Compressive-Sensing-Based Grant-Free Massive Access for 6G Massive Communication

Zhen Gao, Malong Ke, *Member, IEEE*, Yikun Mei, Li Qiao, *Graduate Student Member, IEEE*, Sheng Chen, *Life Fellow, IEEE*, Derrick Wing Kwan Ng, *Fellow, IEEE*, and H. Vincent Poor, *Life Fellow, IEEE*

***Abstract*—The envisioned sixth-generation (6G) of wireless communications is expected to give rise to the necessity of connecting very large quantities of heterogeneous wireless devices, which requires advanced system capabilities far beyond existing network architectures. In particular, such massive communication has been recognized as a prime driver that can empower the 6G vision of future ubiquitous connectivity, supporting Internet of Human–Machine-Things (IoHMT) for which massive access is critical. This article surveys the most recent advances toward massive access in both academic and industrial communities, focusing primarily on the promising compressive sensing (CS)-based grant-free massive access (GFMA) paradigm. We first specify the limitations of existing random access schemes and reveal that the practical implementation of massive communication relies on a dramatically different random access paradigm from the current ones mainly designed for human-centric communications. Then, a CS-based GFMA roadmap is presented, where the evolutions from single-antenna to large-scale antenna array-based base stations, from single-station to cooperative massive multiple-input–multiple-output (MIMO) systems, and from unsourced to sourced random access scenarios are detailed. Finally, we discuss key challenges and open issues to indicate potential future research directions in GFMA.**

***Index Terms*—compressive sensing (CS), grant-free massive access (GFMA), Internet of Human–Machine-Things (IoHMT), Internet of Things (IoT), massive communication, sixth generation (6G).**

Manuscript received 3 September 2023; revised 4 November 2023; accepted 8 November 2023. Date of publication 28 November 2023; date of current version 21 February 2024. The work of Zhen Gao was supported in part by the National Natural Science Foundation of China (NSFC) under Grant U2233216 and Grant 62071044; in part by the Shandong Province Natural Science Foundation under Grant ZR2022YQ62; and in part by the Beijing Nova Program. The work of Derrick Wing Kwan Ng was supported by the Australian Research Council's Discovery Projects under Grant DP210102169 and Grant DP230100603. The work of H. Vincent Poor was supported by the U.S. National Science Foundation under Grant CNS-2128448 and Grant ECCS-2335876. *(Corresponding author: Malong Ke.)*

Zhen Gao is with the MIIT Key Laboratory of Complex-Field Intelligent Sensing, Beijing Institute of Technology, Beijing 100081, China, also with the Yangtze Delta Region Academy, Beijing Institute of Technology (Jiaxing), Jiaxing 314019, China, and also with the Advanced Technology Research Institute, Beijing Institute of Technology, Jinan 250307, China (e-mail: gaozhen16@bit.edu.cn).

Malong Ke is with the Wireless Product Division, Ruijie Network Company Ltd., Fuzhou 350108, China (e-mail: kemalong@ruijie.com.cn).

Yikun Mei is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: meiyikun@bit.edu.cn).

Li Qiao is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China, also with 5GIC & 6GIC, Institute for Communication Systems, University of Surrey, GU2 7XH Guildford, U.K. (e-mail: qiaoli@bit.edu.cn).

Sheng Chen is with the School of Electronics and Computer Science, University of Southampton, SO17 1BJ Southampton, U.K. (e-mail: sqc@ecs.soton.ac.uk).

Derrick Wing Kwan Ng is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: w.k.ng@unsw.edu.au).

