Towards Massive Connectivity Support for Scalable mMTC Communications in 5G networks

Bockelmann, Carsten ; Kiilerich Pratas, Nuno; Wunder, Gerhard; Saur, Stephan; Navarro, Monica; Gregoratti, David; Vivier, Guillaume; De Carvalho, Elisabeth; Ji, Yalei ; Stefanovic, Cedomir; Popovski, Petar; Qi, Wang; Schellmann, Malte; Kosmatos, Evangelos; Demestichas, Panagiotis; Raceala-Motoc, Miruna; Jung, Peter; Stanczak, Slawomir; Dekorsy, Armin

Published in:
IEEE Access

DOI (link to publication from Publisher):
10.1109/ACCESS.2018.2837382

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 5, 2018, accepted April 16, 2018, date of publication May 16, 2018, date of current version June 20, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2837382

Towards Massive Connectivity Support for Scalable mMTC Communications in 5G Networks

CARSTEN BOCKELMANN[1], (Member, IEEE), NUNO K. PRATAS[2], GERHARD WUNDER[3,11], (Senior Member, IEEE), STEPHAN SAUR[4], (Member, IEEE), MÒNICA NAVARRO[5], (Senior Member, IEEE), DAVID GREGORATTI[5], (Senior Member, IEEE), GUILLAUME VIVIER[6], ELISABETH DE CARVALHO[2], YALEI JI[1], ČEDOMIR STEFANOVIĆ[2], (Senior Member, IEEE), PETAR POPOVSKI[2], (Fellow, IEEE), QI WANG[7], MALTE SCHELLMANN[7], EVANGELOS KOSMATOS[8], (Member, IEEE), PANAGIOTIS DEMESTICHAS[9], (Senior Member, IEEE), MIRUNA RACEALA-MOTOC[3], PETER JUNG[10], (Member, IEEE), SLAWOMIR STANCZAK[3], AND ARMIN DEKORSY[1], (Senior Member, IEEE)

[1]University of Bremen, 28359 Bremen, Germany
[2]Aalborg University, 9220 Aalborg, Denmark
[3]Fraunhofer Heinrich-Hertz-Institut Berlin, 10587 Berlin, Germany
[4]Nokia Bell Labs, 70435 Stuttgart, Germany
[5]Centre Tecnològic de Telecomunicacions de Catalunya, 08860 Castelldefels, Spain
[6]Sequans Communications, 92700 Colombes, France
[7]Huawei Munich Research Center, 80992 Munich, Germany
[8]WINGS ICT Solutions, 176 73 Athens, Greece
[9]University of Pireaus, 185 34 Piraeus, Greece
[10]Technical University of Berlin, 10587 Berlin, Germany
[11]FU Berlin, Heisenberg CIT Group, 14195 Berlin, Germany

Corresponding author: Carsten Bockelmann (bockelmann@ant.uni-bremen.de)

This work was supported in part by the Horizon 2020 Project FANTASTIC-5G under Grant ICT-671660 and in part by the European Union.

ABSTRACT The fifth generation of cellular communication systems is foreseen to enable a multitude of new applications and use cases with very different requirements. A new 5G multi-service air interface needs to enhance broadband performance as well as provide new levels of reliability, latency, and supported number of users. In this paper, we focus on the massive Machine Type Communications (mMTC) service within a multi-service air interface. Specifically, we present an overview of different physical and medium access techniques to address the problem of a massive number of access attempts in mMTC and discuss the protocol performance of these solutions in a common evaluation framework.

INDEX TERMS 5G, mMTC, massive access, massive connectivity, random access.

I. INTRODUCTION

The prospect of billions of interconnected devices within the paradigm of the Internet of Things (IoT) has become one of the main drivers of the research and development in the ICT sector. In fact, the 5G requirements for IMT-2020 include the support of a multiplicity of services and applications, with massive Machine Type Communications (mMTC) being one of the three cores services. The other core services being the Ultra Reliable Low Latency (URLLC) and the extreme Mobile Broadband (eMBB) communications.

The focus in this paper is on the massive access protocols and multi-user decoding techniques associated with the support of the mMTC core service. The objective is to attach a large number of low-rate low-power devices, termed Machine-Type Devices (MTDs), to the cellular network. There are multiple factors that demand increased number of connected MTDs: the smart-grid, large scale environment and structure monitoring, asset and health monitoring, etc. Typically, these MTDs connect asynchronously and sporadically to the network to transmit small data payloads. Connected objects include various types with an extremely wide set of requirements: for instance, a connected goggle providing augmented reality would require lower latency and higher throughput compared to a connected smoke detector. However, it is commonly understood that mMTC indicates the family of devices requiring sporadic access to the network to transmit small data payloads. The sporadic access leads to having an unknown, random subset of devices being active at

a given transmission instant or frame, which necessitates the use of some form of random access protocol.

Most of the existing MTC connections, not necessarily massive, are wireless and take place via open standard short range technologies that operate in unlicensed spectrum, such as IEEE 802.15.x and 802.11. Another trend is seen in the proprietary technologies for wide-area IoT, such as SIGFOX [1] and LoRA [2], addressing the physical domain not covered by short-range technologies and thus providing a clear indication of an emerging market that is yet to be filled by service providers. Until recently, the cellular standards could only provide access to MTDs via SMS or GPRS. This approach suffers from coverage limitations (e.g. in deep indoor for instance for gas or water meters), non-optimized hardware and limited subscription models. Moreover, the 2G/3G systems were not designed to handle thousands of sporadically active MTDs. As a result, the 3GPP has extended the support of LTE to MTC with the standardization of cat-M and NB-IoT in 2015-2016. Those standards meet most of the mMTC requirements, but still need to be improved to support the massive number of terminals with low capabilities, sporadic activity patterns, and short packet transmissions.

One of the major obstacles for the proliferation of efficient cellular access for mMTC stems from the deficiencies of the access reservation procedure, a key building block of the cellular access networking. Currently, the access reservation procedure is designed to enable connection establishment from a relatively low number of accessing devices. Additionally, each device has moderate to high data-rate requirements such that the overhead of current access protocols with multiple phases is relatively small. Both assumptions, the low number of devices as well as moderate to high data rates are in contradiction to mMTC needs. Thus, enhancement of the access reservation procedure for mMTC traffic has been in the focus of both the research community [3] and standardization [4]. However, there is a common understanding that the mMTC traffic requirements call for a more radical redesign of the cellular access [5].

Indeed, 3GPP has recently concentrated its standardization efforts in this regard in four parallel tracks, which are (i) LTE for M2M (eMTC) focusing on the modification of LTE radio access network (RAN) for mMTC services and targeted at devices with reduced air-interface capabilities [6], (ii) narrow-band IoT (NB-IoT) which targets low-cost narrow-band devices with reduced functionalities [7], (iii) extended coverage GSM for IoT (EC-GSM-IoT) [8] and (iv) the support of mMTC in 5G. In the efforts (i)-(iii), the goal can be summarized as [9]: improved indoor coverage (15-20 dB when compared to current cellular systems) and outdoor coverage up to 15 km, support of massive number of low data-rate devices with modest device complexity, improved power efficiency to ensure longer battery life, reduced access latency and efficient co-existence with the legacy cellular systems. In (iv), the development towards 5G has started in 3GPP; and while the first phase, to be standardized in Release 15 [10], focuses on extreme MBB (eMBB) services, the URLLC and mMTC will be in the focus of the following phases.

In this paper we summarize several approaches to address the massive access problem for mMTC in 5G and present an evaluation framework to assess the performance of the presented approaches in terms of the access protocol performance. The presented solutions are part of the main innovations and outcomes of the FANTASTIC-5G project [11].[1] First, we will outline the overall mMTC challenges and the specific research questions to be addressed in section II (also see the overview paper [10]) and provide a short overview of the state of the art MTC systems in section III. Then our system level and evaluation approach will be outlined in section IV and detailed technical approaches and their achieved performance for pure MAC protocols will be discussed in section V and combined PHY& MAC approaches in section VI. Finally, we will present the results and compare different solutions in terms of their requirements and advantages. The paper wraps with conclusions in section VII.

II. mMTC CHALLENGES

Many MTC applications are already served by today's communication systems. However, the characteristic properties of mMTC, i.e. the massive number of devices and the very short payload sizes, require novel approaches and concepts. 5G offers the opportunity to tackle the critical challenges in a seamless cellular system combining mMTC and all the other services. In the following we mostly focus on the challenges on the PHY and MAC layers but shortly discuss system level considerations, too.

A. PHY/MAC CHALLENGES

In summary MTC leads to the following mMTC challenges:

- **Control signaling challenge**: In the existing LTE specification an endless cascade of signaling exchange between MTD, eNodeB and core network is initiated if an MTD is in idle mode and intends to send one single small packet. The overall number of sent bits is dominated by control information, and the actual data becomes negligible. Therefore, a 5G system must provide low overhead data transmission modes through novel MAC and PHY design. Additionally, higher layer enhancements, such as radio resource control signaling, are urgently required to lower the overhead on the reconnecting and re-authentication of idle users. Finally, methods that enable the transmission of small data packets over the control plane should be considered.
- **Access capacity challenge**: In LTE the first step to accessing the system or reconnecting when the device in idle mode, is the access reservation protocol. The throughput of the LTE access reservation protocol is

[1] FANTASTIC-5G is the phase 1 project of Horizon 2020 in the framework of 5G PPP dealing with the air interface below 6 GHz with time-line completing on the July 2017.

severely degraded since there is no specific collision resolution procedure in physical (PHY) and medium access (MAC) layer. A 5G system must at least enhance the access reservation protocol through novel MAC and PHY approaches to support a massive number of devices.

- **Power consumption**: The MTDs are often battery powered and require 10+ years of autonomy. For that purpose, the access and communication schemes should be power efficient. This challenge is also related to the type of connection: either always UL triggered (mobile originated) or DL traffic (network originated communications) are considered. For instance, in a Sigfox model, communication is always triggered by an UL request, which helps in terms of power consumption (no need to wake up for paging channels).
- **Multi-Service Integration**: LTE is mostly focused on MBB services and uses a single frame definition and common control channels for these services. In order to enable coexistence of services with very different requirements 5G needs to include flexible frame definitions, a robust waveform and flexible control channel design to allow for dynamic bandwidth sharing and different PHY/MAC approaches. An example is provided in Fig. 1 on the multi-service integration over frequency, time and space resources. In this work, we focus on the mMTC service mainly considering the first three challenges. However, many of the approaches presented in section IV-B can be parameterized for mixed service cases, e.g. to provide higher reliability for lower number of users in a mixed URLLC and mMTC case.

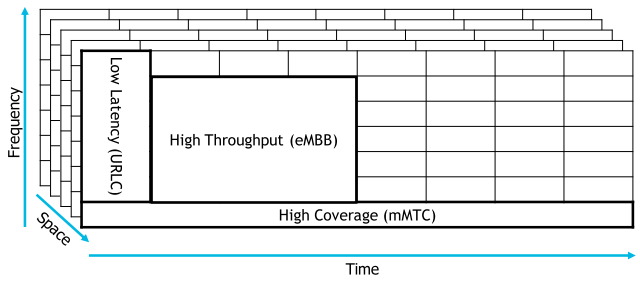

FIGURE 1. Multi service integration.

We focus on the MAC and PHY layer enhancements required to solve the outlined challenges. On the one hand, access protocols with novel waveforms are considered to enable spectral and temporal asynchronicities with very low control overhead; on the other hand, several MAC and PHY approaches and their combination are presented to specifically address the first two challenges ''Control Signaling'' and ''Throughput''. The focus of this paper is on summarizing potential solutions and providing insight on the access protocol throughput and latency of these solutions as will be discussed in section IV. Novel waveforms are only exploited as enabling technology for a novel access protocol here, an exhaustive treatment of the waveforms being considered in a 5G setting is provided in[12] and [13].

B. SYSTEM LEVEL CONSIDERATIONS

When the scenarios under examination are extended to topologies with many cells and in order to take higher layer functionalities into consideration then several system level considerations emerge. In addition, system level scenarios may include cooperative functions between two or more cells (e.g. by using the X2 interface) such as coordinated power control and mobility. In such environments, one can summarize the main system level topics of mMTC as: a) inter-cell interference from devices connected at neighboring cells; b) power control considerations; c) frame structure considerations and; d) intra-cell interference caused by asynchronous transmissions.

In scenarios in which several cells exists, interference emerges both among intra- and inter-cell devices. Regarding intra-cell interference, it emerge in cases of contentions, thus regarding the mMTC access protocols, interference emerges in the access notification stage of multi-stage and two-stage access protocols or during the combined access and data phase of one-stage protocols. On the contrary, inter-cell interference may emerge in any phase of the system including access, connection establishment and data phases, regardless of the selection of the access protocol.

In mMTC scenarios with a single cell, power control mechanisms are targeting to minimize the interference between devices (intra-cell interference) and in increasing the power efficiency to ensure longer battery life. In a multiple cell scenario, power control mechanisms are also targeting to minimize the inter-cell interference in addition to the above. In this direction, several coordinated power control mechanisms exist which study the trade-off between the effectiveness (preciseness of power control) and the overall overhead. Among the innovation of FANTASTIC-5G is the proposition of flexible frame definitions appropriate for a multi-service environment (Fig. 1). In this direction, in contrast to eMBB services which are supported by numerologies with typical LTE TTIs (e.g 1 ms) and URLLC service with strict latency requirements supported by small TTIs (e.g. 0.25 ms), the special requirements of mMTC services can be satisfied by numerologies with long TTIs and short subcarrier spacing in order to increase coverage and decrease device complexity and power consumption.

Regarding the loose uplink synchronization, one main limitation of the mMTC devices with sporadic uplink data is that they use the downlink channel for synchronization. This is not a major problem in small to medium sized cell environments (e.g. with inter-site distance 500 m) and in cases of channel realizations with low delay spread values (e.g. EPA [14]), because in these cases the use of cyclic prefix (CP) compensates for any deviations of the transmission from the detection window reference time. But, in case of large cells (e.g. inter-site distance $>$ 1500m) and for channels with high delay spreads (e.g. ETU [14]), the

deviation can become larger, especially for the devices afar from the base station, and can surpass the selected CP value. In this case, the transmission is considered asynchronous to the detection window and it produces interference to the transmissions adjacent in frequency. The power of this interference is affected by various parameters (e.g. the size of two bursts, the existence of guard bands between them, etc.). In FANTASTIC-5G a set of new waveforms are proposed with properties which can limit and in some cases eliminate the interference effects due to asynchronicity [12], [13].

III. mMTC STATE OF THE ART

Several ongoing efforts aim to support mMTC in commercial communication systems, but most of these only support parts of the mMTC requirements. Short payload packets and extended coverage are already available in some of the solutions. However, the problem of a massive number of devices attempting access has not been solved. In the following, we provide a short overview of mMTC systems currently available or under development covering 3GPP systems as well as Non-3GPP systems.

A. NON-3GPP LOW POWER WIDE AREA NETWORKS

LoRA is a Low Power Wide Area Network (LPWAN) and is typically laid out in a star-of-stars topology in which gateways relay messages between end-devices and a central network server in the network back-end, [2]. The communication between end-devices and gateways is spread out on different frequency channels and data rates. The selection of the data rate is a trade-off between communication range and message duration offering a range of 0.3 kbps to 50 kbps through an adaptive data rate scheme. The access is based on a proprietary chirp based spread spectrum scheme and the MAC protocol is based on frequency and time ALOHA. LoRA operates in the sub-GHz bands and the vendors claim coverage on the order of 10–15 km in rural areas and 3–5 km in urban areas.

Sigfox is also a LPWAN that supports infrequent bi-directional communication, employs ultra narrowband (UNB) wireless modulation as access technology, while the MAC protocol is based on frequency and time ALOHA [1], [15]. The upper layers are proprietary and their definition is not public. The vendor claims coverage on the order of 30–50 km in rural areas and 3–10 km in urban areas.

