Comer, "Interworking with TCP/IP: Principles, Protocols and Architecture Vol. 1," 2006.

[114] S. Floyd, "TCP and explicit congestion notification," *ACM Comput. Commun. Rev.*, vol. 24, no. 5, pp. 8–23, 1994.

[115] R. Adams, "Active Queue Management: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1425–1476, 2013.

[116] K. Nichols and V. Jacobson, "Controlling queue delay," *Communications of the ACM*, vol. 55, no. 7, pp. 42–50, 2012.

[117] M. Alizadeh, A. Kabbani, T. Edsall, B. Prabhakar, A. Vahdat, and M. Yasuda, "Less is more: trading a little bandwidth for ultra-low latency in the data center," in *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. USENIX Association, 2012, pp. 19–19.

[118] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 5, pp. 684–702, Jun 2003.

[119] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 461–471, 2004.

[120] T. Yoo and A. J. Goldsmith, "Capacity and optimal power allocation for fading mimo channels with channel estimation error," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2203–2214, May 2006.

[121] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, 2010.

[122] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, 2013.

[123] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge University Press, 2005.

[124] H. Huh, A. M. Tulino, and G. Caire, "Network MIMO with linear zero-forcing beamforming: Large system analysis, impact of channel estimation, and reduced-complexity scheduling," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 2911–2934, 2012.

[125] K. T. K. Cheung, S. Yang, and L. Hanzo, "Spectral and energy spectral efficiency optimization of joint transmit and receive beamforming based multi-relay mimo-ofdma cellular networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 11, pp. 6147–6165, 2014.

[126] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Coordinated beamforming for multi-cell mimo-noma," *IEEE Communications Letters*, vol. 21, no. 1, pp. 84–87, 2017.

[127] S. He, Y. Huang, L. Yang, B. Ottersten, and W. Hong, "Energy efficient coordinated beamforming for multicell system: Duality-based algorithm design and massive mimo transition," *IEEE Transactions on Communications*, vol. 63, no. 12, pp. 4920–4935, 2015.

[128] H. Q. Ngo, L. N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the Total Energy Efficiency of Cell-Free Massive MIMO," *IEEE Transactions on Green Communications and Networking*, 2017.

[129] J. Zhang and J. G. Andrews, "Adaptive spatial intercell interference cancellation in multicell wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 9, pp. 1455–1468, 2010.

[130] K. Hosseini, W. Yu, and R. S. Adve, "Large-scale MIMO versus network MIMO for multicell interference mitigation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 930–941, 2014.

[131] S. Huberman and T. Le-Ngoc, "Mimo full-duplex precoding: A joint beamforming and self-interference cancellation structure," *IEEE Transactions on Wireless Communications*, vol. 14, no. 4, pp. 2205–2217, 2015.

[132] Y. Ghasempour, C. R. da Silva, C. Cordeiro, and E. W. Knightly, "IEEE 802.11 ay: Next-Generation 60 GHz Communication for 100 Gb/s Wi-Fi," *IEEE Communications Magazine*, vol. 55, no. 12, pp. 186–192, 2017.

[133] H. Shokri-Ghadikolaei, L. Gkatzikis, and C. Fischione, "Beam-searching and transmission scheduling in millimeter wave communications," in *Communications IEEE International Conference on*. IEEE, 2015, pp. 1292–1297.

[134] H. Shokri-Ghadikolaei, F. Boccardi, C. Fischione, G. Fodor, and M. Zorzi, "Spectrum sharing in mmWave cellular networks via cell association, coordination, and beamforming," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 11, pp. 2902–2917, Nov. 2016.

[135] Y. Niu, Y. Li, D. Jin, L. Su, and D. Wu, "Blockage Robust and Efficient Scheduling for Directional mmWave WPANs," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 2, pp. 728–742, Feb. 2015.

[136] A. Kinalis, S. Nikoletseas, D. Patroumpa, and J. Rolim, "Biased sink mobility with adaptive stop times for low latency data collection in sensor networks," *Information fusion*, vol. 15, pp. 56–63, 2014.

[137] F. Boccardi, H. Shokri-Ghadikolaei, G. Fodor, E. Erkip, C. Fischione, M. Kountoris, P. Popovski, and M. Zorzi, "Spectrum pooling in mmWave networks: Opportunities, challenges, and enablers," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 33–39, Nov. 2016.

[138] V. Petrov, M. Komarov, D. Moltchanov, J. M. Jornet, and Y. Koucheryavy, "Interference and SINR in millimeter wave and terahertz communication systems with blocking and directional antennas," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1791–1808, Mar. 2017.

[139] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 436–453, 2016.