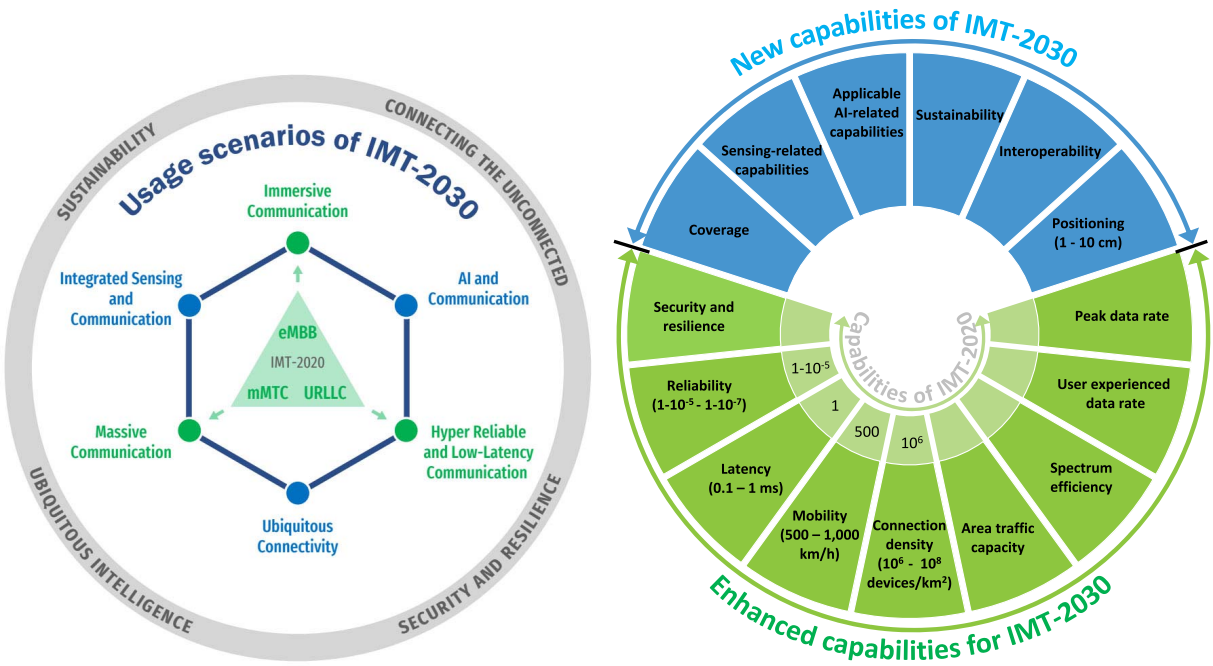
H. Vincent Poor is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08542 USA (e-mail: poor@princeton.edu).

Digital Object Identifier 10.1109/JIOT.2023.3334878

I. Introduction

IN THE future sixth-generation (6G) mobile network vision, the concept of Internet of Things (IoT) is gradually evolving into the Internet of Human–Machine-Things (IoHMT) paradigm, where the interactions across humans, machines, and things are intricately interconnected to create an intelligent ecosystem [1], [2], [3], [4]. In the upcoming IoHMT era, the ubiquitous connectivity of heterogeneous devices is expected to enable a plethora of promising applications, such as smart cities and smart factories, promoting the digitalization of society, and improving the overall efficiency of various vertical sectors [5]. It is predicted that there will be up to 75 billion devices connected in IoHMT ecosystems by 2025, which will lead to significant economic returns of about 11.1 trillion United States (U.S.) dollars each year [6].

The success of IoHMT ecosystems relies on *massive communication*, which is one of the novel usage scenarios of the 6G vision, as shown in Fig. 1(a). Massive communication is evolved from the massive machine-type communications (mMTC) of the fifth-generation (F-G) network, allowing ultra-massive numbers of machine-type devices to exchange their information either with central/distributed servers or with other devices. Due to the highly heterogeneous nature of IoHMT applications, massive communication can be rather different from the conventional human-centric communications that are well supported in the fourth-generation (4G) and 5G cellular networks [8], [9], [10]. Indeed, human-centric communication has the following characteristics: 1) the downlink is usually more heavily loaded than the uplink in the support of data-hungry services provided by the core network; 2) the number of accommodated devices tends to be small, where the devices are relatively homogeneous such as intensively data-oriented smartphones and tablets, and their energy storage is relatively abundant due to the availability of frequent charging; 3) the delay requirements of various application scenarios are less stringent, e.g., the typical real-time transmissions require roughly 10-ms user-plane latency; and 4) the



(a) (b)

Fig. 1. (a) Usage scenarios. (b) Capabilities of 6G vision [7].