IEEE 802.11ah is a WAN, offering low-power and long-range operation. The operating frequency of IEEE 802.11ah is below 1 GHz, allowing a single access point (AP) to provide service to an area of up to 1 km. The PHY and MAC protocol operation is similar to the one present in the 802.11 family of protocols, extended with the introduction of restricted access window during which only certain number of devices are allowed to contend based on their device IDs [16].

There are other network systems built on top of IEEE 802.15.4 (6LoWPAN, ISA100.11a, WirelessHart) which are focused on low number of devices while providing reliability guarantees. Finally, there are other network systems with their own protocol stack such as Ingenu and Weightless.

All of the presented LPWANs assume a rather simple physical layer processing and are not capable of coping with massive number of simultaneously active devices.

B. 3GPP LOW POWER WIDE AREA NETWORKS

Until recently, MTDs were being served by 2G based solutions. However, with the success of non-3GPP technologies as previously described, such as LoRA, Sigfox, Ingenu and Weightless, the cellular industry decided to accelerate the definition of an efficient MTC set of solutions and came up with solutions standardized in 2016. The aim was to introduce new features to the LTE releases that would support IoT-like devices and would exploit the existing 4G coverage around the world. However, these new features would need to align with the new IoT key requirements which can be summarized as following:

- Low cost receiver devices ($2 - 5$ \$);
- Long battery life (> 10 years)
- Extended coverage ($+15$ dB) over LTE-A

In order to achieve the three objectives (cost, power efficiency, extended coverage), design choices were made:

- Single antenna design (to reduce cost)
- Half duplex transmission (to reduce cost)
- Narrow band reception (to reduce cost, power consumption)
- Peak rate reduction (to reduce cost, complexity)
- Limited MCS and limited number of Transmit modes (to reduce complexity)
- Lower transmit power (to reduce power consumption)
- Extended DRX and new power saving modes (to reduce power consumption)
- Transmission repetition (for enhanced coverage)

Three types of IoT devices are currently supported in the 3GPP standards up to Release 13. These are the category M1 (Cat-M1), NB-IoT (NB1) and the extended coverage GSM (EC-GSM). The latter solution targets a very specific market (2G only) and is most likely to stay as a niche technology as the 2G systems spectrum resources are re-farmed into 4G.

1) CAT-M1

The eMTC (now denoted as cat-M1) comes from the need to support simpler devices than the UE types defined currently, while being capable to take advantage of the existing LTE capabilities and network support. The changes in comparison with the LTE system take place both at the device and at the network infrastructure level, where the most important one is the reduction of the device-supported bandwidth from 20 MHz to 1.4 MHz in both downlink and uplink [17]. The main consequence of this change is that the control signals (e.g. synchronization or broadcast of system block information) which are currently spread over the 20 MHz band, will be altered to support the coexistence of both LTE-M UEs and the standard, more capable, UEs.

Another important feature of this new UE category is the reduced power consumption, achieved by the transceiver-chain complexity and cost reduction, such as support of uplink and downlink rate of 1 Mbps, half-duplex operation, use of a single antenna, reduced operation bandwidth of 1.4 MHz, and reduction of the allowed maximum transmission power from 23 dBm to 20 dBm. Furthermore, there is the requirement to increase the cellular coverage of these LTE-M UEs by providing up to 15 dBs extra in the cellular link budget.

The preamble structure and access procedure are the same as in LTE, with the introduction of a simplified procedure without the security overhead. It is focused on increasing coverage, while still keeping LTE-like functionality.

2) NB1

The NB-IoT (also denoted as NB1) pertains to a clean slate design of an access network dedicated to serve a massive number of low throughput, delay tolerant and ultra-low cost devices. NB-IoT can be seen as an evolution of eMTC in respect to the optimization of the trade-off between device cost and capabilities; as well as a substitute to legacy GPRS to serve low rate IoT applications. The main technical features are: (i) reduced bandwidth of 180 kHz in downlink and uplink; (ii) maximum device transmission power of 23 dBm; and (iii) increased link budget by 20 dB extra when compared with commercially available legacy GPRS, specifically to improve the coverage of indoor IoT devices. This coverage enhancement can be achieved by power boosting of the data and control signals, message repetitions and relaxed performance requirements, e.g. by allowing longer signal acquisition time and higher error rate. An important enabler for this coverage enhancement is the introduction of multiple coverage classes, which allow the network to adapt to the device's coverage impairments.

It has a new PRACH structure based on multi-hopping and is not based on Zadoff-Chu sequences like in LTE. There are three versions of the access protocol (full similar to LTE, medium similar to the optimized access in eMTC and light with a preamble followed by data transmission). The main focus of the NB-IoT is on providing extreme coverage, with supported number of users similar to LTE-M [18].

IV. mMTC IN A MULTI-SERVICE AIR INTERFACE

A. SYSTEM MODEL AND ASSUMPTIONS

In general, we assume mMTC to be part of a multi-service air interface suitable to serve all services envisioned in 5G in a single air interface [19], as depicted in Fig. 1. The base physical layer assumption for such a multi-service air interface is a multi-carrier system with a suitable waveform and flexible numerologies as standardized for New Radio (NR) in 3GPPP. Thus, the mMTC service (denoted as MMC in FANTASTIC-5G) may use part of a resource block grid as it is depicted in Fig. 1 and can be organized using all or part of these resources. Of course, the amount of resources available for mMTC and the numerology used will vary according to higher layer management functionalities that balance service requirements in a given scenario or cell. For example, LTE provides only limited resources for the PRACH that facilitates the access reservation protocol in LTE and thereby limits the number of serviceable users.

In contrast to this general view on a multi-service architecture, we aim to present different solutions in a comparable framework such that the access protocol performance can be gauged by different key performance indicators (KPIs). Thus, for evaluation of the proposal described in section V and VI, we consider a single cell scenario using the basic PHY layer assumptions summarized in Table 1. This allows the evaluation of the base performance of different MMC PHY/MAC concepts. Furthermore, we assume a generic OFDM waveform as base assumption that excludes topics like synchronization robustness or service separation solved by appropriate waveform choices [13].

TABLE 1. Basic assumptions for MMC evaluation.

Parameter	Value	Explanation
TTI	1 ms	
Bandwidth	10 MHz	50 PRBs per TTI
Allocation size	1 PRB	1 PRB = 1 ms x 180 kHz
No. of antennas	1	Base assumption is single antenna at UE and BS
Traffic model	Poisson	Arrival rate λ
Packet size	8 Bytes	
Mean waiting time	0.5 ms	Avg. time offset between wake-up of the UE and the beginning of the next TTI when a SR is sent
ACK/NACK response time	3 ms	A request or packet sent in TTI i is followed by ACK/NACK at TTI $i+3$, earliest retransmission then is TTI $i+4$
Random Back-off	0..10 ms	Uniform distribution, back-off after NACK
Max Retransmissions	4	The fourth NACK is the "final" NACK

B. BUILDING BLOCKS

In order to address the mMTC challenges we have identified a number of building blocks that are classified into (i) Physical Layer, (ii) MAC layer, (iii) RRC layer and (iv) Waveforms. The focus of this paper is on the first two, i.e. the physical and MAC layers. However, PHY and MAC enhancements alone will not be able to solve the massive access challenges. Therefore, we also provide a short outlook on RRC and waveforms.

1) PHYSICAL LAYER

The design of access reservation protocols is usually based on idealized assumptions about the PHY performance and behavior. A classical assumption in contention based protocols is that concurrently active users are colliding and cannot be retrieved. Recently, MAC protocol analysis took the capture effect [20] into account, i.e. the decodeability of users with sufficiently different powers such that at least

one can be still decoded. PHY layer technologies that are able to resolve more collisions through advanced receiver processing like successive interference cancellation (SIC) have been in focus to enhance the performance of the overall access protocol. Furthermore, the performance of such technologies in different fading scenarios as well as under the assumption of asynchronous communication, strongly determines the performance baseline of all MAC protocols based on specific PHY solutions. In FANTASTIC-5G we studied different PHY collision resolution techniques in combination with various access protocols. On the one hand classical multi-user detection (MUD) as well as Compressive Sensing based enhancements are considered, and on the other hand also Compress- or Compute-and-Forward based schemes are considered that can be closely related to or even combined with advanced protocols like Coded Random Access.

2) MEDIUM ACCESS CONTROL LAYER

We distinguishes three types of access protocols: (a) multi-stage; (b) two-stage; and (c) one-stage. These can be interpreted very differently, and each of the three types may contain several access protocol variants. We depict these in Fig. 2.

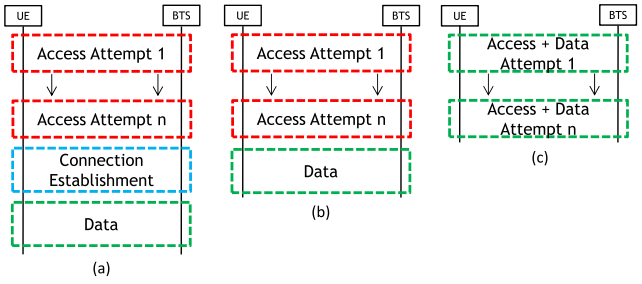

FIGURE 2. High level description of the three considered access protocols types: (a) Multi-stage access protocol with an access, connection establishment and data phase; (b) Two-stages access protocol with access and data phases; and (c) One-stage access with combined access and data phase.

A multi-stage access protocol (a), for which the current LTE connection establishment protocol is a prime example, is composed of at least three phases, the access, connection establishment (including authentication and security) and finally the data phase. A two-stage access protocol (b) allows, the UE to separate the access notification stage with its data delivery stage, e.g. through an intermediate feedback message. This leaves room for feedback and resource allocation to the UE from the eNB. The feedback could be power control and timing alignment. What is meant by a one-stage access protocol (c) is that both the access notification and data delivery need to be done in a single transaction, e.g. using one or several consecutive packets or in a single transmission. All three types of access protocols can lead to scheduled access mode, where the devices after establishing a connection to the network do not need to re-establish access in future attempts.

The work presented in this paper focuses on two-stage and one-stage access protocols. As a common assumption the signalling associated with the first connection establishment (mostly the establishment of mutual authentication and security) is assumed to be reused from a previous session where some incarnation of a full multi-stage access protocol took place.

3) RADIO RESOURCE CONTROL LAYER

A major observation beyond PHY and MAC layer was that the transition from idle mode to connected mode and vice versa used in today's systems must be simplified or even avoided. Connectionless transmission of small packets from UEs once registered and authenticated in the network may reduce the required number of signalling messages significantly. In this case, a small packet must comprise both source and destination addresses and payload. An important component in the reduction of the required signalling, upon connection establishment, is the addition of new RRC states such as the RRC extant state (see [21]) which will allow the devices to maintain the security context active over a long period.

4) WAVEFORMS

Another major conclusion is that due to properties of new 5G waveforms tight uplink synchronization is not required anymore for the small packets typical in mMTC (see [12], [13]). This allows to compress or even avoid broadcast messages that are usually required for synchronous operation. One example is the random access response (RAR) in LTE which consists of 56 bits for each UE that has sent a detected preamble. Essentially, RAR comprises a temporary identifier, a time offset value, and a grant for the subsequent signalling messages. While for the one-stage access protocols discussed here the RAR can be completely omitted, others like the contention-based two-stage variants combined with new waveforms can significantly reduce such overhead.

C. KEY PERFORMANCE METRICS

To evaluate the performance of our various contributions on the PHY and MAC layers detailed in sections V and VI we consider two key performance indicators:

- The **Protocol Throughput** (TP) denotes the total number of served devices per TTI. It directly addresses the massive access problem by showing how many users can be served given a certain access load.
- The **Access Latency** (AL) measures the amount of time (measured in TTIs) between T_1 the time instant when a device has new data to transmit (packet arrival at the device) and T_2 the time instant when the device's data is received successfully (packet arrival at the receiver, which in most cases is the Base Station). Here, it complements the throughput to provide a complete view. Without latency considerations the throughput could be arbitrarily enhanced by aggregation of access opportunities and longer back-off times. Therefore, technologies

can only be fairly compared if both KPIs are considered together.

This manuscript gives a compact overview of the proposed protocols and highlights evaluation results obtained in the EU funded project FANTASTIC-5G. More details on the evaluation of the proposed protocols and additional results can be found [21].

V. MAC PROTOCOL PROCEDURES

In this section we present three different MAC layer approaches using idealized models of the physical layer. First, we present results for One-Stage vs Two-Stages Access Protocols (OSTSAP) with different number of preambles and additionally exploiting decoding of multiple collisions (capture effect) showing that one-stage protocols offer better latency whereas two-stage protocols with collision resolution allow for much higher throughput. Second, we present Signature based Access with Integrated Authentication (SBAIA) that extends the idea of random access preambles like in standard LTE to a signature formed of multiple preambles enabling much higher throughput with added functionality like authentication. Finally, we present Non-Orthogonal Access with Time-Alignment Free Transmission (NOTAFT) that exploits the relaxed timing constraints of Pulse-shaped OFDM and MIMO processing to lower the signalling overhead for MTDs and enable massive access.

A. ONE-STAGE VS TWO-STAGES ACCESS PROTOCOLS (OSTSAP)

In this section we describe implementation variants of the generic two-stage and one-stage schemes shown in Fig. 2(b) and (c), respectively. In contrast to the other solutions presented in this paper, the performance evaluation is limited to pure protocol performance. In case of single-user detection (SUD), this means that two packets collided on the same data resource are always lost, whereas a single packet is always successful. In case of multi-user detection (MUD), we apply an idealized model to get the upper bound of the potential performance gain [22]. We assume that at most two superimposed packets on the same data resource can be decoded given that the UEs have utilized different preambles. Unpredicted overlapping of more than two packets leads to the loss of all of them. Of course, this scheme can be easily extended to more than two users. In a more general view, the probability of successful decoding of any packet $P(n)$, given that n packets overlap, depends on multiple parameters, e.g. the distribution of receive power at the BTS, the modulation and coding scheme (MCS) and the multiple access method on the PHY layer itself [20].

A detailed introduction of the protocol options can be found in [23]. The following paragraphs briefly summarize the two-stage protocol in Fig. 2(b): The UE sends a random preamble sequence, also referred to as service request. We assume a set of S sequences which can be uniquely detected and separated at the BTS through a correlation receiver. However, the BTS cannot distinguish whether just one single UE or several UEs have sent the same sequence. The latter case is referred to as preamble collision. With increasing S, this probability can be reduced at cost of a larger amount of required radio resources M_S. Given a constant number of resource units per time slot, $M = M_S + M_D$, increasing M_S reduces the available data resources M_D accordingly. Without loss of generality we assume in the following a preamble signal generation and transmission scheme equivalent to the Physical Random Access Channel (PRACH) in LTE, and resource units mimic a Physical Resource Blocks (PRB) stacked in frequency dimension.

The BTS broadcasts information related to the assignment of radio resources. In the simplest case this is a binary vector V of length S indicating whether or not the originator of the respective sequence is allowed to transmit its data packet in the second stage. This implies a fixed mapping between preamble sequence and data resource. Typically, the number of sequences S exceeds the number of data resources M_D by an over-provisioning factor N, i.e. $S = NM_D$. Consequently, N sequences point to one single data resource. The BTS without multi-user capability will therefore acknowledge just one detected preamble and reject the remaining. In case of MUD, a second detected preamble is acknowledged as well. A more sophisticated feedback scheme comprises a resource index instead of just one bit ACK/NACK, allowing the BTS for a fully flexible assignment of the detected service requests to the available data resources at cost of a larger downlink signalling overhead. A further enhanced scheme includes additionally the queue length of waiting UEs that could not yet be served. This enables a distribution of the detected service requests in both frequency and time domain, i.e. surplus service requests are automatically shifted to the next free time slot. In the second stage the acknowledged data packet transmission takes place. In case of any error, the retransmission scheme with parameters in Table 1 is initiated.

