IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 19, 2017, accepted May 11, 2017, date of publication May 23, 2017, date of current version June 27, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2707319

Soft Resource Reservation for Low-Delayed Teleoperation Over Mobile Networks

MASSIMO CONDOLUCI[1], (Member, IEEE), TOKTAM MAHMOODI[1], (Senior Member, IEEE), ECKEHARD STEINBACH[2], (Fellow, IEEE), AND MISCHA DOHLER[1], (Fellow, IEEE)

[1]Centre for Telecommunications Research, Department of Informatics, King's College London, London WC2R 2LS, U.K.
[2]Chair of Media Technology, Technical University of Munich, 80333 Munich, Germany

Corresponding author: Toktam Mahmoodi (toktam.mahmoodi@kcl.ac.uk)

ABSTRACT The emerging Tactile Internet (TI) will enable control-oriented networks for remotely accessing or manipulating objects or devices. One major challenge in this context is how to achieve ultra-low-delay communication between the local operator and the remote object/device to guarantee the stability of the global control loop and to maximize the user's quality-of-experience (QoE). Being one of the major human-in-the-loop applications of the TI, haptic teleoperation inherits its delay-sensitive nature and requires the orchestration of communication and control approaches. In this paper, we focus on the radio access protocol, and its impact on the latency of wireless communication. We propose a novel soft resource reservation mechanism for the uplink scheduling of mobile networks that can significantly reduce the latency compared with the current legacy scheme. By leveraging the characteristics of teleoperation data traffic, and reserving resources accordingly, the proposed soft reservation scheme maintains the spectral efficiency while the human operator's QoE is improved. The simulation results confirm the efficiency of the proposed scheme.

INDEX TERMS Teleoperation, QoE, resource allocation, resource reservation, scheduling, Tactile Internet.

I. INTRODUCTION

The past decade has witnessed a tremendous growth of the Mobile Internet which connects millions of mobile devices on a global scale. More recently, we witnessed the emergence of the Internet of Things [1] which enables the transition from the network of mobile communication devices to the network of billions of physical devices, objects, animals, and human beings. These different Internet embodiments embrace the rise of the Tactile Internet (TI) [2], [3], which aims at providing ultra-low-delay and ultra-high-reliable communications enabling a paradigm shift from conventional content-oriented communication to a "*control*"-oriented communication. The TI is of particular relevance for the realization of "*human-in-the-loop*" remote teleoperation applications which are highly delay sensitive and require a tight integration of the communication and control mechanisms.

Teleoperation systems allow a human user to immerse into a distant or inaccessible environment to perform complex tasks, i.e., the human user operates remotely without the need of being physically located where the operation is taking place. A typical haptic teleoperation system enhances a legacy teleoperation system by introducing haptic feedback (forces, torques, position, velocity) allowing the human teleoperator to improve the knowledge of the environment where he/she is teleoperating. This means that, in a haptic session, each teleoperator's action will generate a haptic feedback that changes what the teleoperation is feeling. As illustrated in Fig. 1, typical haptic teleoperation system comprises a slave and a master device, which exchange haptic information, video signals, and audio signals over a communication network [4]. In particular, the bidirectional communication of haptic information (position/velocity and force/torque signals) imposes strong demands on the communication network as it closes a global control loop between the operator and the remote robot. As a result, the system stability is highly sensitive to communication delay [5].

In addition, high-fidelity teleoperation requires a sampling rate of 1 kHz or even higher for haptic signals to ensure a high quality interaction and system stability. In order to keep the communication delay as small as possible, haptic sensor readings are typically packetized and transmitted once available. Teleoperation systems, hence, require 1000 or more haptic data packets/s to be transmitted in both directions between the master and the slave device. For Internet-based communications, stability for such high packet rates is hard to be maintained.

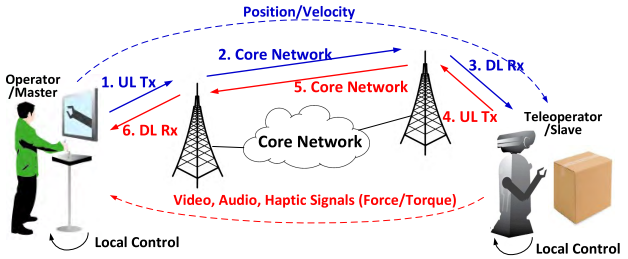

FIGURE 1. Illustration of a bilateral haptic teleoperation session over mobile networks.

State-of-the-art solutions that address the aforementioned teleoperation challenges (low delay and high packet rate) are based on the assumption that the communication latency can not be controlled, and focus on combining different stability-ensuring control schemes with haptic packet rate reduction methods [6], [7].

Although teleoperation systems could be considered as yet another domain in machine-type communications (MTC), their traffic pattern is strongly different from typical MTC traffic. Generally, MTC traffic is characterized by small packets, short-lived flows and storms of access requests [8]. In contrast, the haptic traffic, while depending on the used control scheme, has a different nature. Existing experimentation, for example, shows bursty and irregular patterns of data traffic between the haptic master and slave device [9].

To this end, we thoroughly study the pros and cons of a number of resource request and allocation solutions, such as the scheduling request (SR) [10], [11], the semi-persistent scheduling [12], [13], and the contention-based approach [14], in terms of addressing the specific needs of the haptic data traffic, and the corresponding control scheme. Since the quality-of-experience (QoE) of the human operator degrades for increasing communication delay, the design of transmission/reception procedures should by all means explore the possibility to reduce the communication delay in order to achieve a satisfactory QoE performance during teleoperation.

In this paper, we propose a *soft resource reservation* in the SR procedure of mobile networks, in which the uplink (UL) grant from one transmission is softly reserved for the following transmissions, and therefore the latency is significantly reduced while the spectral efficiency is maintained, thanks to the "*soft*" nature of the resource reservation (this is in strong contrast to e.g., the semi-persistent scheduler). The key contributions of our proposal can be summarized as follows:

- Through extensive analysis, we show that the proposed soft resource reservation mechanism achieves a noticeable round-trip latency improvement. This means that our scheduler is able to improve the QoE of the human teleoperator in bilateral haptic teleoperation scenarios.
- With the exploitation of the characteristics of haptic traffic, the proposed scheduler is able to provide a satisfactory QoE performance for teleoperation without involving inefficiency from a resource management point of view.

The remainder of this paper is organized as follows. In Section II, we briefly demonstrate the characteristics of the haptic traffic in bilateral teleoperation systems, in order to derive the features to be considered for the design of haptic-oriented scheduling and resource allocation procedures. Section III provides a thorough comparative study on the legacy procedures for UL and downlink (DL) transmission, highlighting pros and cons of other approaches proposed in literature. In Section IV, we present a model to analyze UL and DL latencies, and we propose a novel strategy for low latency communications. The performance is studied in Section V. Research background is provided in Section VI. Finally, concluding remarks are presented in Section VII.

II. CHARACTERISTICS OF TELEOPERATION TRAFFIC

In order to handle the teleoperation traffic efficiently, transmission procedures need to take into account characteristics of the traffic generated by this application. For this, we recovered traffic traces during teleoperation experiments using the setup explained in [9]. The teleoperation experiments conducted in [9] exploits a real Phantom Omni haptic device as a master device, connected to a slave device realized in a virtual environment (VE) developed based on the Chai3D library. The slave in the VE acts as a single haptic interaction point with negligible mass. In these experiments, packets on both directions (master-to-slave and slave-to-master) are 24 bytes. A variable delay between master and slave is emulated to represent the communication network in different conditions, ranging from master and slave located in the same local network (with an ideal latency of 0ms) to the case when the two sides are located in different continents (emulated with a latency of 200ms).