signaling overhead incurred by requesting access is not the main obstacle even for high-mobility scenarios. In contrast, mMTC exhibits the following features: 1) the uplink typically generates dominant data traffic and its performance becomes the main bottleneck due to the exceedingly large amount of signaling overhead for massive access, while it is necessary for devices to establish connectivity with the base station (BS) via the uplink access before initiating their downlink traffic; 2) heterogeneous devices exhibit periodic, continuous, or event-triggered uplink traffic, and the number of simultaneously served devices can be massive; and 3) the associated diverse delay requirements ranging from time-critical use cases (less than 1 ms) to delay-tolerant applications (up to 100 ms) are common. On the other hand, more stringent requirements on the capabilities have been put forward for the 6G vision, including connection density of 10^6–10^8 device/km^2, latency of 0.1–1-ms, reliability of $1–10^{-5}$ – $1–10^{-7}$, and so on, as illustrated by Fig. 1(b). Clearly, there still exists a huge performance gap to bridge even with the state-of-the-art technologies. As a result, it is urgently desired to design a new random access paradigm to embrace the IoHMT era, since the existing ones designed for human-centric communications do not facilitate the long-term evolution of machine-type communications [9].

To elaborate a little further, anticipated future IoHMT applications can be classified into the families of massive IoT[1] and critical IoT according to their different service requirements, as illustrated in Fig. 2. The massive IoT family generally involves a massive number of low-cost and energy-constrained devices, supporting uplink-dominated low-data rate transmissions. Its typical applications include smart wireless sensors for monitoring, alerts, and tracking in the fields of agriculture, city management, building automation, logistics, etc. [11]. In contrast, the critical IoT family requires ultrahigh reliability and ultralow latency for both fixed and mobile IoT scenarios. The typical applications encompass remote manufacturing/training/surgery, intelligent transport systems, smart grid, industrial automation, wearable devices, etc. [12]. Therefore, the two families have significantly diverse service requirements, which can be summarized from various key application cases as follows.

1) *Smart sensing, metering, and monitoring* require ultrahigh density device deployments with fixed locations in a large coverage, where the devices have ultralow power consumption, cost, and complexity. Moreover, the devices generate periodic or event-triggered low-rate traffic with small payload sizes and high delay tolerance. The data traffic is uplink dominated with few downlink control signalings being required.
2) *Building automation, logistics, and wearable IoT* require high-density device deployments supporting low mobility, where the devices have relatively low power consumption, cost, and complexity. Moreover, the devices exhibit medium-rate bidirectional traffic while having certain delay requirements for the uplink transmission. Also, mobile devices can report geo-locations to the servers for positioning.
3) *Intelligent transport systems* require medium-density device deployments with high mobility, where the devices generate periodic or event-triggered medium-rate traffic, and also have stringent requirements on the latency and reliability for the uplink transmission and downlink control.

[1]In accordance with existing standards and academic research, the term "IoT" will be consistently used in the text that follows. This terminology encompasses both the existing IoT and the potential IoHMT in the future.

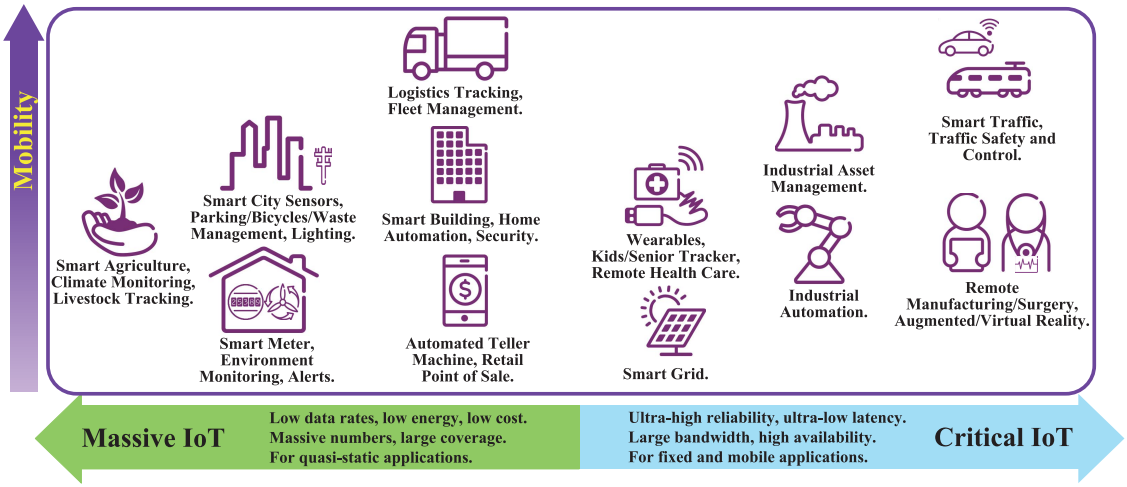


Fig. 2. Application cases of future IoT can be divided into the families of massive IoT and critical IoT.