In the one-stage protocol shown in Fig. 2(c), the intermediate feedback after preamble detection is missing. The main advantage is the acceleration of the complete process. Preamble for activity detection and data packet can be transmitted in the same time slot. However, the capability of the two-stage protocol to control data packet transmissions and to reduce collisions is not present any more. It is therefore straightforward to combine one-stage access with MUD. A significantly high over-provisioning factor N allows the BTS to separate the service requests and to gain awareness how many data packets overlap on each of the M_D resources.

In Figure 3 the achievable protocol throughput with SUD and MUD is depicted as a function of the arrival rate λ for different large sets of preamble sequences S. Obviously, the two-stage protocol outperforms its counterpart with respect to throughput. Main reason for this result is the possibility to assign the available data resources through the intermediate feedback after preamble detection, and consequently to reduce collisions. With a larger set of preambles S, the performance can be significantly improved, especially in high load

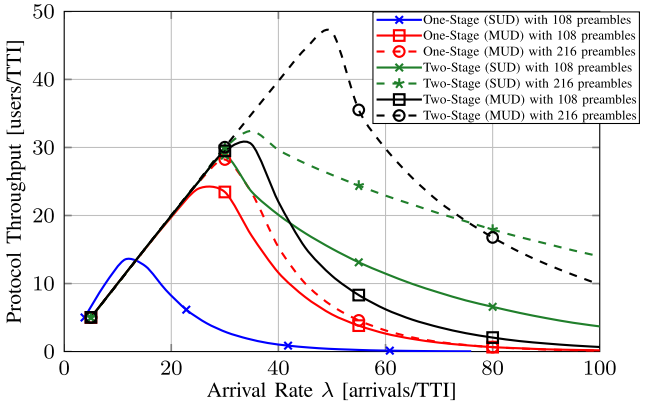

FIGURE 3. Protocol throughput of the one-stage and two-stage variants with SUD and MUD depending on the number of preambles S as a function of the arrival rate λ.

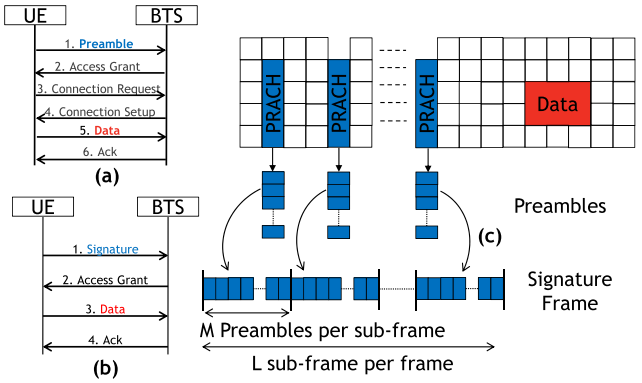

FIGURE 5. (a) LTE-based two-stages random access (b) Signature-based two-stages random access; (c) Physical random access resources mapping to random access preambles and Signature frame constructed from *L* sub-frames composed each by *M* random access preambles.

situations (arrival rate $\lambda > 30$), and motivates further efforts to optimize the preamble sequence design for 5G. MUD improves the performance for both the one-stage and the two-stage access protocol because the number of resources is virtually increased. We remark that only cases with the same over-provisioning factor N, i.e. the same ratio of available preambles and data resources can be directly compared, e.g. two-stage SUD with 108 preambles (solid dark green) and MUD with 216 preambles (dotted light green).

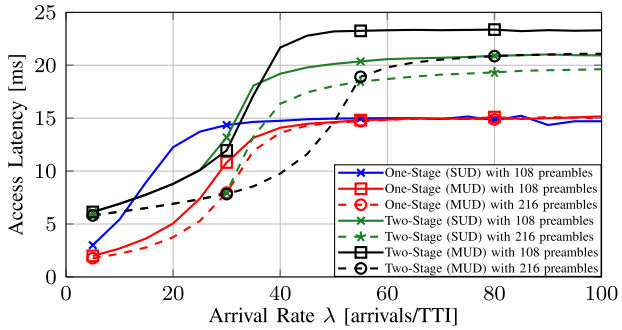

FIGURE 4. Access latency of the one-stage and two-stage variants with SUD and MUD depending on the number of preambles S as a function of the arrival rate λ.

Figure 4 shows the achievable access latency of successful packet transmissions with SUD and MUD. We see that the one-stage protocol overall can achieve significantly smaller delays if the traffic load is very low. A combination with MUD further reduces the access latency. The good result for very high load is misleading in this respect because the corresponding throughput in Fig. 3 is close to zero. In the range around $\lambda = 25$, the two-stage protocol benefits from the lower collision probability, i.e. smaller retransmission rate. We further see that a larger set of preambles S can also provide some gain regarding access latency and that combination with MUD is advantageous as well.

B. SIGNATURE BASED ACCESS WITH INTEGRATED AUTHENTICATION (SBAIA)

In the LTE(-A/Pro) random access protocol, depicted in Fig. 5(a), each device contends for access within a Physical Random Access Channel (PRACH) by selecting randomly one of the M available preambles. In case the device's access attempt is not successful (i.e. the preamble selected by the device was also activated by at least one other device or it was not detected at all), then the device will back-off and re-attempt access later. This procedure is repeated until the device is either successful or the amount of allowed retransmissions is exceeded. In case the access attempt is successful, the device has then to inform the network about its identity and how many resources it requires to transmit its data payload. This protocol step is necessary only because the transmission of a preamble does not encode any information about the device nor its requirements.

In contrast, in the proposed random access scheme, depicted in Fig. 5(b), we allow each device to contend with a predefined sequence of preambles over multiple PRACHs, which we denote as the device's *signature*. These signatures, i.e. the preamble activation pattern over multiple PRACHs, are constructed based on information unique to each device (such as the device's identity). From a protocol standpoint, this signature can then be used to identify the device and its requirements (e.g. the amount of resources required to transmit its data payload). This in turn allows a significant reduction of the amount of exchanges in the access protocol to achieve the same functionality, as it can be seen when comparing Fig. 5(a) and (b). These signatures are transmitted synchronously over a frame composed of several PRACHs, as depicted in Fig. 5(c). This is made possible only if the preambles in each PRACH: (i) are orthogonal to each other; (ii) can be detected simultaneously; and (iii) allow the base station to detect a preamble even when it is transmitted by multiple devices [24], i.e. a collision in the "preamble space" is still interpreted as an activated preamble. This last property can be interpreted as the OR logical operation, since each preamble is detected as activated if there is *at least* one device that transmits the preamble. This observation was the motivation for the use of Bloom filters – a data structure based on the OR operation for testing set membership [25] – for

the construction of the access signatures. Specifically, the device's identity is hashed over multiple independent hash functions and the resulting output used to select which preamble in which PRACH to activate. Finally, all the above properties can be obtained from preambles generated from spread sequences such as the Zadoff-Chu sequences.

In the following we describe briefly the signature construction, transmission and detection. Assume that a device's identity is given by $\mathbf{u}$ and its corresponding signature as $\mathbf{s}^{(h)} = f(\mathbf{u})$. Where $f(.)$ corresponds to the operation of hashing over multiple independent hash functions. The resulting signature can be represented as a binary vector, in such way that the bits at '0' correspond to inactive preambles, while bits at '1' represent the active preambles. As the transmission of all the devices' signatures occurs in a synchronous fashion, then the base station receiver will observe a superposition of all the transmitted signatures as,

$$\mathbf{y} = \bigoplus_{h=1}^{N} \hat{\mathbf{s}}^{(h)}, \tag{1}$$

where $\hat{\mathbf{s}}^{(h)}$ is the detected version of $\mathbf{s}^{(h)}$. The detection if a given signature is active is done by testing if the following holds

$$\mathbf{s} = \mathbf{s} \bigotimes \mathbf{y}, \tag{2}$$

where $\bigotimes$ is the bit-wise AND.

The drawback of this signature construction is that even in the case of perfect preamble detection and no false detection, the base station can still detect signatures that have not been transmitted (i.e. the corresponding device is not active) for which (2) holds. In other words, the base station may decode *false positives*. The signatures can then be designed in terms of the number of active preambles and the signature length; and in doing so control the number of false positives generated.

The signature decoding can be performed in an iterative manner, since the base station will receive each PRACH sequentially; and compare each of the observed active preambles with the valid signatures. This approach is inspired by the fact that the active preambles, which constitute a signature, are randomly spread over the PRACHs of the signature frame and, in principle, the base station does not need to receive all of them to detect that the signature has been transmitted. As the signature of a device is detected, the device is notified and granted access to the channel, following the access protocol depicted in Fig. 5(a).

In the following we provide a comparison in terms of protocol throughput and access latency compared with an LTE(-A/Pro) baseline. The PRACH configuration follows the details in Table 1. The mean number of arrivals is assumed to be known, and the signature based scheme dimensioned for it. The probability of preamble detection by the base station is set to $p_d = 0.99$ and the probability of false detection of a preamble is set to $p_f = 10^{-3}$ [26]. In the baseline, i.e. LTE(-A/Pro) scheme, we assume the typical values for the backoff window of 20 ms, a maximum number of 10 access attempts, 10 ms until the grant message is received and 40 ms until the connection setup (collision resolution) is received. We assume that PRACH occurs every 1 ms, where there are 54 available preambles for contention per PRACH in the LTE baseline which require 6 dedicated PRBs; while for the proposed scheme we assume that 216 preambles are available per PRACH that require 12 of the available PRBs for their generation.

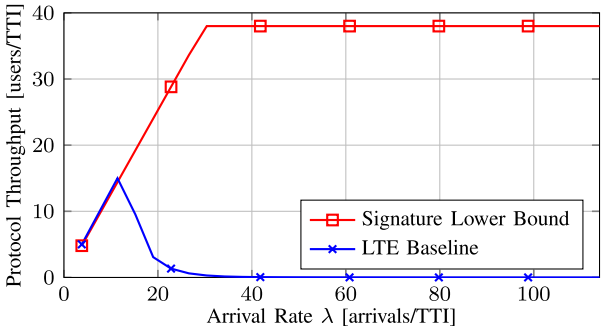

FIGURE 6. Protocol throughput for signature based access with 216 preambles.

The protocol throughput achieved by this scheme is provided in Fig. 6. Note that the result provided is the lower bound throughput, yet for higher loads the throughput will not go beyond 38 packets per TTI as this corresponds to the maximum available PRBs per TTI.

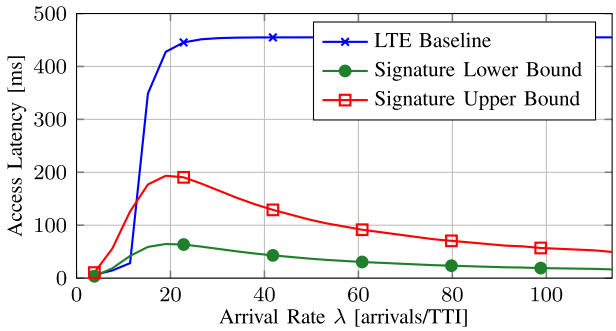

FIGURE 7. Access latency of signature based access for 216 preambles.

Fig. 7 provides the upper and lower bounds of the access latency achieved by the signature scheme, where it can be observed that both bounds decrease with the increasing arrival rate. This decrease is due to the signature length decreasing with the access load, which has a direct impact on the access latency.

Signature based random access is a novel access scheme that allows the reduction of the exchanges required to transmit small payloads in wireless access protocols. The functionality of the described protocol can be extended to include authentication and security establishment and prioritization of traffic [27], [28]. This is possible, since the access pattern can be made in such a way to encode any kind of information.

C. NON-ORTHOGONAL ACCESS WITH TIME ALIGNMENT FREE TRANSMISSION (NOTAFT)

In the current LTE system, both CP-OFDM and DFTs-OFDM impose strict synchronization requirements to the system. In order to guarantee reliable link performance, the timing inaccuracy of the receiving window needs to be kept within the range of the cyclic prefix. In the cellular uplink, however, the mobility of the users yields a continuous change in the propagation delay of their transmission signals, and thus introduces time-variant timing offsets. In order to tackle such random and variable timing misalignment, a closed-loop time alignment (TA) procedure is implemented in the LTE systems for enabling the BS to track each individual user's uplink timing during an active connection. However, for MTC with stringent power consumption limitations and sporadic activity with rather short data packets, it is desirable to design a simplified access procedure that can enable a grant-free and TA-free transmission of a short data packet in a single shot, yielding the one-stage access according to Fig. 2.

The first requirement derived from the above problem statement is the time asynchronous transmission, which is a feature supported by enhanced multi-carrier schemes like pulse shaped OFDM [29]. Pulse-shaped OFDM (P-OFDM) fully maintains the signal structure of CP-OFDM, while allowing for pulse shapes other than the rectangular pulse to balance the localization of the signal power in the time and frequency domain. Let M be the FFT size, N be the number of samples within one symbol period and T_s be the sampling period. We consider the time-frequency rectangular lattice for the OFDM system (T, F), with $T = NT_s$ denoting the symbol period and $F = (MT_s)^{-1}$ the subcarrier spacing. The P-OFDM transmit signal can be given as

$$s(t) = \sum_{n=-\infty}^{+\infty} \sum_{m=1}^{M} a_{m,n} g(t - nT) e^{j2\pi mF(t-nT)}. \quad (3)$$

Here, $a_{m,n}$ is the complex-valued information bearing symbol with sub-carrier index m and symbol index n, respectively, and $g(t)$ represents the transmit pulse shape. At the receiver, demodulation of the received signal $r(t)$ is performed based on the receive pulse shape $\gamma(t)$:

$$\hat{a}_{m,n} = \int_{n=-\infty}^{+\infty} r(t)\gamma(t - nT) e^{-j2\pi mF(t-nT)}. \quad (4)$$

By carefully designing the pulse shapes $g(t)$ and $\gamma(t)$, the power localization in the time and frequency domain of a pulse can be adjusted. In this work, robustness against distortions from large timing offsets is desired. To this end, following the design approach elaborated in [30], an orthogonalized Gaussian pulse which spreads four symbol periods is adopted as the transmit and receive pulse. In comparison to CP-OFDM, it can be shown that this pulse exhibits a high resilience against timing offsets. This allows for asynchronous transmission without timing adjustment within cell coverage. Therefore, the timing alignment procedure during the random access phase can be omitted.

In contrast to the baseline assumptions outlined in Table 1, we assume pulse shaped OFDM (P-OFDM) [29] coupled with a space division multiple access (SDMA) scheme relying on multiple antennas at the BS. Coupling these two technologies facilitates a non-orthogonal grant-free access scheme supporting collision resolution based on MIMO detection techniques on the BS side. To this end, we assume that each spatial layer carries a demodulation reference signal (DMRS) orthogonal to those of the other layers.

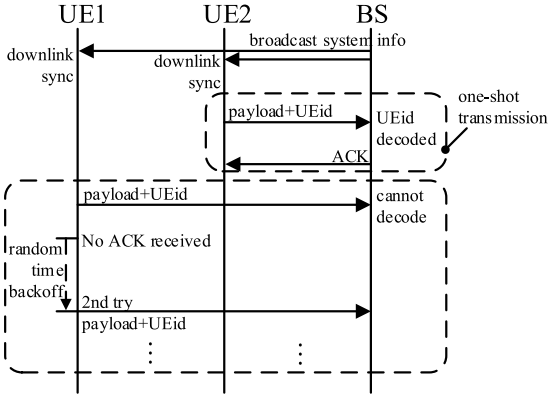

FIGURE 8. Proposed random access procedure with non-orthogonal time alignment free transmission.

The proposed random access scheme with non-orthogonal TA-free transmission is illustrated in Fig. 8 and can be described as follows:

1) UE establishes downlink synchronization to the primary cell and obtains the system configuration by decoding broadcast channel information. The broadcast information may include cell-specific reference signal setting, maximum number of retransmission and default transmission scheme.
2) The UE randomly selects a resource block and transmits its short packet data payload including its UE identifier. Here, the resource block consists of the time-frequency resource on a spatial layer which is identified by its DMRS.
3) The BS decodes the received signal. With a successfully decoded data payload, the UE can be identified and an acknowledgement is fed back.
4) If a UE receives an ACK, the NOTAFT transmission is completed.
5) If no ACK is received, a UE takes a random time backoff, and then steps 2-3 are repeated until either an ACK is received or the maximum number of retransmissions is reached.