[140] C. Zhang, D. Guo, and P. Fan, "Tracking angles of departure and arrival in a mobile millimeter wave channel," in *Communications IEEE International Conference on*, May 2016, pp. 1–6.

[141] J. Zhao, F. Gao, W. Jia, S. Zhang, S. Jin, and H. Lin, "Angle domain hybrid precoding and channel tracking for millimeter wave massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 10, pp. 6868–6880, 2017.

[142] R. Ford, M. Zhang, M. Mezzavilla, S. Dutta, S. Rangan, and M. Zorzi, "Achieving ultra-low latency in 5G millimeter wave cellular networks," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 196–203, 2017.

[143] D. Soldani and A. Manzalini, "Horizon 2020 and beyond: on the 5G operating system for a true digital society," *IEEE Vehicular Technology Magazine*, vol. 10, no. 1, pp. 32–42, 2015.

[144] J. I. Choi, M. Jain, K. Srinivasan, P. Levis, and S. Katti, "Achieving single channel, full duplex wireless communication," in *Proceedings of the sixteenth annual international conference on Mobile computing and networking*. ACM, 2010, pp. 1–12.

[145] Z. Zhang, K. Long, A. V. Vasilakos, and L. Hanzo, "Full-duplex wireless communications: challenges, solutions, and future research directions," *Proceedings of the IEEE*, vol. 104, no. 7, pp. 1369–1409, 2016.

[146] Z. Li, M. Moisio, M. A. Uusitalo, P. Lundén, C. Wijting, F. Sanchez Moya, A. Yaver, and V. Venkatasubramanian, "Overview on initial METIS D2D concept," in *5G for Ubiquitous Connectivity, 1st International Conference on*. IEEE, 2014, pp. 203–208.

[147] Z. Li, M. Moisio, M. A. Uusitalo, P. Lundén, C. Wijting, F. S. Moya, A. Yaver, and V. Venkatasubramanian, "Overview on initial metis d2d concept," in *5G for Ubiquitous Connectivity (5GU), 2014 1st International Conference on*. IEEE, 2014, pp. 203–208.

[148] P. Gandotra and R. K. Jha, "Device-to-device communication in cellular networks: A survey," *Journal of Network and Computer Applications*, vol. 71, pp. 99–117, 2016.

[149] J. Qiao, X. S. Shen, J. W. Mark, Q. Shen, Y. He, and L. Lei, "Enabling device-to-device communications in millimeter-wave 5g cellular networks," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 209–215, 2015.

[150] Y. Niu, L. Su, C. Gao, Y. Li, D. Jin, and Z. Han, "Exploiting device-to-device communications to enhance spatial reuse for popular content downloading in directional mmwave small cells," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 7, pp. 5538–5550, 2016.

[151] C. Chen, "C-RAN: the Road Towards Green Radio Access Network," White Paper, 2011.

[152] P. Chanclou, A. Pizzinat, F. Le Clech, T.-L. Reedeker, Y. Lagadec, F. Saliou, B. Le Guyader, L. Guillo, Q. Deniel, S. Gosselin *et al.*, "Optical fiber solution for mobile fronthaul to achieve cloud radio access network," in *Future Network and Mobile Summit*. IEEE, 2013, pp. 1–11.

[153] N. Cvijetic, A. Tanaka, K. Kanonakis, and T. Wang, "SDN-controlled topology-reconfigurable optical mobile fronthaul architecture for bidirectional CoMP and low latency inter-cell D2D in the 5G mobile era," *Optics express*, vol. 22, no. 17, pp. 20 809–20 815, 2014.

[154] H. Beyranvand, M. Lévesque, M. Maier, and J. A. Salehi, "FiWi enhanced LTE-A HetNets with unreliable fiber backhaul sharing and WiFi offloading," in *Proc. IEEE International Conference on Computer Communications*. IEEE, 2015, pp. 1275–1283.

[155] A. S. Tanenbaum and D. Wetherall, *Computer networks*. Prentice hall, 1996.

[156] I. Tinnirello and S. Choi, "Efficiency analysis of burst transmissions with block ack in contention-based 802.11 e wlans," in *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*, vol. 5. IEEE, 2005, pp. 3455–3460.

[157] C. Shanti and A. Sahoo, "DGRAM: a delay guaranteed routing and MAC protocol for wireless sensor networks," *IEEE Transactions on Mobile Computing*, vol. 9, no. 10, pp. 1407–1423, 2010.

[158] I. Chih-Lin, S. Han, Z. Xu, S. Wang, Q. Sun, and Y. Chen, "New Paradigm of 5G Wireless Internet," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 474–482, 2016.

[159] D. Kreutz, F. M. Ramos, P. Esteves Verissimo, C. Esteve Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: A comprehensive survey," *Proceedings of the IEEE*, vol. 103, no. 1, pp. 14–76, 2015.