4) *Remote manufacturing, training, and surgery* require low-density device deployments with known hot spot locations, where the devices generate continuous or event-triggered high-rate traffic and also demand stringent requirements on both the latency and reliability for bidirectional payload and signaling communications.

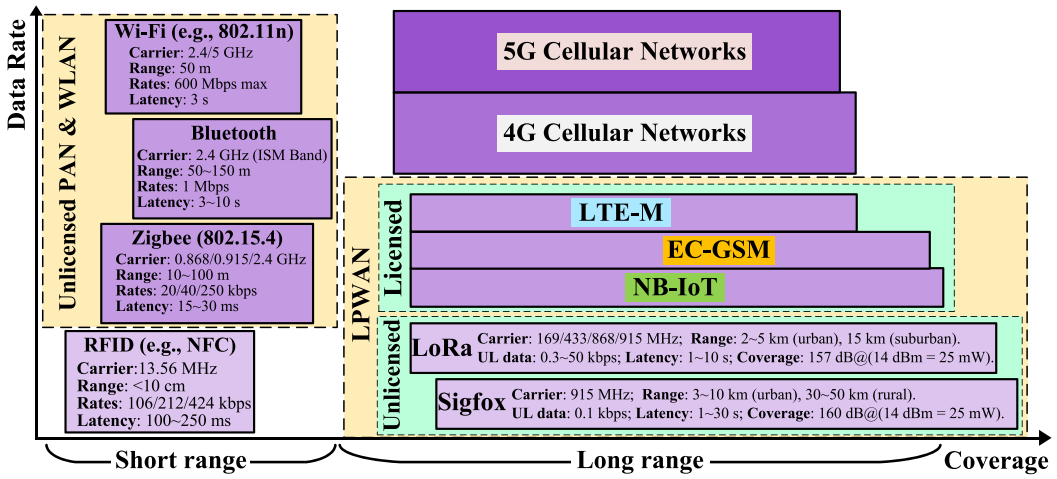
Clearly, one of the salient challenges of massive communication lies in designing a more efficient random access paradigm, which is expected to accommodate massive numbers of devices, reduce access latency, improve detection reliability, and satisfy heterogeneous service requirements. Generally speaking, there is a nontrivial tradeoff between latency and reliability in critical IoT applications [9], [10]. Most existing survey papers on massive access for mMTC only review the most recent advances from the academic community, while the overview of IoT standards in the industry community is limited [13], [14], [15]. Although different grant-based/grant-free nonorthogonal multiple access (NOMA) schemes have been reviewed, a comprehensive overview on the more specific compressive sensing (CS)-based grant-free massive access (GFMA) is still absent. This article seeks to fullfill this gap.

In this article, we first review the state-of-the-art IoT standards and mMTC solutions in the industry community and specify the major limitations of the existing random access schemes. Then, a comprehensive CS-based GFMA roadmap is presented, where the evolutions from single-antenna to large-scale antenna array-based BSs, from single-station to cooperative massive multiple-input–multiple-output (MIMO) systems, and from unsourced to sourced random access scenarios are detailed. Finally, the key challenges and open issues are summarized. For the convenience of readers, a list of major abbreviations used in this article is provided in the Appendix.