We examine the uplink transmission in a single macrocell scenario without timing adjustment. Due to the radio propagation delay, a timing misalignment is present upon the arrival of the uplink signal at the BS. Assuming a cell radius of 2 km, this timing misalignment is calculated according to the propagation delay of the round trip, laying approximately in the range of $[0, 13]\,\mu s$. Link performance

evaluation in [29] shows no significant loss for such scenario when P-OFDM with an appropriately designed pulse spanning four symbol durations is employed. Therefore, for the protocol evaluation, we assume that packet loss is not caused by the timing misalignment, but only by the resource collision, i.e. if two UEs select the same spatial layer on the same resource block. Since access preamble is used, the total number of available resource blocks, i.e. 50 PRBs, can be employed for non-orthogonal data transmission for 10 MHz mode. With a typical setting of four antennas on the BS side, this amounts to a total of 200 random access opportunities per TTI. This scheme is compared to the multi-stage access scheme with TA, depicted in Fig. 2. Parameters listed in Table 1 are applied.

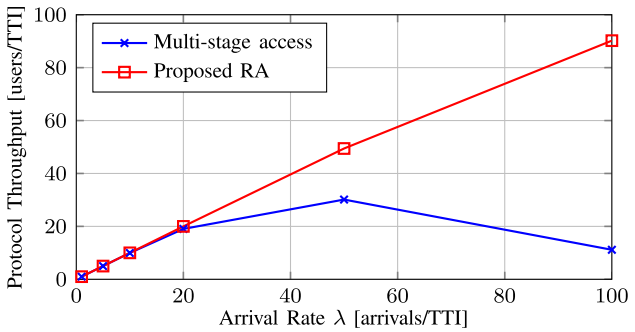

FIGURE 9. Protocol throughput of the non-orthogonal access with time alignment free transmission.

Fig. 9 depicts the achievable packet throughput as a function of the arrival rate. Since no resource is allocated for the random access procedure, all PRBs are utilized for data transmission. Given a much higher number of random access opportunities, the proposed NOTAFT scheme offers significantly higher throughput especially when the arrival rate is relatively high.

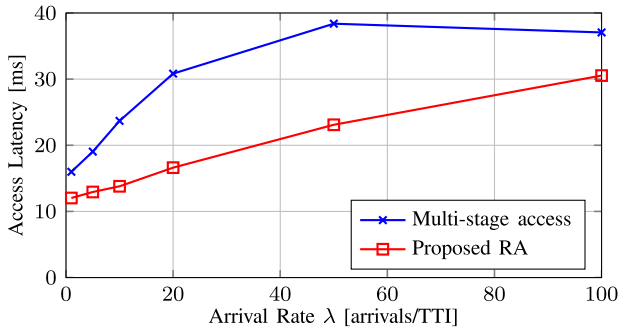

FIGURE 10. Access latency of the non-orthogonal access with time alignment free transmission.

As shown in Fig. 10, since the timing adjustment procedure is removed, the proposed one-shot transmission scheme exhibits lower access latency compared to the baseline approach.

In summary, the proposed access procedure facilitates a 'single-shot transmission,' enabling a reduced end-to-end latency as well as a lower signalling overhead for short packet transmissions. Thanks to this, it could substantially extend the battery life of devices for a better sleep/wake-up operation.

VI. PHY AND MAC INTEGRATED SCHEMES

In the following we present four approaches that extend the pure MAC protocol view of the previous section in terms of the physical layer assumptions. Here, all presented results include simulation of physical layer transmission at least including coding and modulation and in most cases also channel estimation. First, we present Compressive Sensing Multi-User Detection (CSMUD) which exploits sparsity due to sporadic activity in mMTC enabling efficient Multi-User detection in each random access slot of a slotted ALOHA setting. Second, we present Coded Random Access with Physical Layer Network Coding (CRAPLNC) extending the CSMUD approach to frames using ideas from network coding, which results in a high throughput Coded Random Access scheme. Third, we present Compressive Sensing Coded Random Access (CCRA) that combines Coded Random Access CSMUD with an underlay control channel significantly reducing control overhead. Finally, we present Slotted Compute and Forward (SCF) focusing on very dense networks with high numbers of mini base stations forwarding messages to a full base station to efficiently enable mMTC scenarios.

A. COMPRESSIVE SENSING MULTI-USER DETECTION (CSMUD)

The massive access problem outlined in section II is characterized by a massive number of MTDs that do not send information continuously but rather sporadically in large time intervals or even event driven. As already outlined from a MAC perspective different access protocols can structure such a sporadic access pattern. Still, the physical layer design of the access procedure remains open and naturally depends on the MAC protocol choice. Focusing on a one-stage protocol early works on sporadic access in combination with Code Division Multiple Access (CDMA) already noted that intermittent user activity leads to a multi-user detection (MUD) problem with sparsity that required novel algorithmic solutions [31]. Most importantly, with the development of compressive sensing (CS) a new mathematical tool was available to solve MUD with sparsity [32]. A major advantage of combining compressive sensing ideas and MUD lies in the theoretical guarantees of CS for under-determined detection problems. Prior to the so-called Compressive Sensing Multi-user Detection (CSMUD) most MUD problems with sparsity focused on fully-determined systems where the number of resources and users coincide. With CS detection guarantees can be given even if the number of resources is strictly smaller than the number of users which enables user detection even in highly overloaded CDMA setups. From this basic idea CSMUD has been extended in multiple directions ranging from non-coherent communication [33] to channel estimation with user activity detection. In the following we will

revisit the CSMUD ideas for channel estimation with simultaneous user activity detection [34] and present numerical evaluation result in combination with a simple one-stage protocol. The basic CSMUD ideas presented in the following also serve as an introduction to the presented solutions in section VI-B and VI-C.

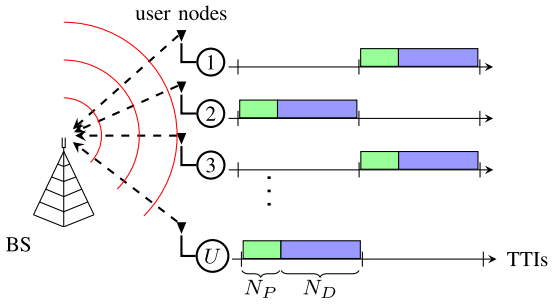

FIGURE 11. Sporadic uplink transmission of multiple devices sending N_P pilot symbols and N_D data symbols to a BS.

Following the assumptions laid out earlier, i.e. a certain time and frequency budget is allocated to the MMC service and it is well separated and robust by choice of an appropriate waveform, Fig. 11 depicts a schematic view on the MMC access protocol. Each TTI all N_{act} active users out of the overall U users access the system by transmitting N_P pilots and N_D data symbols both spread over the whole bandwidth through one of N_S pseudo-noise (PN) spreading sequences $\mathbf{s}_i \in \mathbb{C}^{L_S}\ \forall\ i = 1, \ldots, N_S$. The number of available spreading sequences N_S and their length L_S determine the physical layer performance of CSMUD. If the number of active users N_{act} is in the order of or larger than the number of available spreading sequences N_S, collisions will occur. If fewer active users access the system than spreading sequences are available the system's performance will be dominated by CSMUD performance, i.e. the separation of N_S PN sequences of length L_S. Obviously, the longer the spreading sequence, the lower the achievable data rate given TTI length and bandwidth from Table 1 but the higher the robustness and separability of spreading sequences. The resulting trade-off between MUD performance and collision probability in dependence of retransmission is highly non-trivial. Only the physical layer design trade-off between N_P and N_D was already investigated [34], but the interaction with different MAC protocols is still an open problem. Hence, we will restrict the presented evaluation results to a single parameterization that is designed to achieve the packet size of Table 1.

As indicated in Fig. 11 each user sends a packet of two parts. The first part consists of N_P pilot symbols that are unique per user and serve to estimate channel and activity through CSMUD. The second part consists of the spread N_D data symbols that can be detected and decoded through standard approaches. Each slot is assumed to occupy 10 MHz and 1 ms per table 1. To formalize the task of CSMUD we summarize the user channels $\mathbf{h}_i \in \mathbb{C}^{N_h}\ \forall i = 1, \ldots, U$ in a stacked channel vector $\mathbf{h} = [\mathbf{h}_1^T, \ldots, \mathbf{h}_U^T]^T \in \mathbb{C}^{UN_h}$. Due to the sporadic activity, channels of inactive users will be modeled as zeros, i.e. for inactive user $\mathbf{h}_i = \mathbf{0}_{N_h}\ \forall i \in \bar{\mathcal{Z}}$, where $\bar{\mathcal{Z}}$ and $\mathcal{Z}$ denote the index set of all active and inactive users, respectively. This leads to additional structure in the detection problem, i.e. the vector $\mathbf{h}$ is strictly group-sparse with groups of size N_h. The joint channel and activity signal model is then,

$$\mathbf{y} = \mathbf{S}\mathbf{h} + \mathbf{n}, \tag{5}$$

where $\mathbf{S} \in \mathbb{C}^{M \times N}$ denotes the preamble matrix containing all user preambles, $\mathbf{y} \in \mathbb{C}^M$ denotes the received signal consisting of the superimposed N_P pilots of all users at the base station and $\mathbf{n} \in \mathbb{C}^M$ summarize all noise sources as AWGN. The preamble matrix $\mathbf{S}$ exhibits a Toeplitz structure per user describing the convolution of channel and pilots, i.e. $\mathbf{S} = [\mathbf{S}_i, \ldots \mathbf{S}_U]^T$ with $\mathbf{S}_i$ being a Toeplitz matrix of user i pilots $\mathbf{s}_i$.

Depending on the underlying system assumptions (asynchronicity, waveform, channel model, etc.) the exact values of M and N vary, but are dependent on the number of users U, the pilot length N_P and the length of the channel impulse response N_h. To simplify notation we focus on a one-tap Rayleigh fading channel, i.e. $N_h = 1$. Then, the detection problem can be cast as

$$\hat{\mathbf{h}} = \underset{\mathbf{h} \in \mathbb{C}^N}{arg\,min}\ \|\mathbf{h}\|_0 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{S}\mathbf{h}\|_2 < \epsilon, \tag{6}$$

which is easily extended to N_h-tap Rayleigh fading channels if a group sparsity constraint is introduced (cf. [34]). The minimization in (6) targets the sparsest vector denoted by the "pseudo-Norm" $\|\mathbf{h}\|_0$ that counts the number of non-zeros given an ℓ_2-norm constraint to adhere to a given noise level dependent on ϵ. The solution of (6) can be approached in many different ways like convex relaxation or sub-optimal Greedy approaches which are meanwhile very well covered in the literature.

To evaluate CSMUD with respect to the KPIs and assumptions outlined in section IV the physical layer approach CSMUD was combined with a simple one-stage protocol with random backoff according to the parameters of table 1. Each active user transmits its data in the current TTI and repeats this transmission in case of failure up to four times. A combined MAC and PHY numerical simulation was conducted over 10^4 trails including the full physical layer processing (encoding, modulation, channel estimation, multi-user detection, demodulation, decoding) with BPSK, a $[5; 7]_8$ convolutional code, least squares multi-user equalization and BCJR decoding. The activity and channel estimation step is achieved by the group orthogonal matching pursuit algorithm (GOMP). The spreading factor is $N_S = 32$ and up to $K = 64$ unique spreading sequences / preambles are considered. The traffic model follows a Poisson arrival process with an arrival rate as shown on the x-axis of Fig. 12 and 13. Both KPIs depend on the signal-to-noise ratio (SNR), which is here assumed to be either 0, 5 or 10 dB and identical for all users.

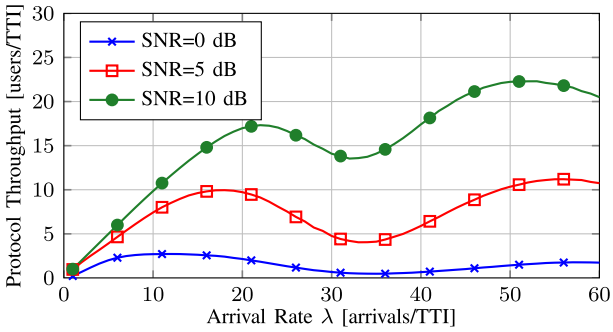

FIGURE 12. Protocol throughput for CSMUD.

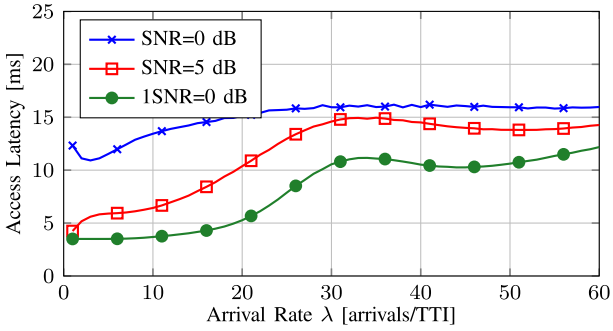

FIGURE 13. Access latency for CSMUD.

Hence, we implicitly assume some form of open-loop power control with idealized conditions.

Fig. 12 depicts the protocol throughput of the CSMUD scheme which clearly shows a nearly linear scaling with increased arrival rate for the 10 dB case up to an arrival rate of 16 at which the probability of successfully detecting and decoding a user begins to decline due to the interference level and a strong increase in retransmissions. Surprisingly, beyond $\lambda = 32$ the throughput increases again. This can be explained by the performance of the GOMP algorithm that is employed to solve (6). Up to half of the available sequences K the detection performance declines because the number of non-zeros to be estimated increases up to the maximum potential for errors at exactly $K/2$. Due to the chosen stopping criteria the estimated channels $\hat{\mathbf{h}}$ can be 100% wrong, i.e. all active users are estimated as inactive (missed detections) and all inactive users are estimated as active (false alarm). Beyond $K/2$, however, detection performance increases again with the decreasing number of zeros in the estimated vector. This is finally limited by the least squares performance of a two times overloaded CDMA system at 64 active users. Naturally, this behavior is also reflected for the lower SNRs of 5 dB and 0 dB with overall decreased performance. Note, that the CSMUD approach used here does not exploit retransmission in any way. A combined decoding approach like presented in Section VI-B can strongly improve performance in cases where single slots are overloaded. However, this is highly dependent on the specific parameters of the system [35]. Furthermore, a comparison with the results presented in Section V-A indicates that the numerical simulations presented here behave differently than the pure MAC performance given orthogonal resources. Especially, the slope is lower, but the performance peak also occurs later and seems broader hinting at a more robust behavior.

Fig. 13 presents the access latency which is very low for all presented working points and shows much lower overall latencies than other schemes. Obviously, a single transmission is sufficient most of the time for 10 dB, which is increased with lower SNR and higher arrival rates. The discussed GOMP behavior does not influence the latency as strongly but leads to small variations around the maximum latency. The access latency is much lower than for example using signature based access or the frame focused PLNC enhanced scheme described in Section VI-B. This is easily explained by the fact that the both have to aggregate multiple TTIs to facilitate a successful access compared to the setup used here.

B. CODED RANDOM ACCESS WITH PHYSICAL LAYER NETWORK CODING (CRAPLNC)

This proposal is inspired by a random access scheme aiming at reduced signalling. More specifically it considers physical layer techniques aiming to increase collision resolution through advanced receiver processing, and their integration with the MAC protocol. It targets one-stage protocols, although the PHY layer solution can be also exploited in two-stage protocols by allowing more than one packet transmission per radio resource block and resolving collisions through advanced receivers. The solution falls under the category of coded random access [36], where features of channel coding are exploited both at the slot and frame level. In particular, the scheme partially presented in [37] and [38] is extended for massive access, with emphasis on the transmission of short packets. The proposed scheme assumes a minimum coordination that ensures packet synchronization. The strategy focuses on collision resolution in a frame slotted ALOHA medium access scheme, where users are granted certain level of redundancy per transmission attempt. It exploits two features of coded schemes: the first one relates to the property that in the finite-field $\mathbb{F}_2$, although the individual messages cannot be correctly decoded, a linear combination of them (the bitwise XOR of a set of messages) may be. This property led to the so-called compute-and-forward [39], [40], which proved achievable gains, from an information-theory point of view. The second one exploits the increase in the diversity order of a linear system of equations if it is defined over an extended Galois Field $\mathbb{F}_q$ with field order $q = 2^n$.