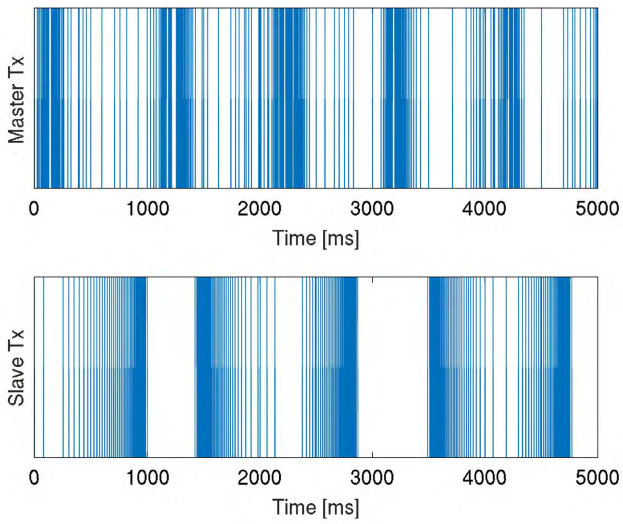

FIGURE 2. An example of the forward and backward haptic traffic over a time window of 5s from one of the subjects in the experiment in [9], for the case when latency is set to 0ms.

We conducted some analyses based on the traces of the experiments in [9]. Fig. 2 shows the forward/backward haptic

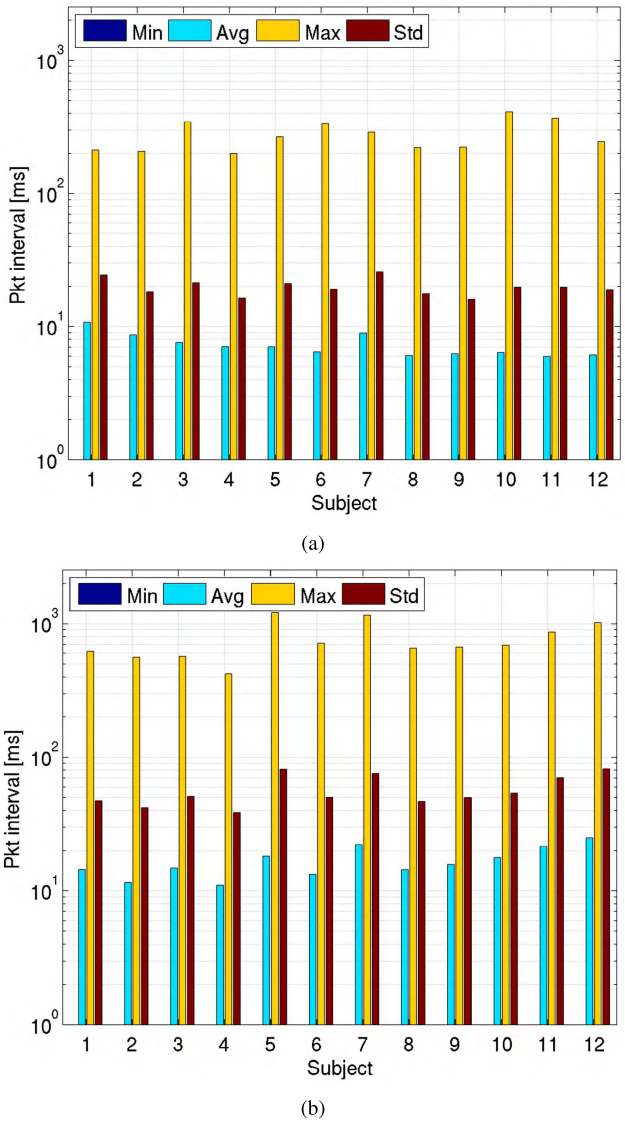

FIGURE 3. Packet inter-arrival times analyzed from the traces of the experiments in [9]. (a) Forward (master-to-slave). (b) Backward (slave-to-master).

traffic over a time window of 5s, where the y-axis shows when a packet transmission occurs. Fig. 3 analyzes the packet inter-arrival time at both master and slave devices by analyzing the experiments of 12 different subjects (x-axis). From both figures, we can observe the bursty behavior and the irregular updates between the master and the slave. However, the packet update rate of the backward channel (force feedback from slave-to-master) is smaller.

Observing from this analysis, teleoperation traffic requires the allocation of resources with a very short interval during bursty periods, where the lowest packet inter-arrival time is 1ms as shown in Fig. 3. During the non-bursty intervals, where teleoperator is not in contact with the remote environment (seen in Fig. 2), packet inter-arrival times can reach values in the range 200ms-300ms (seen in Fig. 3).

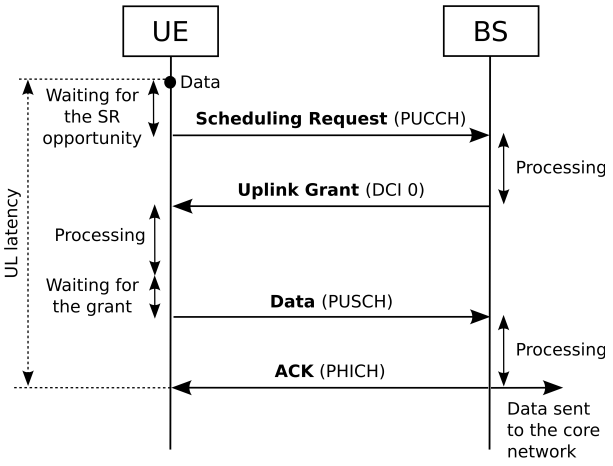

FIGURE 4. The scheduling request procedure in LTE.

Hence, flexibility in the resource allocation to quickly provide resources during bursty intervals while saving resources during non-bursty intervals, can play a crucial role in timely delivery of the traffic information system and maintaining efficiency of wireless resource utilization.

III. DATA TRANSMISSION AND RECEPTION

In this section, we focus on analyzing different sources of delay in mobile communications in order to identify how those delays could be reduced/removed for teleoperation traffic using the insight from the teleoperation traffic pattern.

Focusing on the bidirectional master-slave communication depicted in Fig. 1, the different components of the delay in the haptic session are: *uplink (UL) transmission*, for the transmission of packets from either the master or the slave to the base station; *downlink (DL) transmission*, for the reception of packets sent from the base station to either master or slave devices; *core network*, exploited to inter-connect the base stations involved in the session.