II. Overview of State-of-the-Art IoT Standards

Seamless and stable wireless connectivity is a fundamental prerequisite for designing IoT ecosystems. Owing to their heterogeneous service requirements and limited physical resources, there does not exist a single connectivity solution that can fit all emerging IoT applications. Currently, a proportion of IoT devices have been interconnected via low-cost commercial technologies, such as radio-frequency identification (RFID) [16], Bluetooth [17], ZigBee [18], wireless fidelity (Wi-Fi) [19], etc. However, these technologies only support short-range wireless communications, i.e., up to hundreds of meters, which severely hinders their practical implementations in future IoT applications that require ubiquitous coverage for widely distributed devices [14], [15]. Indeed, a significant number of IoT devices will have to be connected by low-power wide-area networks (LPWANs) for better coverage. To this end, the wireless communication industry has been standardizing several LPWAN solutions, which can be divided into unlicensed and licensed LPWANs, respectively [20]. In particular, the former category, also known as (a.k.a.) noncellular LPWAN, includes Sigfox and long-range radio (LoRa), while the latter category, a.k.a. cellular LPWAN, includes extended coverage global system for mobile communications (EC-GSM), long-term evolution for machine (LTE-M), and narrow-band IoT (NB-IoT) [21]. This section reports the whole landscape of existing IoT standards, in terms of data rate and coverage, and characterizes their key performance indicators, as illustrated in Fig. 3(a).

In practice, noncellular LPWANs are the emerging proprietary wireless connectivity solutions designed for low-cost devices in massive IoT [6]. Their advantages and drivers are low complexity and low cost, but at the expense of much lower throughput, higher latency, and susceptibility to the interference in unlicensed bands. In general, these low-cost options are still attractive to numerous enterprises interested in cheap IoT deployment. Particularly, Sigfox relies on a unified network that has been globally deployed and operated by the owing company for covering more than 60 countries and regions. Yet, the used chipset is open source since the company freely provides the protocol specifications to chip manufacturers as long as certain business terms are agreed upon [22]. In contrast, LoRa allows the customers to flexibly establish their private networks, but the involved physical-layer techniques of the chipset are proprietary to the U.S. corporation Semtech [23].



PAN: Personal area network; **WLAN:** Wireless local area network; **ISM:** Industrial scientific medical; **NFC:** Near-field communication; **UL:** Uplink. For the coverage, 160 dB@(14 dBm = 25 mW) indicates a service area with a maximum coupling loss of 160 dB and a transmit power of 14 dBm.

(a)

Solution / Parameter	EC-GSM	NB-IoT		LTE-M	
	Release 13	Release 13 (LTE Cat NB1)	Release 14 (LTE Cat NB2)	Release 13 (LTE Cat M1/eMTC)	Release 14 (LTE Cat M2/FeMTC)
Bandwidth	200 kHz (2.4 MHz for total bandwidth)	200 kHz (180 kHz in-band with 1 PRB)		1.4 MHz (1.08 MHz in-band with 6 PRBs)	5 MHz (4.32 MHz in-band with 24 PRBs)
Deployment	In-band GSM; Half-duplex FDD	In-Band & guard-band LTE, standalone; Half-duplex FDD		In-band LTE; Half-duplex FDD, full-duplex FDD, TDD	
Peak Rate (DL/UL)	DL/UL rates: 0.35~70 kbps (GSMK); up to 240 kbps (8PSK)	DL rates: ~26 kbps DL TBS: 680 bits UL rates: ~62 kbps/20 kbps (multi/singe tone) UL TBS: 1000 bits	DL rates: ~80/127 kbps (1 HARQ/2 HARQ) UL rates: ~105/159 kbps (1 HARQ/2 HARQ) UL/DL TBS: 2536 bits	DL rates: 0.3/0.8 Mbps UL rates: 0.375/1 Mbps Max BTS: 1000 bits	DL rates: 4 Mbps DL TBS: 4008 bits UL rates: 7 Mbps UL TBS: 6968 bits
Power Saving	PSM, ext. I-DRX	PSM, ext. I-DRX, C-DRX		PSM, ext. I-DRX, C-DRX	
DL Modulation	TDMA/FDMA, GMSK & 8PSK (optimal), 1Rx	OFDMA, 15 kHz tone spacing, QPSK, 1Rx		OFDMA, 15 kHz tone spacing, Turbo Code, QPSK/16QAM, 1Rx	
UL Modulation	TDMA/FDMA, GMSK & 8PSK (optional)	Single tone: 15/3.75 kHz spacing, $\pi/2$ BPSK, $\pi/4$ QPSK SC-FDMA: 15 kHz spacing, QPSK		SC-FDMA, 15 kHz single-tone spacing Turbo code, QPSK/16QAM	
Coverage	164 dB@33 dBm 154 dB@23 dBm	164 dB deep penetration	155 dB, 164 dB deep penetration	155.7 dB	
Power Class	33 dBm, 23 dBm	23 dBm, 20 dBm	23 dBm, 20 dBm, 14 dBm	23 dBm, 20 dBm	
Voice	No	No		VoLTE	
Mobility	Support	Non-support (only cell reselection in idle mode)	More mobility compared to Cat. NB1	Limited-to-full mobility	
Positioning	Cell ID	Cell ID	Enhanced cell ID and observed time difference of arrival	Enhanced cell ID	Enhanced cell ID and observed time difference of arrival
Latency	700 ms ~ 2 s	1.6 s ~ 10 s	At least the same as Cat NB1	10 ms ~ 15 ms	At least the same as Cat M1
Cost	< 7 $	< 5 $		< 10 $	
Battery Life	10 years	10 years		10 years	