[160] "Network Functions Virtualisation (NFV); Architectural Framework," ETSI, Tech. Rep., 2016.

[161] N. F. Virtualisation, "NFV performance & portability best practises," ETSI, Tech. Rep., 2014.

[162] B. Briscoe, "Network Functions Virtualisation (NFV)-NFV Security: Problem Statement," ETSI, Tech. Rep., 2014.

[163] M. S. Siddiqui, A. Legarrea, E. Escalona, M. C. Parker, G. Koczian, S. D. Walker, G. Lyberopoulos, E. Theodoropoulou, K. Filis, A. Foglar *et al.*, "Hierarchical, virtualised and distributed intelligence 5G architecture for low-latency and secure applications," *Transactions on Emerging Telecommunications Technologies*, vol. 27, no. 9, pp. 1233–1241, 2016.

[164] T. Kurimoto, S. Urushidani, H. Yamada, K. Yamanaka, M. Nakamura, S. Abe, K. Fukuda, M. Koibuchi, H. Takakura, S. Yamada *et al.*, "SINET5: A low-latency and high-bandwidth backbone network for SDN/NFV Era," in *Communications, IEEE International Conference on*. IEEE, 2017, pp. 1–7.

[165] Y. Rekhter, T. Li, and S. Hares, "A border gateway protocol 4 (BGP-4)," Tech. Rep., 2005.

[166] A. Arins, "Latency factor in worldwide ip routed networks," in *Information, Electronic and Electrical Engineering, 2014 IEEE 2nd Workshop on Advances in*. IEEE, 2014, pp. 1–4.

[167] H. E. Egilmez, S. Civanlar, and A. M. Tekalp, "A distributed QoS routing architecture for scalable video streaming over multi-domain OpenFlow networks," in *Image Processing, 19th IEEE International Conference on*. IEEE, 2012, pp. 2237–2240.

[168] P. Lin, J. Bi, S. Wolff, Y. Wang, A. Xu, Z. Chen, H. Hu, and Y. Lin, "A west-east bridge based SDN inter-domain testbed," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 190–197, 2015.

[169] F. Paolucci, V. Uceda, A. Sgambelluri, F. Cugini, O. G. De Dios, V. Lopez, L. Contreras, P. Monti, P. Iovanna, F. Ubaldi *et al.*, "Interoperable Multi-Domain Delay-aware Provisioning using Segment Routing Monitoring and BGP-LS Advertisement," in *42nd European Conference on Optical Communication; Proceedings of*. VDE, 2016, pp. 1–3.

[170] K.-S. Lui, K. Nahrstedt, and S. Chen, "Routing with topology aggregation in delay-bandwidth sensitive networks," *IEEE/ACM Transactions on Networking*, vol. 12, no. 1, pp. 17–29, Feb 2004.

[171] R. Singh and E. Modiano, "Optimal routing for delay-sensitive traffic in overlay networks," *ArXiv e-print arXiv:1703.07419*, Mar. 2017.

[172] X. Ji, "Models and algorithm for stochastic shortest path problem," *Applied Mathematics and Computation*, vol. 170, no. 1, pp. 503–514, 2005.

[173] M. S. Talebi, Z. Zou, R. Combes, A. Proutiere, and M. Johansson, "Stochastic online shortest path routing: The value of feedback," *IEEE Transactions on Automatic Control*, to be published.

[174] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *INFOCOM, 2010 Proceedings IEEE*. IEEE, 2010, pp. 1–9.

[175] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[176] ——, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1029–1040, 2015.

[177] K. N. Doan, T. Van Nguyen, T. Q. Quek, and H. Shin, "Content-aware proactive caching for backhaul offloading in cellular network," *IEEE Transactions on Wireless Communications*, vol. 17, no. 5, pp. 3128–3140, 2018.

[178] B. Tan and L. Massoulié, "Optimal content placement for peer-to-peer video-on-demand systems," *IEEE/ACM Trans. Netw.*, vol. 21, no. 2, pp. 566–579, 2013.

[179] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE International Conference on Computer Communications*. IEEE, 2012, pp. 1107–1115.

[180] E. Bastug, M. Bennis, and M. Debbah, "Living on the Edge: The role of proactive caching in 5G wireless networks," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 82–89, 2014.

[181] S. Bhattacharjee, K. L. Calvert, and E. W. Zegura, "Self-organizing wide-area network caches," in *INFOCOM'98. Seventeenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 2. IEEE, 1998, pp. 600–608.

[182] B. Li, M. J. Golin, G. F. Italiano, X. Deng, and K. Sohraby, "On the optimal placement of web proxies in the internet," in *INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 3. IEEE, 1999, pp. 1282–1290.