Within the radio access, the transmission of packets in the UL direction is performed by means of two mechanisms: random access or schedule request. The random access procedure [15] is performed when the device is not synchronized with or connected to the network. This procedure is, thus, usually performed for the initial attachment of the device to the network. In the haptic session, this procedure happens only once, i.e., during the session setup phase, and it does not affect the data transmission from both master and slave sides. On the contrary, the scheduling request (SR) [10], [11] is performed when a device has an active connection and it needs to request resources for UL transmission to the base station. The SR procedure is depicted in Fig. 4. The device is configured by the base station with a SR period, varying from 1ms up to 320ms. In case the device has data to transmit, it should wait until the next available SR opportunity in order to inform the base station about the amount of data in its buffer. To avoid congestion at the base station, the *sr-ProhibitTimer-r9* (whose duration can vary from 0 to 7 SR

periods) has been introduced to avoid too many requests to be sent in subsequent SR opportunities. At the reception of the SR, the base station processes the incoming request in order to allocate the resource blocks (RBs) to the device, then it sends an UL grant to the device with the information about the allocated resources. Finally, the device processes the received grant and sends its data on the RBs the base station has allocated to its transmission.[1]

As analyzed in [14] and [19], the SR procedure can guarantee (with a given margin) deterministic delays and it is thus capable of guaranteeing a stable latency for haptic communications. The SR procedure, however, introduces a considerable delay due to its own nature, as explained. The SR is a device-to-base station handshake procedure, and can maintain high spectral efficiency at the expense of the latency in data transmission. As mentioned above, recent advances from 3rd Generation Partnership Project (3GPP) aim to reduce the latency of this procedure by bringing the SR periodicity down to 1ms. This will cut the waiting delay before triggering the SR procedure, but not the delay experienced during the procedure itself. One possible solution to cut the SR delay is moving to semi-persistent scheduling, which is based on the idea of statically allocating resources with a given periodicity to the device [12], [13]. Therefore, the device can send its data without performing the SR and thus cutting the transmission delay. Semi-persistent scheduling is known as an effective way to reduce delays for traffic with deterministic behavior, such as VoIP, where the device generates one voice packet each 20 ms [20]. On the other hand, haptic traffic can drastically vary during the session, according to the plots in Fig. 3. Hence, semi-persistent scheduling would involve wasting of resources during the long off-period of the haptic session, since RBs assigned to the haptic session cannot be exploited by any other devices and this would drastically decrease the spectral efficiency. An alternative solution for avoiding the SR delay is the introduction of a novel contention-based channel, as proposed in [14]. In this case, instead of having resources pre-scheduled for each device, the network assigns resources to be exploited by all haptic devices. These resources, with pre-defined modulation and coding scheme (MCS) thus carrying pre-defined packet size, will be shared among the haptic devices that are active in the cell. In other words, any device with a packet to be sent, waits until the next available resource and then transmit its packet. The main drawback of this approach, however, is the possibility of collision in case multiple devices select the same RB(s) to send their data. This aspect is exacerbated in haptic session when devices send packets every 1ms during their bursty interval, as observed in Fig. 3. This means that, to effectively avoid collisions and thus retransmissions during the bursty interval, the contention-based approach needs to reserve a significant amount of RBs which is spectrally inefficient.

[1] Another aspect to consider is the HARQ procedure, exploited to inform the device about the effective reception of the transmitted signal and to trigger a re-transmission in case of a failure. From a haptic session point of view, this procedure introduces delays in case of re-transmissions. Nevertheless, solutions such as [16]–[18] can be exploited to improve the transmission reliability for haptic sessions by exploiting the static nature of the involved devices. In this paper, the delay associated with the HARQ procedure is not considered, since the main focus is on the analysis of data transmission/reception strategies on the radio segment.

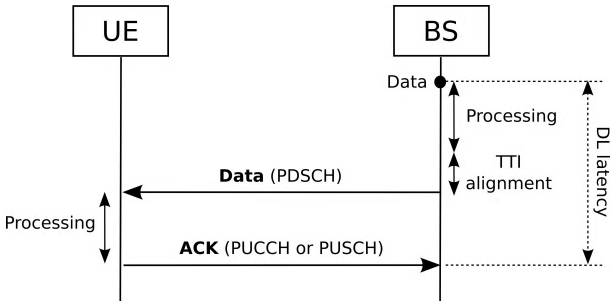

FIGURE 5. Downlink reception procedure in LTE.

Despite the explained latency in the UL direction, the DL data communication is considerably shorter and the procedure is less complex than UL. Fig. 5 shows the DL procedure in LTE [19]. Since the base station is already aware of the amount of data to be delivered to the device, it only needs to inform the device about the assigned resources and subsequently send the data. At the core network, delay can vary between 1 ms and 20 ms based on the 3GPP study on latency reduction in the LTE [19]. There has been various efforts in reducing the core network and the end-to-end latency by either moving radio access functionalities to the cloud [21] or by deploying the concepts of edge-cloud and optimal placement of virtual network functions [22] across the end-to-end path. In this paper, however, we focus on reducing latency caused by the transmission procedures at the radio access, and given the significance of UL latency, we devise our attention to the UL. Hence, a novel solution for the UL SR is discussed in the next section.

IV. SYSTEM MODEL

We consider a scenario with K master-slave pairs. For the sake of simplicity, we assume all K masters are co-located and are served by one base station while K slaves are served by another base station. An illustration of our deployment scenario with one master-slave pair can be seen in Fig. 1. The notations used in the paper are listed in Table 1.

We focus our attention on the transmission and reception procedures within one cell, with R_{DL} and R_{UL} resource blocks reserved, respectively, for data reception in DL and data transmission in UL. The transmission time on the radio interface is denoted with T_{TX}, and the time needed for processing received data is given by T_{PRO}. By considering a generic device k within the cell, s_k and σ_k represent the size of the packet to be transmitted/received, and the experienced SINR, respectively. Finally, $\mathcal{B}_k$ indicates the transmission buffer for device k for uplink transmission while, for downlink reception, it indicates the buffer at the base station with the packets to be delivered to device k. The amount of resource

TABLE 1. List of notations

Notation	Definition	Value
K	Number of devices	-
N	Maximum number of SRs that base station can handle simultaneously	18 [19]
R_{DL}	Number of DL RBs	25
R_{UL}	Number of UL RBs	25
T_{SR}	SR period	5ms [19]
T_{DL}	DL scheduling period	5ms [19]
T_{TX}	Transmission time	1ms [19]
T_{PRO}	Processing time	1ms [19]
T_{Al}	TTI alignment	0.5ms [19]
$\mathcal{B}_k$	Buffer for device k with the packets to be sent (uplink) or to be received (downlink)	-
$\mathcal{A}_{k,t}$	Set of packets of device k at the t-th time slot	-
$\mathcal{D}_{k,t}$	Set of packets device k transmits/receives in the t-th time slot	-
σ_k	SINR device k	-
s_k	Packet size	24B Master/Slave
r_k	Number of RBs needed by device k	-
p_k	Counter of packets for device k	-
$\alpha_{k,t}$	Binary value equal to 1 if device k performs the SR in the t-th time slot	(1)
$\beta_{k,t}$	Binary value equal to 1 if device k receive the UL grant in the t-th time slot	(3)
$\gamma_{k,t}$	Binary value equal to 1 if device k transmits its data in the t-th time slot	(5)
$\delta_{k,t}$	Binary value equal to 1 if the base station has a packet for device k at the t-th time slot	0,1
$\eta_{k,t}$	Binary value equal to 1 if the base station schedules device k for DL reception at the t-th time slot	(8)
$\lambda_{k,p}$	Transmission latency of packet p for device k	(6) and (9)

blocks needed to receive the scheduled packets is given by r_k.

A. LEGACY DATA TRANSMISSION

The legacy data transmission is performed by means of the SR procedure, depicted in Fig. 4. This procedure allows the devices to inform the base station about the status of their buffers, and the scheduling requests can be sent from the devices with a periodicity equal to T_{SR}.