The multiple access scheme operates as follows:

- Given a frame size of S slots, users pick at random the slot positions where they will attempt transmission of each redundancy packet[2]. In principle the scheme can operate with a different level of redundancy R

[2]In general, redundancy packets are not the same, they correspond to different codewords per user message.

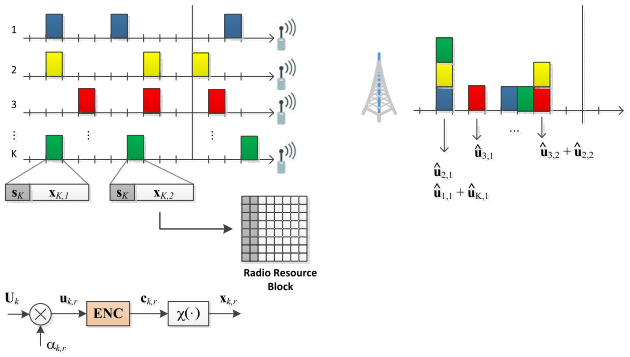

FIGURE 14. Coded Random Access, with PLNC and extended Galois field precoding.

and distribution. Although the illustrative example in Fig. 14, sets $R = 2$ for all users, the scheme can be combined with optimized distributions.

- Each message, previous to channel encoding and modulation[3], allows for a linear precoding, which consist in a symbol-wise multiplication in the extended Galois Field $\mathbb{F}_q$, i.e. $\mathbf{u}_{\mathbb{F}_q}(m) = \alpha_r \times \mathbf{U}_{\mathbb{F}_q}(m)$ where $\mathbf{U}_{\mathbb{F}_q}(m)$ denotes the m-th symbol of the non-binary representation of the binary message $\mathbf{U}$. Precoding coefficients $\alpha_r \in \mathbb{F}_q$ are generated randomly[4].
- User detection and channel estimation is enabled by means of a preamble including the user signature and small overhead for identification of precoding coefficients.
- At the receiver side, for each slot, the receiver performs user detection and channel estimation, followed by the channel decoding stage. Each decoded message or linear combination (in $\mathbb{F}_2$), generates a new row at the frame matrix $\mathbf{A} \in \mathbb{F}_q$. If $\mathbf{A}$ is full rank, collisions can be resolved without the need for having one singleton packet.

For two-stage protocols, only the PHY layer component is used, applying advanced decoding to the reception of the data transmission stage. That is, the data transmission stage can be modified to allow several users to transmit their messages over the same physical resources. It only transmits the payload data since in this scenario, the receiver knows which users are transmitting and simply takes advantage of the increased capture probability provided by the advanced decoding scheme.

Relevant aspects of the scheme rely on the detection, channel estimation and decoding algorithms applied to the received signal within a single slot. In particular, for the detection of colliding users and channel estimation in one-stage protocols, we resort to a CSMUD algorithm, as introduced in section VI-A (see also [41]). More specifically, we consider channels with no delay spread and, thus, the simple CSMUD form in (6) is sufficient. Note that in the case where a packet

[3]The same channel code and modulation among users is assumed.

[4]Note that the system can be configured to include no precoding, $\alpha_r \in \mathbb{F}_2$

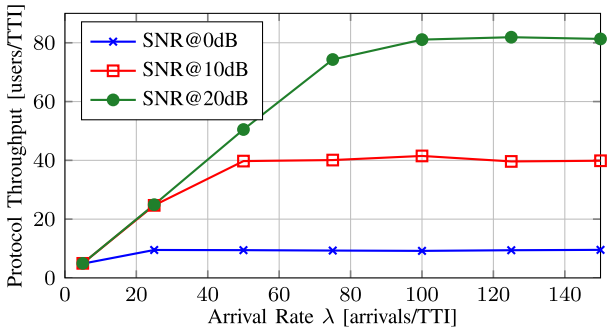

FIGURE 15. Throughput performance of CRAPLNC massive access scheme for several SNRs.

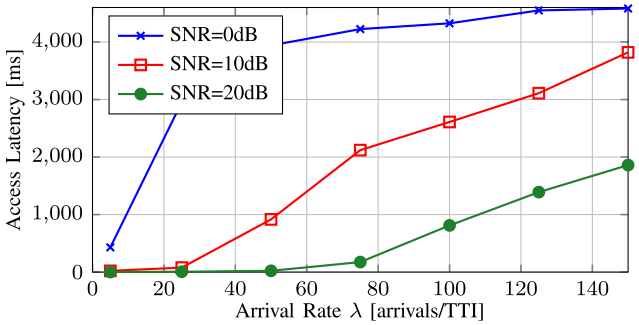

FIGURE 16. Latency performance of CRAPLNC massive access scheme for several SNRs.

fits a single radio resource block, as it is the case for the minimum allocation size of 1 PRB = 180 kHz × 1 ms (see Table I), the channel can be assumed constant. At the receiver side, advanced decoding (joint decoder and the "seek-and-decode" principle) is implemented independently at each slot, applied after standard SIC fails to decode any more messages, thus reducing complexity. Final decision decoding is made at the end of the frame (S slots), although variants to the scheme could allow faster acknowledgements as soon as individual messages are correctly decoded at each slot. Results are shown in Figs. 15 and 16 for very short codes (i.e. binary LDPC with codeword length of 164 coded symbols) and system parameters defined in Table 1 under block fading channels, shows relevant throughput gains against benchmark (slotted ALOHA) for moderate/high loads, even with no precoding. Results are also encouraging in terms of robustness against channel estimation errors, and user misdetection. We shall remark that simulation results include full physical layer implementation (multi-user detection, channel estimation and decoding) over the medium access control (for a frame size of $S = 10$ slots). Further details can be found in [21] including additional KPIs.

C. COMPRESSIVE SENSING CODED RANDOM ACCESS (CCRA)

Recent concepts combine advanced MAC protocols with Compressive Sensing (CS) based multiuser detection [42], [43]. In this section, we introduce a concept for

sparse joint activity, channel and data detection in the context of the Coded ALOHA (FDMA) protocol which we call *Compressive Coded Random Access* (CCRA) extending the work in [42] and [44]–[46]. We will argue that a simple sparse activity and data detection is not sufficient (as many papers do) because control resources are in the order of the data. In addition, we will 1) improve on the performance of such protocols in terms of the reduction of resources required for the user activity, channel estimation and data detection 2) achieve the required channel estimation quality for the successive interference cancellation procedure required in coded ALOHA and CCRA.

Let us assume for simplicity a single time slot and an OFDM system with n subcarriers. This is easily generalized to the case where there are multiple time slots, notably, within the coherence time so that channels are constant over these slots. Let $p_i \in \mathbb{C}^n$ be some signature from a given set $\mathcal{P} \subset \mathbb{C}^n$ and $x_i \in \mathcal{X}^n$ be an unknown (uncoded) data sequence (e.g. BPSK) from the modulation alphabet $\mathcal{X}^n$ both for the i-th user with $i \in \{1, ..., u\}$ and u is the (fixed) maximum set of users in the systems. Note that in our system n is a very large number, e.g. 24k. Due to the random zero-mean nature of x_i we have $\frac{1}{n}E\|p_i + x_i\|_{\ell_2}^2 = 1$, i.e. the total (normalized) transmit power is unity. Provided user i is active, we set:

$$\alpha := \frac{1}{n}\|p_i\|_{\ell_2}^2 \quad \text{and} \quad \alpha' := 1 - \alpha = \frac{1}{n}E\|x_i\|_{\ell_2}^2 \tag{7}$$

Hence, the control signalling fraction of the power is α. If a user is not active then we set both $p_i = x_i = 0$, i.e. either a user is active and seeks to transmit data or it is inactive. Whether or wether not a user is active depends on the traffic model and is discussed below.

Let $h_i \in \mathbb{C}^s$ denotes the sampled channel impulse response (CIR) of user i where $s \ll n$ is the length of the cyclic prefix (further structural assumptions on h_i are also discussed below). Let $[h_i, 0] \in \mathbb{C}^n$ denote the zero-padded CIR. The received signal $y \in \mathbb{C}^n$ is then:

$$y = \sum_{i=0}^{u-1} \mathrm{circ}([h_i, 0])(p_i + x_i) + e \tag{8}$$

$$y_{\mathcal{B}} = \Phi_{\mathcal{B}} y \tag{9}$$

Here, $\mathrm{circ}([h_i, 0]) \in \mathbb{C}^n$ is the circulant matrix with $[h_i, 0]$ in its first column. The AWGN is denoted as $e \sim \mathcal{CN}(0, \sigma^2) \in \mathbb{C}^n$, i.e. $E(ee^*) = \sigma^2 I_n$. $\Phi_{\mathcal{B}}$ denotes some measurement matrix (to be specified) where the active rows indices are collected in $\mathcal{B}$ with cardinality m. Typically, $\mathcal{B}$ refers to some set of subcarriers in case of Fourier (FFT) measurements (Φ is orthonormal matrix) but, mainly for analytical purposes, also Gaussian measurements are considered (Φ is *not* orthonormal matrix).

The key idea of CCRA scheme is that all users' preambles $\hat{p}_i$ $\forall i$ 'live' entirely in $\mathcal{B}$ while all data resides in the complement $\mathcal{B}^C$, i.e. formally $\mathrm{supp}(\hat{p}_i) \subseteq \mathcal{B}$ $\forall i$, (hence, for orthonormal matrix Φ like FFT there is no interference in between). We will call this a *common overloaded control channel* [45] which is used for *the user activity and channel detection*. Since data resides only in $\mathcal{B}^C$ the entire bandwidth $\mathcal{B}^C$ can be divided into B frequency patterns. Each pattern is uniquely addressed by the preamble and indicates where the data and corresponding copies are placed. the scheme works as follows: if a user wants to transmit a small data portion, the pilot/data ratio α is fixed and a preamble is randomly selected from the entire set. The signature determines where (and how many of) the several copies in the B available frequency slots are placed which are processed in a specific way (see below). Such copies can greatly increase the utilization and capacity of the traditional, e.g. ALOHA schemes, and which is used for *the data detection*. An illustration of the scheme is in Fig. 17.

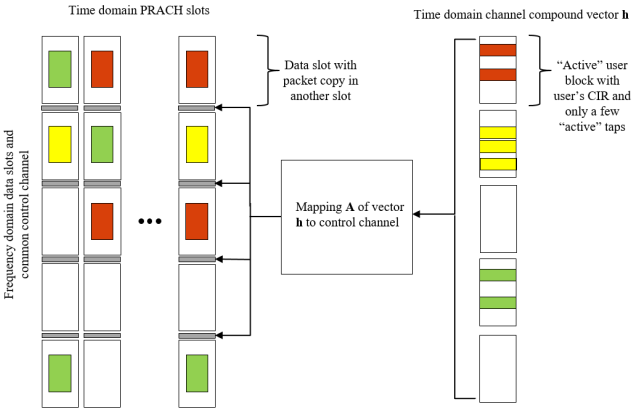

FIGURE 17. Schematic of the CCRA scheme: sets are ... the common control channel.

To derive a proper model for the user activity and channel detection, we can stack the users as:

$$y = D(p)h + C(h)x + e \tag{10}$$

where $D(p) := [\mathrm{circ}^{(s)}(p_1), \ldots, \mathrm{circ}^{(s)}(p_u)] \in \mathbb{C}^{n \times us}$ and $C(h) := [\mathrm{circ}^{(n)}([h_1, 0]), \ldots, \mathrm{circ}^{(n)}([h_u, 0])] \in \mathbb{C}^{n \times un}$ are the corresponding compound matrices, respectively $p = [p_1^T\ p_2^T\ ...p_u^T]^T$ and $h = [h_1^T\ h_2^T\ ...h_u^T]^T$ are the corresponding compound vectors. In general, the measurement map is difficult to analyse since $D(p)$ depends on the specific design of the signatures p_i. One choice of $\mathcal{P}$ that works for a small number of active users and $n \gg us$ is as follows: We choose p_0 to be a sequence with unit power in frequency domain, i.e. such that (up to phases, which can be selected according to other optimization criteria, e.g. PAPR):

$$|\left(\hat{p}_0\right)_i| = \begin{cases} \sqrt{\frac{n}{m}} & i \in \mathcal{B} \\ 0 & \text{else} \end{cases} \tag{11}$$

where $\hat{p}_0 := Wp_0$ denotes the FFT transform of p_0. Since $n \geq us$, the matrix $D(p)$ can be completely composed of cyclical shifts of the sequence p_0, i.e. $p_1 = p_0$, $p_2 = p_1^{(s)}$, $p_3 = p_2^{(s)}$, ... where $p^{(i)}$ is the i times cyclically shifted p. Hence, $D(p)$ is a single circulant matrix, and, in this

situation, we can show, that the control channel is finally represented as:

$$y_{\mathcal{B}} = Ah + z,$$

where A is a subsampled $m \times us$ FFT matrix, which is normalized by a factor of $\sqrt{1/m}$ and $z \sim \mathcal{CN}\left(0, \frac{\sigma^2}{n} I_m\right)$. Now, based on this measurement model, the most important assumptions on the structure of h are:

- Bounded support of h_i (with high probability), i.e. $\text{supp}(h_i) \leq s$ and $s \ll n$
- Sparse user activity, i.e. k_u users out of u are actually active
- Sparsity of h_i, i.e. $\|h_i\|_{l_0} \leq k_s$

Hence, classical sparsity of h is $k := k_u k_s$ and the typical arsenal of CS algorithms can be used. In CCRA, though, we are exploiting block-column sparsity: a k-sparse compound vector h is so-called block-column sparse, i.e. (k_u, k_s)-sparse, if it consists of k_u active blocks of length s each k_s-sparse. Block-column sparsity is exploited in the detection of the activity and channel by a new algorithm called Hierarchical HTP (HiHTP). HiHTP uses a so-called block-column thresholding operator $L_{k_s,k_u}(z)$. This operator can be efficiently calculated by selecting the k_s absolutely largest entries in each block and subsequently the k_u blocks that are largest in ℓ_2-norm. The strategy of the HiHTP algorithm is to use the thresholding operator L_{k_u,k_s} to iteratively estimate the support of h and subsequently solve the inverse problem restricted to the support estimate. HiHTP comes with explicit recovery guarantees while exploiting the specific structure of h, see [47].

The data detection algorithm can be seen as in instance of *coded slotted ALOHA* framework [36], tuned to incorporate the particularities of the physical layer addressed in the paper, as described in the previous section. Specifically, the random access algorithms assumes that:

- the users are active in multiple combinations of time-frequency slots, denoted simply as slots in further text,
- the activity pattern, i.e. the choice of the slots is random, according to a predefined distribution,
- every time a user is active, it sends a replica of packet, which contains data,
- each replica contains a pointer to all other replicas sent by the same user.

Obviously, due to the random nature of the choice of slots, the access point (i.e. the base station) observes idle slots (with no active user), singleton slots (with a single active user) and collision slots (with multiple active users). Using a compressive sensing receiver, the base station, decodes individual users from non-idle slots, learns where the replicas have occurred, removes (cancels) the replicas, and tries to decode new users from slots from which replicas (i.e. interfering users) have been cancelled. In this way, due to the cancelling of replicas, the slots containing collisions that previously may have not been decodable, can become decodable. This process is executed in iterations, until there are no slots from which new users can be decoded. The above described operation can be represented via graph. Analytical modelling of the above is the main prerequisite to assess the performance of the random access algorithm, which in turn, allows for the design of the probability distribution that governs the choice slots, and which is typically optimized to maximize the throughput, i.e. the number of resolved packets per slot [36].