[183] "Fog Computing and the Internet of Things: Extend the Cloud to Where the Things Are," 2015.

[184] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the first edition of the workshop on Mobile cloud computing*. ACM, 2012, pp. 13–16.

[185] F. Bonomi, R. Milito, P. Natarajan, and J. Zhu, "Fog computing: A platform for internet of things and analytics," in *Big Data and Internet of Things: A Roadmap for Smart Environments*. Springer, 2014, pp. 169–186.

[186] Q. Li, H. Niu, A. Papathanassiou, and G. Wu, "Edge cloud and underlay networks: Empowering 5G cell-less wireless architecture," in *European Wireless; 20th European Wireless Conference; Proceedings of*. VDE, 2014, pp. 1–6.

[187] N. Bessis and C. Dobre, *Big data and internet of things: a roadmap for smart environments*. Springer, 2014, vol. 546.

[188] H. Gupta, A. Vahid Dastjerdi, S. K. Ghosh, and R. Buyya, "ifogsim: A toolkit for modeling and simulation of resource management techniques in the internet of things, edge and fog computing environments," *Software: Practice and Experience*, vol. 47, no. 9, pp. 1275–1296, 2017.

[189] M. Aazam and E.-N. Huh, "Fog computing and smart gateway based communication for cloud of things," in *Future Internet of Things and*

Cloud (FiCloud), 2014 International Conference on. IEEE, 2014, pp. 464–470.

[190] A. A. Zaidi, R. Baldemair, H. Tullberg, H. Bjorkegren, L. Sundstrom, J. Medbo, C. Kilinc, and I. Da Silva, "Waveform and numerology to support 5G services and requirements," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 90–98, 2016.

[191] S. Mumtaz, A. Alsohaily, Z. Pang, A. Rayes, K. F. Tsang, and J. Rodriguez, "Massive internet of things for industrial applications: Addressing wireless iiot connectivity challenges and ecosystem fragmentation," *IEEE Industrial Electronics Magazine*, vol. 11, no. 1, pp. 28–33, 2017.

[192] D. De Guglielmo, S. Brienza, and G. Anastasi, "IEEE 802.15. 4e: A survey," *Computer Communications*, vol. 88, pp. 1–24, 2016.

[193] "LTE-like performance with Wi-Fi-like simplicity," MulteFire, 2015.

[194] M. Labib, V. Marojevic, J. H. Reed, and A. I. Zaghloul, "Extending LTE into the unlicensed spectrum: technical analysis of the proposed variants," 2017, arXiv preprint arXiv:1709.04458.

[195] "IEEE TSN (Time-Sensitive Networking): A Deterministic Ethernet Standard," TTTech, 2015.

[196] S. Tasaka, "Multiple-access protocols for satellite packet communication networks: A performance comparison," *Proc. IEEE*, vol. 72, no. 11, pp. 1573–1582, Nov 1984.

[197] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmWave frequencies," *arXiv preprint arXiv:1804.01908*, 2018.

[198] T. Levanen, J. Pirskanen, and M. Valkama, "Radio interface design for ultra-low latency millimeter-wave communications in 5G era," in *Globecom Workshops*. IEEE, 2014, pp. 1420–1426.

[199] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. IEEE International Conference on Computer Communications*. IEEE, 2012.

[200] L. Huang and E. Modiano, "Optimizing age-of-information in a multiclass queueing system," in *IEEE International Symposium on Information Theory*. IEEE, 2015.

[201] I. Kadota, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, "Minimizing the age of information in broadcast wireless networks," in *Communication, Control, and Computing, 54th Annual Allerton Conference on*. IEEE, 2016.

[202] *Optimizing Information Freshness in Wireless Networks under General Interference Constraints*. ACM, 2018.

[203] R. Talak, S. Karaman, and E. Modiano, "Minimizing age-of-information in multi-hop wireless networks," in *Communication, Control, and Computing (Allerton), 2017 55th Annual Allerton Conference on*. IEEE, 2017, pp. 486–493.

[204] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, 2014.

[205] O. Goldreich, *Foundations of cryptography: volume 2, basic applications*. Cambridge university press, 2009.

[206] W. K. Harrison, J. Almeida, M. R. Bloch, S. W. McLaughlin, and J. Barros, "Coding for secrecy: An overview of error-control coding techniques for physical-layer security," *IEEE Signal Processing Magazine*, vol. 30, no. 5, pp. 41–50, 2013.

[207] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[208] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[209] L. Sankar, W. Trappe, K. Ramchandran, H. V. Poor, and M. Debbah, "The role of signal processing in meeting privacy challenges: An overview," *IEEE Signal Processing Magazine*, vol. 30, no. 5, pp. 95–106, 2013.