The data transmission buffer at device k, i.e., $\mathcal{B}_k$, contains the list of the packets to be sent.

With the aim of distinguishing the packets, we exploit a counter denoted with p_k (starting from 0), which allow us to trace and measure the data transmission delay for each packet. When a packet reaches the data transmission buffer, it is added to the buffer, i.e., $\mathcal{B}_k \leftarrow \mathcal{B}_k \cup \{p_k\}$, and hence the packet counter will be increased, i.e., $p_k \leftarrow p_k + 1$.

In every SR period, the device checks its own buffer and performs the SR in case the buffer is not empty. The binary parameter $\alpha_{k,t}$ indicates if the device k performed a SR at the t-th SR period. The $\alpha_{k,t}$ can be defined as follows:

$$\alpha_{k,t} = \begin{cases} 1, & \text{if } \mathcal{B}_k \neq \emptyset \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

If $\alpha_{k,t} = 1$, i.e., device k sent a SR in the t-th slot, the base station needs to take into account the buffer state of the device. We define $\mathcal{A}_{k,t}$ to be the number of packets of device k taken into account by the base station in the scheduling procedure at the t-th time slot. Clearly, when the SR occurs, $\mathcal{A}_{k,t} \leftarrow \mathcal{B}_k$.

After the transmission of the SR, the device waits for the reception of the UL grant, otherwise the SR is re-transmitted. According to Fig. 4, given t^* the time slot when the device sent the SR, the time slot t when the device is expected to receive the UL grant can be computed as:

$$t = t^* + \left\lceil \frac{T_{TX} + T_{PRO} + T_{TX}}{T_{SR}} \right\rceil. \tag{2}$$

This allows to take into account the SR transmission time, the processing and the transmission time at the base station, where the ceiling function is used to synchronize the time slot t with the T_{SR} window. The binary parameter $\beta_{k,t}$ indicates whether a device k is receiving an UL grant in the t-th time slot. The $\beta_{k,t}$ is defined as follows:

$$\beta_{k,t} = \begin{cases} 1, & \text{if } \sum_{k'=1}^{k} \alpha_{k',t^*} \leq N \wedge \alpha_{k,t^*} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where N is the maximum number of SRs that base station can handle simultaneously. The $\beta_{k,t} = 1$ means that base station has scheduled the device for data transmission and device k will receive the UL grant in the t-th slot. Assuming $\mathcal{A}_{k,t} = \mathcal{A}_{k,t^*}$ is the set of device k's packets for transmission, the buffer will be updated as follows: $\mathcal{B}_k \leftarrow \mathcal{B}_k \setminus \mathcal{A}_{k,t}$. On the other hand $\beta_{k,t} = 0 \wedge \alpha_{k,t^*} = 1$ means the transmitted SR is not granted and device needs to re-schedule a SR procedure at the next SR opportunity.

We assumed that all devices have the same priority and base station does not prioritize the SR of any device. Hence, the SRs received by the base station are handled in a round robin fashion. Similar assumption holds for data scheduling in (5) and (8).

After the reception of the UL grant, the final step is thus the effective data transmission. According to Fig. 4, given t^* the time slot when device receives the UL grant, the first time slot t available to send data can be computed as:

$$t = t^* + \left\lceil \frac{T_{PRO} + T_{TX}}{T_{SR}} \right\rceil. \tag{4}$$

The Equation allows to take into account the processing time at the device side after the reception of the grant plus the time needed to send the data. The amount of resources needed to transmit the data, i.e., r_k, can be computed by considering the amount of data and the SINR experienced by the device: $r_k = f(|\mathcal{A}_{k,t^*}| \cdot s_k, \sigma_k)$. We exploit the binary parameter $\gamma_{k,t}$ to indicate if a device k is transmitting its data in the t-th time slot. The $\gamma_{k,t}$ is defined as follows:

$$\gamma_{k,t} = \begin{cases} 1, & \text{if } \sum_{k'=1}^{k} r_{k'} \cdot \beta_{k',t^*} \leq R_{UL} \wedge \beta_{k,t^*} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

If $\gamma_{k,t} = 1$, the device has been successfully scheduled to transmit its data in the t-th time slot. We save this information through the set $\mathcal{D}_{k,t}$, hence $\mathcal{D}_{k,t} \leftarrow \mathcal{A}_{k,t^*}$ and $\mathcal{A}_{k,t^*} \leftarrow \emptyset$. If not enough resources are available in the t-th time slot, base station will not schedule device k and hence $\gamma_{k,t} = 0 \wedge \beta_{k,t^*} = 1$. In this case, the devices will be scheduled in the next available time slot, hence $\beta_{k,t^*+1} = 1, \mathcal{A}_{k,t^*+1} \leftarrow \mathcal{A}_{k,t^*}$, $\mathcal{D}_{k,t} \leftarrow \emptyset$ and, finally $\mathcal{A}_{k,t^*} \leftarrow \emptyset$.

When the session ends, p_k represents the last packet counter for device k. This means that we can build a set of packets sent by device k as: $\mathcal{P}_k = \{1, 2, \ldots, p_k\}$. For each packet $p \in \mathcal{P}_k$, we can compute the transmission delay as follows:

$$\lambda_{k,p} = \frac{T_{SR}}{2} + (t_{TX} - t_{SR}) \cdot T_{SR} + T_{PRO} \tag{6}$$

where $T_{SR}/2$ takes into consideration the average waiting time before sending the SR, while T_{PRO} represents the processing time at the base station after data reception. In (6), t_{SR} indicates the time slot when the device sent the SR relevant to the packet p and can be computed as $t_{SR} = t | p \in \mathcal{A}_{k,t}$; similarly, t_{TX} indicates the time slot when the device sent the packet p to the base station and can be computed as $t_{TX} = t | p \in \mathcal{D}_{k,t}$.

B. LEGACY DATA RECEPTION

The legacy data reception is triggered when the base station receives a packet to be delivered to a device within its coverage area. We assume that the base station schedules packets in the downlink direction every T_{DL} ms.

The binary parameter $\delta_{k,t}$ indicates if data has been received ($\delta_{k,t} = 1$) or not ($\delta_{k,t} = 0$) by the base station towards device k at the t-th time slot. A packet that is addressed to device k reaching the base station, is represented by p_k. If $\delta_{k,t} = 1$ the base station adds this packet to the buffer of data to be delivered to the device. For the sake of simplicity, we reuse the notation $\mathcal{B}_k$ to denote this buffer. Hence, $\mathcal{B}_k \leftarrow \mathcal{B}_k \cup \{p_k\}$ and then $p_k \leftarrow p_k + 1$. If $\delta_{k,t} = 1$, we exploit the set $\mathcal{A}_{k,t} = \mathcal{B}_k$ in order to compute the final delivery delay for each packet.