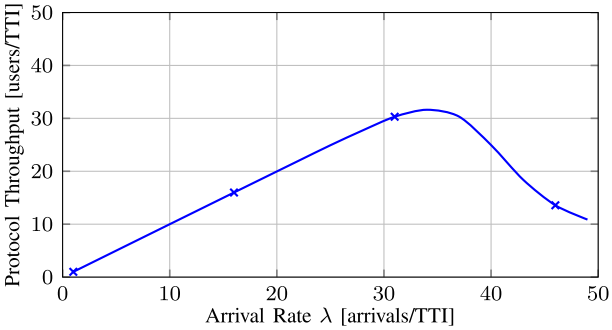

FIGURE 18. Throughput performance of CCRA over arrival rate.

We follow the common simulation assumptions described in Table I. Note that the pilot-to-data ratio is only 13% so the overhead compared to LTE-4G has significantly reduced (reported to be up to 2000% in [48]). For the CCRA throughput evaluation, we use BPSK modulated subcarriers and successive interference cancellation. Fig. 18 shows the throughput of actually successfully recovered packets over different arrival rates using *at most* three replicas per packet (optimum results from testing one to five copies). It can be seen that with three replicas the performance is significant improved over, say, traditional slotted ALOHA (SA) which achieves only max. 40% normalized throughput (i.e. 20 user/TTI). While not shown here in detail, we mention that BER performance for detecting replicas at 15 dB SNR is well below 10^{-1} even for those with three-step interference cancellation detection procedure pointing out the good channel estimation performance. Altogether, we conclude that even for this challenging scenario the CCRA achieves a significant throughput gain with reasonable BER performance per detected and decoded packet and, at the same time, drastically reduces the signalling overhead.

D. SLOTTED COMPUTE AND FORWARD (SCF)

The presented Slotted Compute-and-Forward (SCF) approach is a random access extension of the Compute-and-Forward (CF) relaying scheme introduced in [49]. The approach combines the concept of network densification with physical-layer network coding and a multicarrier transmission scheme (OFDM). Using linear codes it enables the network to exploit channel collisions [50] by decoding linear combinations of the messages transmitted by different devices that access the channel simultaneously in the same frequency band. The scheme assumes a dense network infrastructure with a large number of MTC devices accessing the wireless channel, where each transmitter can be heard by

multiple mini base stations. The data transmission is a two-hop communication with multiple mini base stations acting as relays. They receive individual superpositions of the sent signals, process, decode and forward them to the macro base station. The macro base station then estimates the transmitted messages over a finite field based on the received linear combinations. A simplified example is shown in Fig. 19.

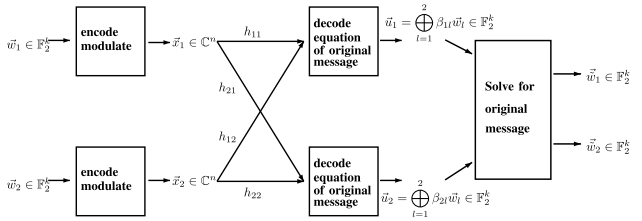

FIGURE 19. Toy example describing the main processing blocks for the 2 transmitters × 2 mini base stations case.

Let us assume that the large set of uniformly distributed MTC devices $\mathcal{M}_{tot}$, with $M_{tot} := |\mathcal{M}_{tot}|$, is supported by a set of mini base stations $\mathcal{B}_{tot}$, which are connected to the macro base station through a wired or wireless communication. Each mini base station has only knowledge of its own channel coefficients, whereas the MTC devices have no channel state information. Let $\mathcal{M} \subset \mathcal{M}_{tot}$, with $M := |\mathcal{M}|$, be a set of active MTC devices that can be heard by each mini base station $b \in \mathcal{B}$ of a predefined subset $\mathcal{B} \subset \mathcal{B}_{tot}$ and $B := |\mathcal{B}|$. Note that, for simplicity, we have assumed here $B = M$. To increase robustness it is often reasonable to choose $B > M$ and solve instead the over-determined system of equations. Each device $m \in \mathcal{M}$, has a length-k complex message $\boldsymbol{w}_m = (\boldsymbol{w}_m^R, \boldsymbol{w}_m^I)$, with $\boldsymbol{w}_m^R$ and $\boldsymbol{w}_m^I$ real, respectively imaginary part drawn from some finite field $\mathbb{F}_p^k$, and maps its message to a length-n codeword $\boldsymbol{x}_m \in \mathbb{C}^n$ subject to an average power constraint $\frac{1}{n}\|\boldsymbol{x}_m\|^2 \leq P$. We model the complex baseband signal $\boldsymbol{y}_b$ received by mini base station $b \in \mathcal{B}$ as

$$\boldsymbol{y}_b = \sum_{m \in \mathcal{M}} h_{bm}\boldsymbol{x}_m + \boldsymbol{z}_b \ , \tag{12}$$

where $\boldsymbol{z}_b \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_n)$ denotes independent Gaussian noise of unit variance (per dimension) and h_{bm} is the complex-valued channel coefficient between MTC device m and base station b.

The mini base station b performs rescaling and **integer–forcing** to obtain a noisy linear combination of the transmitted codewords with integer coefficients:

$$\tilde{\boldsymbol{y}}_b = \alpha_b \boldsymbol{y}_b = \sum_{m \in \mathcal{M}} a_{bm}\boldsymbol{x}_m + \underbrace{\sum_{m \in \mathcal{M}} (\alpha_b h_{bm} - a_{bm})\boldsymbol{x}_m + \alpha_b \boldsymbol{z}_b}_{\text{effective noise}},$$

and decode $\mathbb{F}_p$ linear combinations

$$\boldsymbol{u}_b := \bigoplus_{m \in \mathcal{M}} \beta_{bm}\boldsymbol{w}_m$$

of messages $\boldsymbol{w}_m$ over $\mathbb{F}_p$. The scaling factor α_b and the integer coefficients a_{bm} are chosen such that the effective noise is minimized. The equation coefficients β_{bm} satisfy $\beta_{bm} = [a_{bm}] \bmod p \in \mathbb{F}_p$. Once the mini base stations have successfully decoded the linear equations they forward these along with the respective coefficients $\boldsymbol{\beta}_b = (\beta_{b1}, \ldots, \beta_{bM})$ to the macro base station. If the equation coefficients have been chosen such that the matrix $\mathbb{B} := (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_B)^T \in \mathbb{F}_p^{B \times M}$ is invertible over $\mathbb{F}_p$, the macro base station estimates the original messages by calculating [51].

$$(\hat{\boldsymbol{w}}_1, \ldots, \hat{\boldsymbol{w}}_M)^T = \mathbb{B}^{-1}(\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_B)^T \ . \tag{13}$$

This approach makes the SCF solution especially suited for two-hop communication scenarios where the capacity limited second hop is a bottleneck in the transmission. For a transmission to be successful, the superposition of messages has to be successfully decoded, and $\mathbb{B}$ must be invertible over $\mathbb{F}_p$, meaning that the system of linear equations at the macro base station has to be of full rank. To reduce the probability of rank deficiency, each MTC device transmits the same message over several frequency slots. For our simulations we consider four slots. We further allow for cooperation between mini base stations and macro base station for up to four colliding devices.

To analyse the end-to-end performance of the approach we consider a system which consists of the following blocks: coding and modulation, resource allocation, transmission over the wireless channel, signal reception and processing at the mini base stations, forwarding to the macro base station, data aggregation at the macro base station.

We assume that channel estimation has been performed and all active devices have been identified [52]. For the slotted transmission synchronization within the guard interval is assumed. All nodes are equipped with a single antenna while all devices transmit at an equal rate. The messages are encoded using an LDPC channel code with a code rate of R=1/4. Each transmitter transmits complex messages of 128 bit, sending a total number of 256 information bit. The encoded data is modulated using a QPSK modulation alphabet. The channel is modelled as a four-tap block fading Rayleigh multipath channel. For the sake of computational complexity we assume that no more than 9 devices collide on the same resource block at a given time. Note that the number of active devices during one time-slot can be much higher. Since each device transmits two independent messages over complex channels, up to 18 messages can collide. The traffic model follows a Poisson arrival process with an arrival rate λ per time-slot.

Both KPIs, protocol throughput and access latency are highly dependent on the signal-to-noise ratio (SNR).

The throughput, shown in Fig. 20, is defined as the mean number of successfully transmitted messages for a certain arrival rate λ. No retransmissions have been considered in the simulations when determining the protocol throughput. Since the macro base station is still able to decode, packets are not discarded when two or more collisions occur, leading to

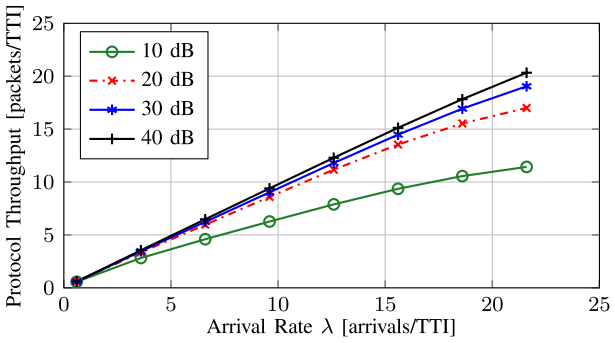

FIGURE 20. Throughput performance of SCF massive access scheme for several SNRs of the first hop.

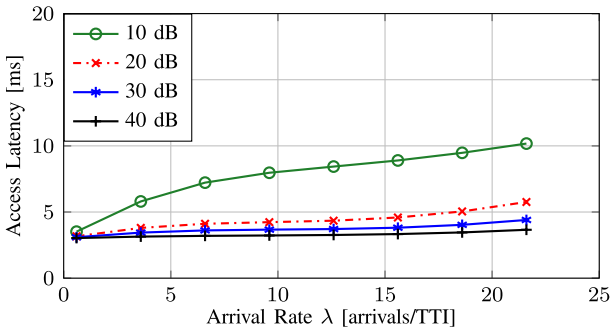

FIGURE 21. Access Latency of SCF massive access scheme for several SNRs of the first hop.

a high throughput even for the lower SNR region. In the considered setup the throughput does not improve significantly for SNR values above 20 dB. In order to determine the access latency, depicted in Fig. 21, we combined the SCF physical layer approach with a random backoff protocol. Transmission is repeated in case of failure until a successful transmission or until the maximum number of retransmissions is reached. If a random access is successful at the first attempt, the expected latency includes the wake-up time and the time to perform a successful random access. If the random access is successful at a later attempt, the access latency includes the latency caused by the unsuccessful attempts prior to the successful transmission, the back-off time between retransmissions and the latency of the last successful random access.

Since users manage to transmit their messages on average in one or two transmissions, the SCF access latency can be kept very low.

E. MASSIVE MIMO

We now consider a massive access solution that takes advantage of the Massive MIMO capabilities, where the base station of a massive MIMO system is equipped with a very large number of antennas and can create a very large number of spatial Degrees of Freedom (DoF) under favorable propagation conditions. Those DoFs are naturally suited to efficiently serve a very large number of devices such as in machine-type communications, not only by spatially multiplexing a dense crowd of devices but also by improving contention resolution in resource access. We target a multiple antenna system at legacy frequency band (below 6 GHz) where the devices are assumed to have a small number of antennas due to their size. The use of a larger number of antennas at the devices is in principle possible at millimeter-wave bands. However, the cost of devices equipped with multiple antennas and beamforming capabilities at those bands is currently a limitation in MTC applications.

This solution addresses two important aspects in machine-type communications: acquisition of Channel State Information (CSI) and data communications for uplink traffic. CSI is estimated at the BS based on training via pilot sequences. The pilot sequences available are assumed to be mutually *orthogonal*. For UL machine-type traffic, CSI estimation suffers from two fundamental limits. First, the duration of pilot sequences is limited by the (time-frequency) coherence interval of the channel, as well as the transmit power of the device. For orthogonal pilot sequences and in crowd scenarios, it means that the number of sequences could be in severe shortage. Therefore, allocation policy of the pilot sequences becomes a central question. Second, the data traffic is intermittent and only a subset of the devices is active simultaneously. Hence a fixed pilot allocation to all the devices in the system would be highly inefficient. Pilot allocation has rather to adapt and scale with the traffic activity pattern and not to the actual number of devices present in the system. A natural choice is to decentralize pilot access to the devices and make it random.

Random access to pilot sequence leads to pilot collision, also known as *pilot contamination*. Pilot contamination is a major impairment in massive MIMO system: when used for data decoding, contaminated channel estimates lead to interference that can be significant. The basic idea of the proposed joint pilot and data access is to randomize the effect of pilot contamination over multiple transmission slots, so that the the effect of contamination-induced interference is averaged out and becomes predictable. Related work can be found [53]–[55].

Uplink transmission is organized into transmission frames made out of multiple transmission slots. A transmission slot is a time-frequency unit where the channel can be approximated as constant. Fig. 22 depicts a simplified example with four active devices and two orthogonal pilot sequences, where τ_u is the duration of a transmission slot and τ_p is the duration of the pilot sequences in symbols. A block fading model is assumed, with independent realization in each slot and for each device.

A device with data to transmit waits for the start of a new transmission frame. Each active device encodes its data into one codeword that is divided into multiple parts and transmits one pilot sequence followed by one part of the codeword within a transmission slot. The pilot sequence serves to estimate the channel that is then used for soft decoding of the associated codeword portion. Within a transmission frame, a number of K_a devices are active out of a total number of K devices. The activation probability of a device is p_a.

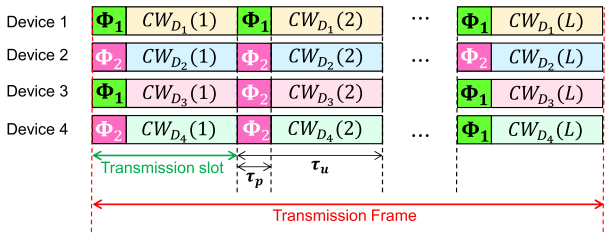

FIGURE 22. Illustration of the transmission frame with four active devices $\{D_1, D_2, D_3, D_4\}$ and two mutually orthogonal pilot sequences $\{\Phi_1, \Phi_2\}$.

In order to randomize the effect of pilot contamination, *pilot hopping* is performed. In each transmission slot, each active device selects one pilot sequence from the set of orthogonal pilot sequences according to a pseudo-random pilot-hopping pattern that is unique to the device. Hence, in each transmission slot and for one given device, contamination-induced interference comes from different sets of devices. The codeword of the device experiences all possible contamination events from the K_a active devices, provided that the number of transmission slots duration is sufficiently long. Likewise, for an asymptotic large number of transmission slots, the additive noise at the BS and fading is averaged out. Under those asymptotic conditions, a maximal achievable rate per device can be defined within each transmission frame. Achieving this rate assumes the following features: a) estimation of the number of active users at the BS, b) estimation of the average channel energy per device at the BS and at the device, c) BS broadcasts the rate associated to each value of average channel energy. With conditions a) and b), the BS computes a maximal achievable rate per device. The BS broadcasts both the channel energy and its associate rate for each active device. As the device itself knows its channel energy, it can associate its assigned rate.

For each transmission slot, the following steps are performed. First, the BS detects which pilot sequences are in use. This is done by correlating the received signal with each sequence available. The pilot detection outcomes are buffered in order to be utilized for device activity detection. Second, for each pilot sequence detected, the corresponding channel estimate is computed. In this work, MMSE channel estimation is performed. When there is pilot collision, channel estimation is contaminated. Third, for each pilot sequence detected, a multiple antenna processing based on the channel estimate is applied to the data symbols in the slot and its output is buffered along with its associated pilot index. In this work, Maximum Ratio Combining (MRC) is utilized.