At the reception of the packet, the base station needs to schedule the data reception on the radio channel. By denoting with t^* the slot when the base station received the data addressed to device k, the first available time slot t for data delivery can be computed as:

$$t = t^* + \left\lceil \frac{T_{PRO} + T_{Al} + T_{TX}}{T_{DL}} \right\rceil \tag{7}$$

where t takes into account the processing time at the base station, the time needed for Transmit Time Interval (TTI) alignment (here denoted with T_{Al}) plus the time spent by the base station to send the packet(s). The amount of resources needed to transmit the data, i.e., r_k, can be computed by considering the amount of data and the SINR experienced by the device: $r_k = f(|\mathcal{A}_{k,t^*}| \cdot s_k, \sigma_k)$. Finally, we exploit the binary parameter $\eta_{k,t}$ to indicate if a device k is receiving its data in the t-th time slot; $\eta_{k,t}$ is defined as follows:

$$\eta_{k,t} = \begin{cases} 1, & \text{if } \sum_{k'=1}^{k} r_{k'} \cdot \delta_{k',t^*} \leq R_{DL} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

If $\eta_{k,t} = 1$, the device has been successfully scheduled to receive its data in the t-th time slot; we save this information through the set $\mathcal{D}_{k,t}$. In detail, we set $\mathcal{D}_{k,t} \leftarrow \mathcal{A}_{k,t^*}$ and then we set $\mathcal{A}_{k,t^*} \leftarrow \emptyset$. In case the base station did not schedule device k for data reception in the t-th time slot, i.e., $\eta_{k,t} = 0 \wedge \delta_{k,t^*} = 1$, this means that not enough resources are available in this time slot and as a consequence the devices will be scheduled in the next available one. This means that $\delta_{k,t^*+1} \leftarrow 1, \mathcal{A}_{k,t^*+1} \leftarrow \mathcal{A}_{k,t^*}$, and, finally $\mathcal{A}_{k,t^*} \leftarrow \emptyset$.

When the session ends, p_k represents the last packet counter for device k. We can thus denote with $\mathcal{P}_k = \{1, 2, \ldots, p_k\}$ the set of packets received by device k. For each packet $p \in \mathcal{P}_k$, we can compute the reception delay as follows:

$$\lambda_{k,p} = (t_{RX} - t_{BS}) \cdot T_{PRO} \tag{9}$$

where T_{PRO} represents the processing time at the device side after data reception. In (9), t_{BS} indicates the time slot when the base station receives the packet p to be delivered to device k and can be computed as $t_{BS} = t | p \in \mathcal{A}_{k,t}$. The parameter t_{RX} indicates the time slot when the device receives the packet p from the base station and can be computed as $t_{RX} = t | p \in \mathcal{D}_{k,t}$.

C. ENHANCED DATA TRANSMISSION

The main novelty of our proposal is the introduction of a 3GPP-compliant *soft resource reservation* in the SR procedure. As depicted in Fig. 6, the soft resource reservation is composed of two steps. In the first step the device performs the legacy SR procedure for the transmission of the first haptic packet. Therefore, the base station becomes aware of the packet size relevant to the haptic session. When the base station assigns the UL grant to the device, this grant is *soft reserved* for the device. This means that the device will use this grant for the following transmissions and hence the second step of our proposed procedure is shaped. It is worth mentioning that this does not require the introduction of any new message from the device to the BS, fields of existing messages just need to be modified to inform the BS that this a soft resource reservation SR. Information such as requested amount of traffic is already present in the SR procedure. Our proposal goes in the direction of current advances in the design of 5G networks, where flexibility is one of the key targets to achieve. From this point of view, our proposal allows flexibility in the SR procedure by allowing a device to ask for either a legacy SR or soft resource reservation SR without requiring for a novel procedure from scratch.

In the second step, and when the device wants to transmit extra packets, it will send a SR to the base station in order to inform the base station about an incoming data transmission. However, device already knows the UL grant that is

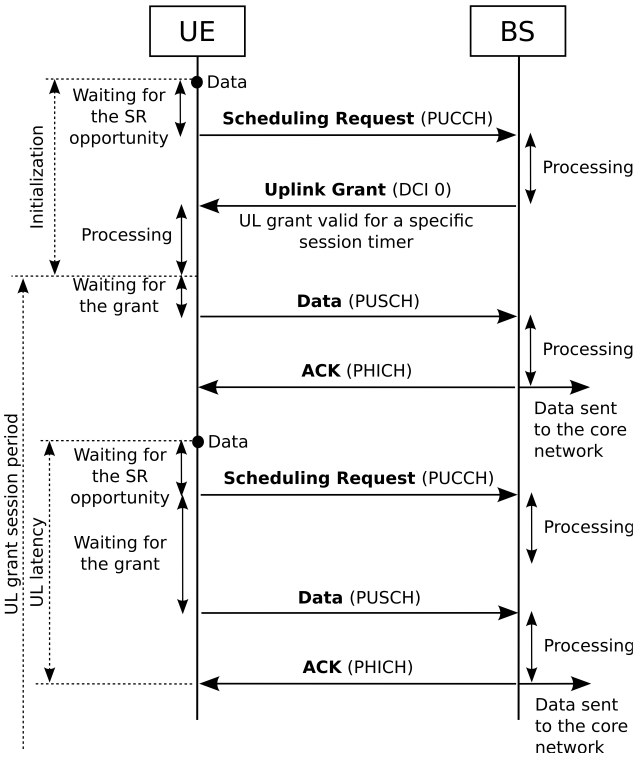

FIGURE 6. Our proposed enhanced scheduling request procedure in LTE.

softly reserved for its transmission and hence can directly transmit its data in the UL grant assigned during the previous round. Considering the fact that, potentially, the device could generate packets every 1ms, there might be multiple packets for transmission at the device. Therefore, given the SR periodicity T_{SR}, the reserved resources are allocated in order to accommodate T_{SR} packets. For example, in the case of one transmission every 1ms, the device will send T_{SR} packets every T_{SR} ms. Hence, the soft UL grant will be composed of $r_k = f(|T_{SR}| \cdot s_k, \sigma_k)$ RBs.

The term "*soft*" in this procedure aims to underline the main difference with respect to the semi-persistent scheduling. In case the base station does not receive any SR from the device, it is aware that the device does not have any data to send and can thus use the resources reserved for this device for other communications within the cell. This means that, from a spectral efficiency point of view, our proposed approach does not introduce any drawbacks compared to the legacy SR, as resources not used by a device involved in a teleoperation session will be assigned by the BS to other devices in the cell. The UL grant configuration step can be repeated in case of changes such as the exploitation of a different control scheme or changes in the channel quality of the device.

We now present the model of our enhanced uplink transmission procedure, by focusing on the period when the soft UL grant is active. The buffer of each device, i.e., $\mathcal{B}_k$, contains the list of packets to be transmitted. Similar to the above, we exploit a packet counter denoted with p_k. When a packet generated from the application layer reaches the buffer of data transmission, we consider that the novel packet is added to the buffer, i.e., $\mathcal{B}_k \leftarrow \mathcal{B}_k \cup \{p_k\}$, and then the counter of the next packet to be sent is increased, i.e., $p_k \leftarrow p_k + 1$. Every SR period, the device checks its own buffer and sends a SR in case it is not empty: in this case, the device basically informs the base station about the number of packets to be sent, in order to make the base station aware of how many resources of the soft UL grant will be used. We exploit the binary parameter $\alpha_{k,t}$ to indicate if the device k performed a SR at the t-th SR period, and $\alpha_{k,t}$ is defined as in (1). If $\alpha_{k,t} = 1$, i.e., the device informed the base station that it is going to transmit $\mathcal{A}_{k,t} = \mathcal{B}_k$ packets in its soft UL grant.

After the transmission of the SR, the device does not need to wait for the UL grant reception, as it already knows the resources allocated to it. By denoting with t^* the time slot when device sent the SR, $t = t^* + 1$ represents the instant when the device will be able to send its data by exploiting the resources of the soft UL grant. This means that $\mathcal{D}_{k,t} \leftarrow \mathcal{A}_{k,t^*}$.

When the session ends, p_k represents the last packet counter for device k. This means that we can build a set of packets sent by device k as: $\mathcal{P}_k = \{1, 2, \ldots, p_k\}$. For each packet $p \in \mathcal{P}_k$, we can compute the transmission delay as for the legacy procedure, i.e., as in (6), the only difference is that $t_{TX} - t_{SR}$ now has a shorter value compared to the legacy procedure. This is because the device, which has a soft reservation of resources, will be transmitting its data quicker and thus t_{TX} is smaller compared to the legacy procedure.

V. PERFORMANCE EVALUATION

In the considered scenario, we assume all devices support QPSK. This means that a 24B data packet generated by a haptic device needs only one LTE RB to be transmitted as QPSK allows to transmit up to 35B with one RB [10]. By considering (5), this means that $r_k = 1 \forall k$.

For the sake of simplicity, we consider only haptic devices in our evaluation. Nevertheless, it is worth mentioning that, in case of presence of other human devices, access control mechanisms can be applied to guarantee that traffic handled by the base station does not exceed the maximum supported one while prioritization can be applied to guarantee high-priority scheduling for haptic devices. Furthermore, it is worth noticing that our proposed strategy intrinsically supports haptic traffic prioritization by means of soft reserving resources for haptic data transmission. This means that we expect to obtain results similar to those shown in the remainder of this Section also in the presence of background human traffic.

We also assume a delay of 1 ms in the core network [19]. Configuration of main network parameters of interest for our evaluation is reported in Table 1.

A. ONE-WAY COMMUNICATION BETWEEN MASTER AND SLAVE DEVICES

Fig. 7(a) and 7(b) analyze the latencies in the UL and DL directions and for both master and slave sides by considering

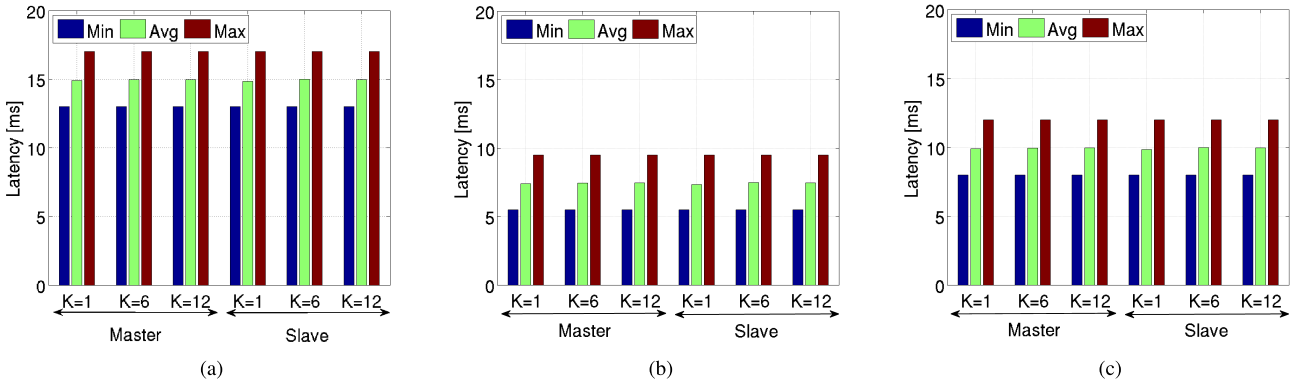

FIGURE 7. Latency in the legacy procedures and in our proposed soft resource reservation. (a) Latency in uplink direction for legacy procedure. (b) Latency in downlink direction for legacy procedure. (c) Latency in uplink direction for our proposed procedure.

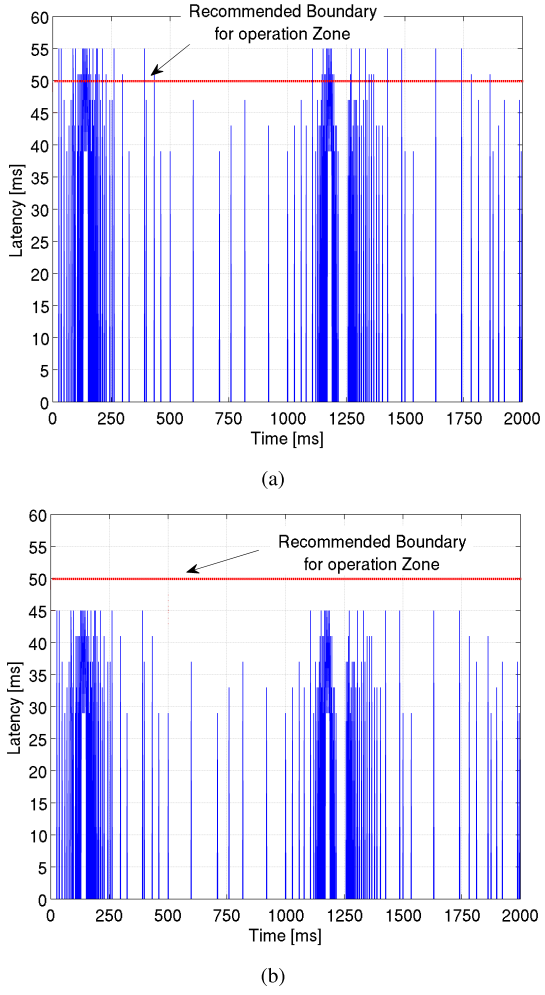

FIGURE 8. Round trip time for the case of one subject. (a) Legacy Procedure. (b) Proposed soft reservation strategy.

the legacy procedures. We study three cases, i.e., when the session is composed of 1, 6, and 12 master-slave pairs, and several interesting conclusions can be drawn from this study. It can be seen that the increase in the number of devices does not affect the performance, due to the fact that haptic traffic does not represent a large source of traffic load for the cell. Observing from Fig. 7(a) and 7(b), the UL latency is between 13 ms and 17 ms, while the DL one varies from 5.5 ms to 9.5 ms.

Fig. 7(c) shows the UL performance achieved using our proposed soft reservation procedure. It can be seen that the latency varies from 8 ms to 12 ms, i.e., a reduction ranging from 30% to 40% w.r.t. the legacy procedure. The improvement is obtained by considering that the device does not need to wait for the UL grant, while it sends data directly after the transmission of the SR to effectively reserve the already soft reserved RBs.

B. ROUND-TRIP COMMUNICATION PATH, SINGLE HAPTIC DEVICE

After the analysis of the single directions of the latency, we now focus our attention on the Round Trip Time (RTT), since this is a delay observed by the human operator of the teleoperation system, i.e. master-slave-master path.

Fig. 8 shows the RTT obtained with the legacy transmission procedures as well as with our proposed UL strategy, in the case of one subject. Observing from this figure, the legacy procedure experiences RTT increases by up to 55 ms during the bursty periods. Furthermore, Fig. 8 shows the RTT obtained by exploiting our proposed UL transmission strategy, where the RTT delay is reduced to a maximum value of approximately 45 ms.

It is worthwhile mentioning that subjective studies [9] in teleoperation system show 50 ms as a threshold above which the QoE of the human operator will be significantly affected. Given the assumption of our simulation model (most important one being 1 ms latency of the core network), the proposed UL soft reservation will bring the RTT below this threshold.