A unique pseudorandom pilot-hopping pattern is assigned to each device. The pilot-hopping patterns are known at the BS and serve for device identification at the BS. In order to detect the transmitting devices, the BS combines the pilot sequence detection outcomes from the slots that follow the pattern. Based on the identifying pilot-hopping patterns, the BS identifies which MRC outputs to combine to decode the data of each transmitting device.

Our main performance metric is the system uplink sum rate. It is the sum rate per transmission frame averaged over the activation statistics of the device population. We work on an approximation of the uplink sum rate $\mathcal{R}$ that is tight thanks to channel hardening and when the total number of devices is large. This metric depends on the total number of BS antennas M and the number of pilot sequences, τ_p: the larger those quantities, the more devices can be multiplexed. Bound $\mathcal{R}$ is also a function of the device activation probability, p_a. To maximize the sum rate, one can optimize p_a and τ_p. When the number of antennas M and the duration of transmission slot τ_u are of the same order, the sum rate scales of $\sqrt{M\tau_u}$. Heuristic solutions indicate that one third of the transmission slot should be devoted to training while the average number of active devices should be of the order of $\sqrt{M\tau_u}$.

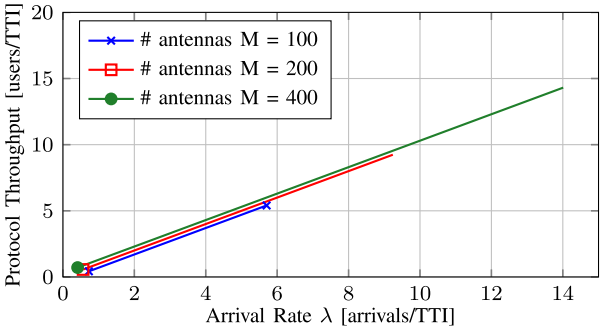

FIGURE 23. Protocol throughout as a function of the arrival rate for $K = 400$ and for $M = 100, 200, 400$.

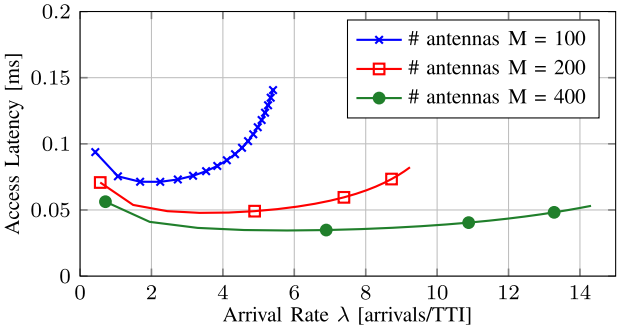

FIGURE 24. Access latency (ms) as a function of the arrival rate for $K = 400$ and for $M = 100, 200, 400$.

Fig. 23 and Fig. 24 show the performance metrics for a scenario with K = 400 and M = 100, 200, 400 for an SNR of 10 dB. The transmission slot duration is fixed to $\tau_u = 300$ and $\tau_p = 100$: this ratio is chosen as it leads to a near-optimal solution (see above). We compute the average sum rate per device from which we determine the average delay to transmit 8 bytes per device over a bandwidth of 1 MHz. This study relies on an information theory framework, where the devices are guaranteed to transmit their data reliably. Therefore, the average number of active devices that have successfully transmitted is also the average number of active users in the TTI. The performance metrics are plotted against the arrival rate per TTI.

VII. CONCLUSIONS

The shown performance results show that each solution provides a trade-off between throughput and latency. Yet we can conclude that very significant gains can be achieved if the following techniques are applied in the design of massive access protocols:

- Physical layer: (i) compressive sensing for multi-user detection (CSMUD, CRAPLNC), (ii) multi-user decoding (OSTSAP, CRAPLNC), (iii) redesign of access preambles (OSTSAP) and (iv) multiple spatial layers (NOTAFT);
- Medium access layer: (v) coding over retransmissions (SBA, CRAPLNC), (vi) back-off schemes (OSTSAP, CRAPLNC);
- Protocol Design: (vii) one-stage protocols (CSMUD) and (viii) low overhead network synchronization (NOTAFT).

One final remark is that for all the schemes a large part of the complexity is at the receiver of the base station, while the transmitter operation at the devices does not suffer an increase in complexity (with the exception of the CRAPLNC scheme).

For massive Machine Type Communications to take place, there is the need for efficient access protocols capable of withstanding a massive number of devices contending for network access. We have proposed several random-access schemes of one-stage and two-stages types. Several physical layer and medium access layer techniques have been considered. The physical layer techniques include multi-user detection using compressive sensing techniques, collision resolution and harness of interference using physical layer network coding and non-orthogonal access with relaxed time-alignment. The medium access layer techniques include coded random access and signature based access, one/two-stage random access and fast uplink access protocols with a focus on latency reduction. A common evaluation framework has been defined and individual performance results provided. These results will help on the design of a robust massive access solutions, by identifying which techniques lead to higher protocol performance, and doing so provide recommendations on the protocol design for the NR in 3GPP.

Acknowledgement

The authors would like to acknowledge the contributions of their colleagues in FANTASTIC-5G.

Emil Björnson, Jesper H. Sørensen and Erik G. Larsson are contributors of the work presented in section V.C.

REFERENCES

[1] (2015). *SigFox—Global Cellular Connectivity for the Internet of Things*. [Online]. Available: http://www.sigfox.com/en/

[2] (2015). *LoRa Alliance—Wide Area Networks for the Internet of Things*. [Online]. Available: https://www.lora-alliance.org/

[3] 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, 1st Quart., 2014.

[4] *Study on Enhancements to Machine-Type Communications (MTC) and Other Mobile Data Applications; Radio Access Network (RAN) Aspects*, document TR37.869, 3GPP, 2013.

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

[6] *Further LTE Physical Layer Enhancements for MTC*, document RP-151186, 3GPP, 2015.

[7] *Narrowband IOT*, document RP-151621, 3GPP, 2015.

[8] *Technical Specification Group Radio Access Network; General Packet Radio Service (GPRS); Overall Description of the GPRS Radio Interface*, document TS 43.064, 3GPP, 2016.

[9] *Service Requirements for Machine-Type Communications (MTC)*, dcument TR22.368, 3GPP, 2014.

[10] F. Schaich *et al.*, "FANTASTIC-5G: Flexible air interface for scalable service delivery within wireless communication networks of the 5th generation," *Trans. Emerg. Telecommun. Technol.*, vol. 27, no. 9, pp. 1216–1224, Jun. 2016.

[11] *FANTASTIC-5G Website*. Accessed: May 16, 2018. [Online]. Available: www.fantastic5g.eusss

[12] *Preliminary Results for Multi-Service Support in Link Solution Adaptation*, document D3.1, FANTASTIC-5G, 2016.

[13] *Final Report on the Holistic Link Solution Adaptation*, document D3.2, FANTASTIC-5G, 2016.

[14] *E-UTRA Base Station (BS) Radio Transmission and Reception*, document TS 36.104, 3GPP, 2017.

[15] M. Centenaro, L. Vangelista, A. Zanella, and M. Zorzi, "Long-range communications in unlicensed bands: The rising stars in the IoT and smart city scenarios," *IEEE Wireless Commun.*, vol. 23, no. 5, pp. 60–67, Oct. 2016.

[16] G. C. Madueño, Č. Stefanović, and P. Popovski, "Reliable and efficient access for alarm-initiated and regular M2M traffic in ieee 802.11ah systems," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 673–682, Oct. 2016.

[17] *Further LTE Physical Layer Enhancements for MTC*, document RAN1#45, RP-151186, 3GPP, May 2015.

[18] M. Lauridsen, I. Z. Kovacs, P. Mogensen, M. Sorensen, and S. Holst, "Coverage and capacity analysis of LTE-M and NB-IoT in a rural area," in *Proc. IEEE 84th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2016, pp. 1–5.

[19] F. Schaich *et al.*, "FANTASTIC-5G: flexible air interface for scalable service delivery within wireless communication networks of the 5th generation," *Trans. Emerg. Telecommun. Technol.*, vol. 27, no. 9, pp. 1216–1224, 2016. [Online]. Available: http://dx.doi.org/10.1002/ett.3050

[20] A. Zanella and M. Zorzi, "Theoretical analysis of the capture probability in wireless systems with multiple packet reception capabilities," *IEEE Trans. Commun.*, vol. 60, no. 4, pp. 1058–1071, Apr. 2012.

[21] *Final Results for Service Specific Multinode/Multi-Antenna Solutions*, document D4.2, FANTASTIC-5G, 2016.

[22] S. Saur and M. Centenaro, "Radio access protocols with multi-user detection for URLLC in 5Gs," in *Proc. Eur. Wireless*, May 2017, pp. 1–6.

[23] S. Saur, A. Weber, and G. Schreiber, "Radio access protocols and preamble design for machine type communications in 5G," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2015, pp. 8–12.

[24] H. Thomsen, N. K. Pratas, Č. Stefanović, and P. Popovski, "Code-expanded radio access protocol for machine-to-machine communications," *Trans. Emerg. Telecommun. Technol.*, vol. 24, no. 4, pp. 355–365, May 2013.

[25] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Commun. ACM*, vol. 13, no. 7, pp. 422–426, Jul. 1970.

[26] *Base Station (BS) Conformance Testing*, document TS36.141, 3GPP, 2016.

[27] N. Pratas, S. Pattathil, C. Stefanovic, and P. Popovski, *Massive Machine-Type Communication (mMTC) Access With Integrated Authentication*. Piscataway, NJ, USA: IEEE Press, 2017.

[28] N. Pratas, C. Stefanovic, G. Madueno, and P. Popovski, "Random access for machine-type communication based on bloom filtering," in *Proc. Global Commun. Conf. (GLOBECOM)*, Washington, DC, USA, 2016.

[29] Z. Zhao, M. Schellmann, X. Gong, Q. Wang, R. Böhnke, and Y. Guo, "Pulse shaped OFDM for 5G systems," *EURASIP J. Wireless Commun. Netw.*, vol. 2017, no. 1, Apr. 2017, Art. no. 74.

[30] Y. Guo, Z. Zhao, and R. Böhnke, "A method for constructing localized pulse shapes under length constraints for multicarrier modulation," in *Proc. IEEE 83rd Veh. Technol. Conf. (VTC Spring)*, May 2016, pp. 1–5.

[31] H. Zhu and G. B. Giannakis, "Exploiting sparse user activity in multiuser detection," *IEEE Trans. Commun.*, vol. 59, no. 2, pp. 454–465, Feb. 2011.

[32] E. J. Candès, "Compressive sampling," in *Proc. Int. Congr. Math.*, 2006, pp. 1433–1452. [Online]. Available: http://www.dsp.ece.rice.edu/cs/

[33] F. Monsees, M. Woltering, C. Bockelmann, and A. Dekorsy, "A potential solution for MTC: Multi-carrier compressed sensing multi-user detection," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2015, pp. 18–22. [Online]. Available: http://www.asilomarsscconf.org/

[34] H. F. Schepker, C. Bockelmann, and A. Dekorsy, "Exploiting sparsity in channel and data estimation for sporadic multi-user communication," in *Proc. 10th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Ilmenau, Germany, Aug. 2013, pp. 1–5. [Online]. Available: http://conference.vde.com/iswcs2013/

[35] Y. Ji, C. Bockelmann, and A. Dekorsy, "Numerical analysis for joint PHY and MAC perspective of compressive sensing multi-user detection with coded random access," in *Proc. IEEE Int. Conf. Commun. (ICC) 3rd Int. Workshop 5G RAN Des.*, Paris, France, May 2017, pp. 1018–1023. [Online]. Available: http://icc2017.ieee-icc.org/

[36] E. Paolini, C. Stefanovic, G. Liva, and P. Popovski, "Coded random access: Applying codes on graphs to design random access protocols," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 144–150, Jun. 2015.

[37] G. Cocco and S. Pfletschinger, "Seek and decode: Random multiple access with multiuser detection and physical-layer network coding," in *Proc. IEEE Int. Conf. Commun.-Workshop*, Jun. 2014, pp. 501–506.

[38] S. Pfletschinger, M. Navarro, and G. Cocco, "Interference cancellation and joint decoding for collision resolution in slotted ALOHA," in *Proc. IEEE Int. Symp. Netw. Coding*, Jun. 2014, pp. 1–6.

[39] B. Nazer and M. Gastpar, "Compute-and-forward: Harnessing interference through structured codes," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6463–6486, Oct. 2011.

[40] J. Goseling, M. Gastpar, and J. H. Weber, "Random access with physical-layer network coding," *IEEE Trans. Inf. Theory*, vol. 61, no. 7, pp. 3670–3681, Jul. 2015.

[41] C. Bockelmann, H. F. Schepker, and A. Dekorsy, "Compressive sensing based multi-user detection for machine-to-machine communication," *Trans. Emerg. Telecommun. Technol.*, vol. 24, no. 4, pp. 389–400, 2013.

[42] Y. Ji, Č. Stefanović, C. Bockelmann, A. Dekorsy, and P. Popovski, "Characterization of coded random access with compressive sensing based multi-user detection," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Austin, TX, USA, Dec. 2014, pp. 1740–1745. [Online]. Available: http://www.arxiv.com/1404.2119

[43] G. Wunder, H. Boche, T. Strohmer, and P. Jung, "Sparse signal processing concepts for efficient 5G system design," *IEEE ACCESS*, vol. 3, pp. 195–208, Dec. 2015. [Online]. Available: http://arxiv.org/abs/1411.0435

[44] G. Wunder, P. Jung, and C. Wang, "Compressive random access for post-LTE systems," in *Proc. IEEE Int. Conf. Commun. (ICC) Workshop*, Sydney, NSW, Australia, May 2014, pp. 539–544.

[45] G. Wunder, P. Jung, and M. Ramadan, "Compressive random access using a common overloaded control channel," in *Proc. IEEE Global Commun. Conf. (GLOBECOM) Workshop*, San Diego, CA, USA, Dec. 2015, pp. 1–6. [Online]. Available: http://arxiv.org/abs/1504.05318

[46] G. Wunder, C. Stefanovic, P. Popovski, and L. Thiele, "Compressive coded random access for massive MTC traffic in 5G systems," in *Proc. 49th Annu. Asilomar Conf. Signals, Syst.*, Pacific Grove, CA, USA, Nov. 2015, pp. 13–17.

[47] I. Roth, M. Kliesch, J. Eisert, and G. Wunder. (Dec. 2016). "Reliable recovery of hierarchically sparse signals and application in machine-type communications," [Online]. Avilable: https://arxiv.org/abs/1612.07806

[48] G. Wunder *et al.*, "5GNOW: Non-orthogonal, asynchronous waveforms for future mobile applications," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 97–105, Feb. 2014.

[49] B. Nazer and M. Gastpar, "Reliable physical layer network coding," *Proc. IEEE*, vol. 99, no. 3, pp. 438–460, Mar. 2011.

[50] M. Goldenbaum and S. Stańczak, "Robust analog function computation via wireless multiple-access channels," *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 3863–3877, Sep. 2013.

[51] M. Raceala-Motoc, J. Schreck, P. Jung, and S. Stanczak, "Robust message recovery for non-cooperative compute-and-forward relaying," in *Proc. 50th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2016, pp. 1303–1307. [Online]. Available: http://dx.doi.org/10.1109/ACSSC.2016.7869585

[52] M. Goldenbaum, P. Jung, M. Raceala-Motoc, J. Schreck, S. Stańczak, and C. Zhou, "Harnessing channel collisions for efficient massive access in 5G networks: A step forward to practical implementation," in *Proc. 9th Int. Symp. Turbo Codes Iterative Inf. Process. (ISTC)*, Sep. 2016, pp. 335–339.

[53] E. de Carvalho, E. Björnson, E. G. Larsson, and P. Popovski, "Random access for massive MIMO systems with intra-cell pilot contamination," in *Proc. IEEE ICASSP*, Mar. 2016, pp. 3361–3365.