C. ROUND-TRIP COMMUNICATION PATH, MULTIPLE HAPTIC DEVICES

Fig. 9 shows the Empirical Cumulative Distribution Function (ECDF) of the RTT for 12 teleoperation traffic

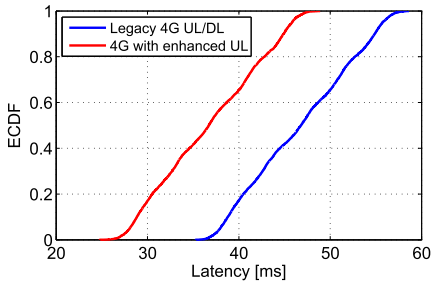

FIGURE 9. Empirical cumulative distribution function (ECDF) of the round trip (master-slave-master) time delay for the case of 12 subjects.

TABLE 2. Analysis of UL latency with shorter TTI.

Step	Legacy TTI 1ms	Legacy TTI 0.2ms	Proposed TTI 1ms	Proposed TTI 0.2ms
Waiting for SR	2.5ms	0.1ms	2.5ms	0.1ms
SR transmission	1ms	0.2ms	1ms	0.2ms
BS Processing	3ms	3ms	-	-
UL grant	1ms	0.2ms	-	-
UE Processing	3ms	3ms	2.5ms	1ms
Data Tx	1ms	0.2ms	1ms	0.2ms
BS Processing	3ms	3ms	3ms	3ms
Total	14.5ms	9.7ms	10ms	4.5ms

TABLE 3. Analysis of DL latency with shorter TTI.

Step	TTI 1ms	TTI 0.2ms
Processing	3ms	3ms
TTI alignment	0.5ms	0.1ms
DL data transmission	1ms	0.2ms
Data encoding	3ms	3ms
Total	7.5ms	6.3ms

flows (these correspond to the 12 participants in the study in [9]). This plot shows how effectively our proposed strategy maintains a round trip delay below 50 ms (the QoE degradation threshold) even in the case of multiple haptic sessions active in the cell. In the legacy procedure, however, more than 35% of the packets experience RTT of higher than 50 ms, translated to low QoE by human operator (the master side).

D. ANALYSIS OF 5G AND SHORTER TTI

Another analysis we present takes into consideration the exploitation of shorter Transmit Time Interval (TTI), envisioned in 5G systems to be reduced to support low latency services. Hence, we consider a TTI duration of 0.2 ms as proposed in [19] and we compared the results with the performance in current deployed systems where the TTI is 1 ms. Tables 2 and 3 analyze the average UL and DL latency consecutively by elaborating different sources of latency. Latency is calculated as in [19], where for each step the latency is given as a function of the TTI duration. For shorter TTIs, we used the same evaluation with an updated TTI value. Observing from these tables, the legacy procedure can achieve a latency of approximately 9.7 ms and 6.3 ms in UL and DL directions, respectively. Our proposed procedure can, however, reduce the UL latency further down to 4.5 ms.

VI. RESEARCH BACKGROUND

The interest towards wireless communications to interconnect sensors and actuators has increased in the last decade as wireless technologies are able to cut deployment costs and time-to-market for IoT applications [23], [24]. Nevertheless, the intrinsically non-deterministic nature of wireless links due to the transmission over a shared medium affected the exploitation of wireless communications for applications with strict QoS requirements. In particular, when coming to haptic teleoperation applications, aspects such as low and stable latency become of primary importance to provide acceptable QoE.

In order to meet requirements of the application delivered by wireless networks, research community has focused on cross-layer approaches [25]. Cross-layer protocols allow parameters of two or more layers to be recalled/altered to achieve some specific targets such as latency minimization or reliability maximization. An example can be found in [26], where scheduling, routing and sampling rates are dynamically optimized to guarantee the stability of control systems. Both [25] and [26] are studied within Wireless Sensor Networks (WSNs). However, haptic teleoperation sessions require low and stable end-to-end latency, and cross-layer protocols in WSNs fail in achieving these goals mainly because of a limited control over the full delivery of traffic, i.e. there is no control over the traffic when it leaves the WSN.

Given the property of mobile networks (i.e. traditional cellular networks) in providing full control over the end-to-end path, there is a growing attention to mobile networks as a possible solution for providing guaranteed end-to-end latency. Within mobile networks, two bodies of research shape the baseline for this paper, including works on scheduling request and on resource allocation procedures. In the first category, as discussed in Sec. III, past works focused on proposing novel strategies to be used instead of the legacy scheduling request (SR) procedure. The scheduling request procedures follows two main approaches: semi-persistent scheduling [12], [13] and contention-based scheduling [14]. The former can guarantee low latency at the expense of spectral efficiency and resource utilization while the latter has limitations in terms of collisions if multiple devices access the same resource as well as flexibility due to the use of fixed MCS and packet sizes. In the second category, application-aware resource allocation is presented in [27] and [28], where traffic is prioritized during resource assignments depending on the application. Although both [27] and [28] are effective solutions to cut delays due to resource allocation in scenarios with constrained resources, both works assume the exploitation of the legacy SR procedure that causes extra delay, as discussed throughout the paper.

VII. CONCLUSION

In this paper, we proposed a soft reservation strategy for the UL scheduling of LTE-based networks aiming at providing ultra-low-delay services to various teleoperation scenarios. The development of the proposed strategy highly depends

on the characterization of haptic traffic (that depends on the control scheme, and the coding used). The simulation results illustrated the efficiency of the proposed soft reservation strategy which reduces the round-trip delay by an average of 10 ms compared with the legacy solution. Because of the delay-sensitive nature of teleoperation systems, this achievement will bring admirable QoE improvements to teleoperation under different application scenarios. These results can be considered as a valuable guidance to control stability mechanisms at the teleoperation devices to allow for appropriate selection of control schemes under different environment dynamics and communication delays.