[54] J. H. Sørensen, E. de Carvalho, and C. Stefanović, and P. Popovski, "Coded pilot access: A random access solution for massive MIMO systems," *IEEE Trans. Wireless Commun.*, to be published. [Online]. Available: https://arxiv.org/abs/1605.05862

[55] E. Björnson, E. de Carvalho, J. H. Sørensen, E. G. Larsson, and P. Popovski, "A random access protocol for pilot allocation in crowded massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2220–2234, Apr. 2016. [Online]. Available: http://arxiv.org/pdf/1604.04248.

CARSTEN BOCKELMANN (S'09–M'12) received the Dipl.-Ing. and Ph.D. degrees in electrical engineering from the University of Bremen, Germany, in 2006 and 2012, respectively. Since 2012. He has been a Senior Research Group Leader with the University of Bremen coordinating research activities regarding the application of compressive sensing/sampling to communication problems. His current research interests include communications in massive machine communication (5G) and industry 4.0, compressive sensing, channel coding, and transceiver design.

NUNO K. PRATAS received the Ph.D. degree in wireless communications from Aalborg University, Copenhagen, Denmark, in 2012. He was an Assistant Professor in wireless communications with the Department of Electronic Systems, Aalborg University. Since 2017, he has been with Intel, Aalborg, Denmark. His current research interests are on wireless communications, and networks and development of analysis tools for machine-to-machine and device-to-device applications. He was a recipient of the best student conference paper award twice in 2010 and has been recognized as an Exemplary Reviewer of the IEEE TRANSACTION ON COMMUNICATIONS.

GERHARD WUNDER (M'04–SM'11) received the Dipl.-Ing. degree (Hons.) in electrical engineering and the Ph.D. degree (Dr.-Ing.) *(summa cum laude)* Technical University of Berlin in 1999 and 2003, respectively, and the Habilitation degree (venia legendi) in 2007. He became a Research Group Leader at the Fraunhofer Heinrich-Hertz-Institut, Berlin. In 2007, he became a Privatdozent (Associate Professor) and he was a Visiting Professor with the Georgia Institute of Technology (Prof. Jayant), Atlanta, GA, USA, and Stanford University (Prof. Paulraj), Palo Alto, CA, USA. In 2009, he was a Consultant with Alcatel-Lucent Bell Labs (Prof. Stolyar), Murray Hill, NJ, USA, and at Alcatel-Lucent Bell Labs (Dr. Valenzuela), Crawford Hill, NJ, USA. In 2015, he became a Heisenberg Fellow and granted for the first time to a Communication Engineer. He is currently the Head of the Heisenberg Communications and Information Theory (Heisenberg CIT Group), Free University of Berlin. He has been a Coordinator and a Principal Investigator in the FP7 Project 5GNOW on 5G new waveforms (received Outstanding Excellence from EC) and PROPHYLAXE on IoT Physical Layer Security funded by German BMBF. He has been a member of the project management teams of H2020 projects FANTASTIC-5G (also received Outstanding Excellence) and ONE5G, both regarded as flagship projects within the European 5GPPP framework. He is a member of the IEEE TWC's Executive Editorial Committee. In 2011, he was a recipient of the 2011 Award for Outstanding Scientific Publication in the field of communication engineering at the German Communication Engineering Society (ITG Award 2011). He has co-chaired numerous international renowned workshops, conference tracks, special issues, particularly in the context of 5G. He was the Co-Chair of the IEEE GLOBECOM 2017 Signal Processing for Communications Symposium. He has been nominated together with Dr. Müller (BOSCH Stuttgart) and Prof. Paar (Ruhr University Bochum) for the Deutscher Zukunftspreis 2017 on behalf of the PROPHYLAXE Project.

STEPHAN SAUR (M'08) received the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering from the University of Stuttgart, Germany, in 2000 and 2008, respectively. He joined Alcatel Research and Innovation in 2006, where he is currently a Member of Technical Staff with Nokia Bell Labs Wireless Research, where he is entrusted with the development of novel PHY, MAC, and system concepts for future cellular mobile radio systems. His current research topics are efficient radio access for sporadic low-rate traffic and tracking algorithms for enhanced road safety.

MÒNICA NAVARRO (S'96–M'98–SM'08) received the M.Sc. degree in telecommunications engineering from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 1997 and the Ph.D. degree in telecommunications from the Institute for Telecommunications Research, University of South Australia, in 2002. From 1997 to 1998, she was a Research Assistant with the Department of Signal Theory and Communications, UPC, where she was involved in the development of fractal shape multiband antennas for wireless cellular communications systems. She has also been a part-time Lecturer with the Universitat Pompeu Fabra, Barcelona. She is currently the Head of the Communication Systems Division and a Senior Researcher with the Communication Systems Division, Centre Tecnològic de Telecomunicacions de Catalunya. She has strong expertise on new radio air interfaces for cellular networks. Over the last 10 years, she has lead projects funded by the European Commission, Spanish and Catalan Governments, and the European Space Agency. Her primary areas of interest are on information processing with applications to wireless communications and positioning, particularly on iterative information processing, adaptive transmissions, coding techniques, and random access protocols for massive access. She served at the Editorial Board of Emerging Telecommunications Technologies from 2013 to 2016.

DAVID GREGORATTI (S'02–M'10–SM'15) received the M.Sc. degree in telecommunications engineering from the Politecnico di Torino, Italy, in 2005, and the Ph.D. degree in signal theory and communications from the Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), Universitat Politécnica de Catalunya (UPC), Barcelona, Spain, in 2010. Since 2006, he has been with CTTC, where he is currently a Research Associate with the Advanced Signal and Information Processing Department. During his academic career, he also visited the Institut Eurecom, Sophia Antipolis, France, from 2003 to 2004, Qualcomm Inc., San Diego, CA, USA, in 2004, the Telecom Italia Lab, Torino, Italy, in 2005, and the Télécom ParisTech, Paris, France, in 2009. He has been actively participating in diverse research projects (with both public and private funding) and in the organization of the IEEE-sponsored events. His current research interests cover wireless communications, multicarrier modulations, and optimization with sparsity.

GUILLAUME VIVIER received the M.Sc. degree in telecommunication engineering from Télécom ParisTech, Paris, France, in 1993, and the Ph.D. degree from the University of Pierre and Marie Curie, Paris, in 2003. After various positions at Alcatel and Motorola, he joined Sequans Communications in 2008 to drive innovation into products, where he is currently the Director of advanced technology and the Head of CTO office. He coordinates standardization activities and Sequans research collaborations in European and national funded programs. He initiated 5G activity to prepare future generation of Sequans' chipsets.

ELISABETH DE CARVALHO received the Ph.D. degree in electrical engineering from Télécom ParisTech, France, in 1999. After her Ph.D., she was a Post-Doctoral Fellow with Stanford University, USA, and then worked in industry in the field of DSL and wireless LAN. Since 2005, she has been an Associate Professor with Aalborg University, Copenhagen, Denmark, where she has led several research projects in wireless communications. She has co-authored the text book *A practical guide to the MIMO radio channel*. Her main expertise is in signal processing for MIMO communications with recent focus on massive MIMO including channel measurements, channel modeling, beamforming, and protocol aspects.

YALEI JI received the bachelor's degree in communications engineering from Shandong University, Jinan, China, in 2010, and the M.Sc. degree in communication and information technology from the University of Bremen, Germany, in 2013, where he is currently pursuing the Ph.D. degree with the Department of Communications Engineering. He joined the Department of Communications Engineering, University of Bremen, in 2013. His research interests are compressive sensing in wireless communications, especially the joint MAC- and PHY-layer perspective of compressive sensing multi-user detection technique in wireless communication systems.

ČEDOMIR STEFANOVIĆ received the Dipl.-Ing., Mr.-Ing., and Dr.-Ing. degrees in electrical engineering from the University of Novi Sad, Serbia, in 2001, 2006, and 2011, respectively. He is currently an Associate Professor with the Department of Electronic Systems, Aalborg University, Denmark. In 2014, he received an individual postdoctoral grant by the Danish Council for Independent Research. He has been involved in a number of national and EU projects on IoT and 5G communications. His research interests include communication theory and wireless and smartgrid communications.

PETAR POPOVSKI (S'97–A'98–M'04–SM'10–F'16) received the Dipl.-Ing. degree in electrical engineering and the M.-Ing. degree in communication engineering from the Saints Cyril and Methodius University of Skopje, Skopje, Macedonia, in 1997 and 2000, respectively, and the Ph.D. degree from Aalborg University, Denmark, in 2004. He is currently a Professor in wireless communications with Aalborg University. He has over 300 publications in journals, conference proceedings, and edited books. He holds over 30 patents and patent applications. His research interests are in the area of wireless communication and networking and communication/information theory. He is currently a Steering Committee Member of the IEEE SmartGridComm and previously served as a Steering Committee Member for the IEEE Internet of Things Journal. He is a holder of a Consolidator Grant from the European Research Council. He was a recipient of the Danish Elite Researcher Award and a member of the Danish Academy for Technical Sciences. He is currently an Area Editor of the IEEE Transactions on Wireless Communications.

QI WANG received the B.Eng. degree from the Beijing University of Posts and Telecommunications, China, in 2005, the M.Sc. degree from Linköping University, Sweden, in 2007, and the Dr.techn. degree from the Vienna University of Technology, Austria, in 2012. She has been a Senior Research Engineer with the Huawei Munich Research Center, Germany, since 2014. She has been involved in several European research projects such as Fantastic-5G and 5GCar. Her research interests include radio access technologies and signal processing aspects such as waveform design, channel estimation, and localization techniques.

MALTE SCHELLMANN received the Dipl.-Ing. (M.S.) degree in electrical engineering and information technology from the Technische Universität München, Germany, in 2003, and the Dr.-Ing. (Ph.D.) degree from the Technische Universität Berlin, Germany, in 2009. From 2004 to 2009, he was with the Fraunhofer Heinrich Hertz Institute, Berlin. He is currently a Principal Research Engineer with the Huawei Munich Research Center, Germany. His research addresses radio access technologies for 5G with particular focus on air interface design for vehicular communications. He has been actively involved in several European research projects, among those the 4G flagship projects WINNER, WINNER II, WINNER+, and the 5G flagship projects METIS, where he was leading the work package on radio link concepts, METIS II and FANTASTIC-5G.

EVANGELOS KOSMATOS (M'15) received the Dipl.-Ing. and Ph.D. degrees from the School of Electrical and Computer Engineering, National Technical University of Athens, Greece, in 2002 and 2008, respectively. He has participated in several EU projects (IST AQUILA, ENAMORADO, SMS, INCASA, STRONGEST, IDEALIST, FANTASTIC-5G, SPEED-5G, ONE5G, and Clear5G) and national projects (CONFES and WisePON). His research interests include 4G, 5G networks, optical networks (EPON, GPON, and WDM-PON), network control and management, including SDN and NFV, radio-over-fiber networks, protocols and algorithms supporting QoS and QoE in both wireless and optical networks, sensors/actuators networks (Internet of Things), and middleware and distributed technologies. He has several publications in international journals refereed conferences related to these fields. He is a member of the Technical Chamber of Greece.

PANAGIOTIS DEMESTICHAS (M'95–SM'12) received the Diploma and Ph.D. degrees in electrical engineering from the National Technical University of Athens, Greece, in 1993 and 1996, respectively. Since 2012, he has been a Full Professor with the Department of Digital Systems, University of Pireaus, Greece, where he served as the Head from 2011 to 2015. From 2015 to 2016, he was on Sabbatical, collaborating with the University of Surrey and, in particular, the 5G Innovation Center. Since 2015, he has contributed to the development of systems for the SMEs WINGS ICT Solutions, including Artificial Intelligence-powered solutions for vertical industries, and Incelligent (proactive network and customer experience management for the telecommunications and various fintech-related sectors). He has over 25 years of experience in research and development in the fields of wireless/mobile broadband networks, network planning and management, smart utilities, smart cities, and environment management. Recent interests include 5G aspects, and especially, the exploitation of spectrum beyond 6 GHz, overall spectrum management, 5G architectures, artificial-intelligence-based and predictive management, and virtualization technologies based on SDN and NFV. He has several publications in these areas in international journals and refereed conferences. At the European level, he has been actively involved in, and coordinated as a project manager, a deputy project manager, and a technical manager, a number of international research and development programs. He also co-organized the European Conference on Networks and Communications, Athens, Greece, in 2016, and had as a general theme The Dawn of 5G. In terms of standardization, he has contributed to various standardization bodies, such as ETSI and IEEE. He is a member of the IEEE Working Group on 5G publications. He is a member of the ACM and the Technical Chamber of Greece.

MIRUNA RACEALA-MOTOC received the Dipl.-Ing. (M.S.) degree in industrial engineering from the Technical University of Berlin (TU Berlin), Germany, in 2013, where she is currently pursuing the Dr.-Ing. (Ph.D.) degree in electrical engineering. She studied at Politehnica University, Bucharest, Romania, and at TU Berlin. Since 2011, she has been with the Fraunhofer Heinrich Hertz Institute. Her research interests include digital communications, signal processing, and wireless multiuser communications with focus on adaptive transmission and coding techniques.

PETER JUNG (M'05) received the Dipl.-Phys. in high energy physics from Humboldt University, Berlin, Germany, in 2000, in cooperation with DESY Hamburg, and the Dr. rer.nat (Ph.D.) degree on Weyl–Heisenberg representations in communication theory from the Technical University of Berlin (TUB), Germany, in 2007. Since 2001, he has been with the Department of Broadband Mobile Communication Networks, Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut, and since 2004, he has been with the Fraunhofer German-Sino Laboratory for Mobile Communications. He is currently working under DFG grants at TUB, where he is involved in the field signal processing and information and communication theory. His current research interests are in the area compressed sensing, machine learning, time–frequency analysis, dimension reduction, and randomized algorithms. He is giving lectures in compressed sensing, estimation theory, and inverse problems. He is a member of VDE/ITG.

SLAWOMIR STANCZAK studied electrical engineering with specialization in control theory from the Wroclaw University of Technology and from the Technical University of Berlin (TU Berlin). He received the Dipl.-Ing. and Dr.-Ing. degrees *(summa cum laude)* in electrical engineering from TU Berlin in 1998 and 2003, respectively, and the Habilitation degree (venia legendi) in 2006. Since 2015, he has been a Full Professor for network information theory with TU Berlin and the Head of the Wireless Communications and Networks Department. He has been involved in research and development activities in wireless communications since 1997. In 2004 and 2007, he was a Visiting Professor with RWTH Aachen University, Germany. In 2008, he was a Visiting Scientist with Stanford University, Stanford, CA, USA. He has co-authored two books and over 200 peer-reviewed journal articles and conference papers in the area of information theory, wireless communications, signal processing, and machine learning. He received research fellowships from the German Research Foundation and the Best Paper Award from the German Communication Engineering Society in 2014. He was a Co-Chair of the 14th International Workshop on Signal Processing Advances in Wireless Communications in 2013. From 2009 to 2011, he was an Associate Editor of the *European Transactions for Telecommunications* (information theory) and an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2012 to 2015.

ARMIN DEKORSY (SM'18) received the Dipl.-Ing. degree from Fachhochschule Konstanz, Germany, in 1992, the Dipl.-Ing. degree from the University of Paderborn, Germany, in 1996, and the Ph.D. degree from the University of Bremen, Germany, in 2000, all in communications engineering. He is currently the Head of the Department of Communications Engineering, University of Bremen. He is distinguished by his over 10 years of industrial experience in leading research positions at Deutsche Telekom, Alcatel-Lucent (Bell Labs), and Qualcomm successfully conducting (inter)national research projects (18 BMBF/BMWI/EU projects). He published over 160 journal and conference publications and holds over 19 patents in the area of wireless communications. He investigates new lines of research in wireless communications and signal processing for transmitter baseband design which can readily be transferred to industry. His current research directions are cooperative and distributed communications, compressive sensing, and in-network processing. He is a Vice Chairman of the VDE/ITG expert committee Information and System Theory and represents the ETSI, University of Bremen, and NetWorld2020 ETP. He acted as an Editor of the IEEE COMMUNICATIONS LETTER.