REFERENCES

[1] F. Xia, L. T. Yang, L. Wang, and A. Vinel, "Internet of things," *Int. J. Commun. Syst.*, vol. 25, no. 9, p. 1101, 2012.
[2] G. P. Fettweis, "The tactile Internet: Applications and challenges," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 64–70, Mar. 2014.
[3] M. Dohler *et al.*, "Internet of Skills, where robotics meets AI, 5G and the tactile Internet," in *Proc. EUCNC*, 2017, p. 1.
[4] W. R. Ferrell and T. B. Sheridan, "Supervisory control of remote manipulation," *IEEE Spectr.*, vol. 4, no. 10, pp. 81–88, Oct. 1967.
[5] D. A. Lawrence, "Stability and transparency in bilateral teleoperation," *IEEE Trans. Robot. Autom.*, vol. 9, no. 5, pp. 624–637, Oct. 1993.
[6] X. Xu, C. Schuwerk, B. Cizmeci, and E. Steinbach, "Energy prediction for teleoperation systems that combine the time domain passivity approach with perceptual deadband-based haptic data reduction," *IEEE Trans. Haptics*, vol. 9, no. 4, pp. 560–573, Oct. 2016.
[7] X. Xu, B. Cizmeci, A. Al-Nuaimi, and E. Steinbach, "Point cloud-based model-mediated teleoperation with dynamic and perception-based model updating," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 11, pp. 2558–2569, May 2014.
[8] M. Condoluci, G. Araniti, T. Mahmoodi, and M. Dohler, "Enabling the IOT machine age with 5G: Machine-type multicast services for innovative real-time applications," *IEEE Access*, vol. 4, pp. 5555–5569, May 2016.
[9] X. Xu, Q. Liu, and E. Steinbach. (May 2017). "Toward QoE-driven dynamic control scheme switching for time-delayed teleoperation systems: A dedicated case study." [Online]. Available: https://arxiv.org/abs/1705.05613
[10] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures*, document TR 36.213, 3GPP, 2009.
[11] *Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) Protocol Specification*, document TR 36.321, 3GPP, 2010.
[12] H. Jin, C. Cho, N. O. Song, and D. K. Sung, "Optimal rate selection for persistent scheduling with HARQ in time-correlated Nakagami-m fading channels," *IEEE Trans. Wireless Commun.*, vol. 10, no. 2, pp. 637–647, Feb. 2011.
[13] J. B. Seo and V. C. M. Leung, "Performance modeling and stability of semi-persistent scheduling with initial random access in LTE," *IEEE Trans. Wireless Commun.*, vol. 11, no. 12, pp. 4446–4456, Dec. 2012.
[14] S. Andreev *et al.*, "Efficient small data access for machine-type communications in LTE," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2013, pp. 3569–3574.
[15] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4–16, Feb. 2014.
[16] R. Kwan and J. C. Ikuno, "Effective HARQ code rate modelling for LTE," *Electron. Lett.*, vol. 49, no. 7, pp. 462–464, Mar. 2013.
[17] M. R. Raghavendra, S. Nagaraj, K. V. Pradap, and P. Fleming, "Robust channel estimation and detection for uplink control channel in 3GPP-LTE," in *Proc. Global Telecommun. Conf. (GLOBECOM)*, Nov. 2009, pp. 1–5.
[18] M. Centenaro and L. Vangelista, "HARQ in LTE uplink: A simple and effective modification suitable for low mobility users," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 3763–3769.
[19] *Study on Latency Reduction Techniques for LTE*, document TR 36.881, 3GPP, 2010.
[20] D. Jiang, H. Wang, E. Malkamaki, and E. Tuomaala, "Principle and performance of semi-persistent scheduling for VoIP in LTE system," in *Proc. Int. Conf. Wireless Commun., Netw. Mobile Comput.*, Sep. 2007, pp. 2861–2864.
[21] G. Mountaser, M. Lema, T. Mahmoodi, and M. Dohler, "On the feasibility of MAC and PHY split in Cloud RAN," in *Proc. IEEE WCNC*, Mar. 2017, p. 1.
[22] P. Vizarreta, M. Condoluci, C. M. Mahuca, T. Mahmoodi, and W. Kellerer, "QoS-driven function placement reducing expenditures in NFV deployments," in *Proc. IEEE ICC*, May 2017, pp. 1–5.
[23] M. Z. Hasan, H. Al-Rizzo, and F. Al-Turjman, "A survey on multipath routing protocols for QoS assurances in real-time wireless multimedia sensor networks," *IEEE Commun. Surveys Tuts.*, to be published.
[24] A. A. K. S, K. Ovsthus, and L. M. Kristensen, "An industrial perspective on wireless sensor networks—A survey of requirements, protocols, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1391–1412, 3rd Quart., 2014.
[25] I. Al-Anbagi, M. Erol-Kantarci, and H. T. Mouftah, "A survey on cross-layer quality-of-service approaches in WSNs for delay and reliability-aware applications," *IEEE Commun. Surv. Tuts.*, vol. 18, no. 1, pp. 525–552, 1st Quart., 2016.
[26] P. Park, P. D. Marco, and K. H. Johansson, "Cross-layer optimization for industrial control applications using wireless sensor and actuator mesh networks," *IEEE Trans. Ind. Electron.*, vol. 64, no. 4, pp. 3250–3259, Apr. 2017.
[27] H. Shajaiah, A. Abdelhadi, and T. C. Clancy, "Towards an application-aware resource scheduling with carrier aggregation in cellular systems," *IEEE Commun. Lett.*, vol. 20, no. 1, pp. 129–132, Jan. 2016.
[28] X. Wang, M. J. Sheng, Y. Y. Lou, Y. Y. Shih, and M. Chiang, "Internet of Things session management over LTE: Balancing signal load, power, and delay," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 339–353, Jun. 2016.

MASSIMO CONDOLUCI received the B.Sc. and M.Sc. degrees in telecommunications engineering and the Ph.D. degree in information technology from the Mediterranea University of Reggio Calabria, Italy, in 2008, 2011, and 2016, respectively. He is currently a Research Associate with the Centre for Telecommunications Research, King's College London, U.K. His current research interests include mobile and fixed network convergence, flexible functionality split, machine-type communications, and multicasting over 5G wireless networks.

TOKTAM MAHMOODI received the B.Sc. degree in electrical engineering from the Sharif University of Technology, Iran, and the Ph.D. degree in telecommunications from the King's College London, U.K. She was involved in the European FP7 and EPSRC Projects aiming to push the boundaries of the next generation mobile communications forward. She was also involved in the mobile and personal communications industry from 2002 to 2006, and in the Research and Development Team on developing DECT standard for WLL applications. She was a Mobile VCE Researcher from 2006 to 2009, a Post-Doctoral Research Associate with the ISN Research Group, Electrical and Electronic Engineering Department, Imperial College, from 2010 to 2011, and a Visiting Research Scientist with F5 Networks, San Jose, CA, in 2013. She is currently Lecturer in telecommunications with the Department of Informatics, King's College London.

ECKEHARD STEINBACH (F'15) received the Dipl.Ing. degree from the Karlsruhe Institute of Technology, Karlsruhe, Germany, in 1994, and the Dr.-Ing. degree from the University of Erlangen-Nuremberg, Erlangen, Germany, in 1999. From 1994 to 2000, he was a Research Staff Member with the Image Communication Group, University of Erlangen-Nuremberg. From 2000 to 2001, he was a Post-Doctoral Fellow with the Information Systems Laboratory, Stanford University, Stanford, CA, USA. In 2002, he joined the Department of Electrical Engineering and Information Technology, Technical University of Munich, Munich, Germany, where he is currently a Full Professor of Media Technology. His current research interests include audio-visual-haptic information processing and communication and networked and interactive multimedia systems.

MISCHA DOHLER (F'14) is currently a Full Professor of Wireless Communications with the King's College London, the Head of the Centre for Telecommunications Research, and the Co-Founder and a member of the Board of Directors of the smart city pioneer Worldsensing. He is a Frequent Keynote, a Panel, and a Tutorial Speaker. He has pioneered several research fields, contributed to numerous wireless broadband and IoT/M2M standards, holds a dozen patents, organized and chaired numerous conferences, has over 200 publications, and authored several books. He acts as a Policy, Technology, and Entrepreneurship Adviser, examples being Richard Branson's Carbon War Room, the House of Lords U.K., the EPSRC ICT Strategy Advisory Team, the European Commission, the ISO Smart City Working Group, and various start-ups. He is also an Entrepreneur, an Angel Investor, a Passionate Pianist, and fluent in six languages. He has talked at TEDx. He had coverage by national and international TV and radio. His contributions have featured on BBC News and the *Wall Street Journal*. He is a Distinguished Lecturer of the IEEE.