PRB: Physical resource block; **FDD:** Frequency division duplexing; **TDD:** Time division duplexing; **DL:** Downlink; **GSMK:** Gaussian filtered minimum shift keying; **PSK:** Phase shift keying; **TBS:** Transport block set; **HARQ:** Hybrid automatic repeat request.

(b)

Fig. 3. (a) Whole landscape of existing IoT standards and their key features and (b) massive/critical IoT connectivity can be achieved by cellular LPWANs, including EC-GSM, LTE-M, and NB-IoT standardized by the 3GPP, where the corresponding features are listed [6], [14], [15].

On the other hand, cellular LPWANs, as listed in Fig. 3(b), are standardized by the third-generation partnership project (3GPP) exploiting licensed bands. Different cellular LPWANs complement each other in terms of technology availability, service requirements, and practical deployment conditions. Here, we summarize the standards of cellular LPWANs as follows.

1) *EC-GSM* was introduced in 3GPP Release 13 by adding new control and data channels to the conventional GSM networks, which can be readily achieved by applying a simple software update to the existing GSM systems [24]. Note that EC-GSM still dominates many mobile markets and has been supporting a majority of cellular IoT applications via general packet radio services (GPRS). As a benefit of its backward compatibility, deploying EC-GSM based on the global coverage of traditional GSM networks can result in an extensive coverage from day one, expediting its market penetration. In general, EC-GSM has a higher uplink capacity and wider downlink coverage than legacy GSM systems, at the expense of its higher power consumption and higher complexity at the devices.

2) *NB-IoT* is a clean-slate solution specifically tailored for massive low-throughput, low-cost, and energy-constrained IoT devices [25], [26]. It has engaged a new power-saving mode (PSM) and the extended discontinuous reception (eDRX) for prolonging the battery life of IoT devices to ten years or more, and achieves an extra 20 dB of power boost over the legacy GPRS [27]. NB-IoT has the advantages of reduced cost and improved energy efficiency over LTE-M. It also outperforms both Sigfox and LoRa in terms of throughput, response speed, and Quality of Services (QoS). It can be deployed either exploiting the guard-band, or within the existing 4G LTE spectrum, or as a standalone carrier relying on the second-generation (2G) spectrum. However, the handovers among different cells would be a problem for NB-IoT, due to the high control signaling overhead, which makes it the best suited for static setting rather than mobile devices.
3) *LTE-M* is the most flexible LPWAN solution supporting a full breadth of IoT application cases, varying from low-end static sensors to high-end mobile devices requiring high throughput [28]. It also has PSM and eDRX strategies that can be multiplexed onto the full LTE carriers. Hence, it is eminently suitable for co-existence with existing cellular networks [27]. The advantages of LTE-M over EC-GSM and NB-IoT include higher data rate, higher mobility, and the support of voice communications, albeit at the expense of requiring more bandwidth and a higher implementation cost.

Note that both NB-IoT and LTE-M were introduced in Release 13 and have evolved to Release 17 at the time of writing. Fig. 3(b) compares EC-GSM, NB-IoT (Releases 13 & 14), and LTE-M (Releases 13 & 14), and the readers can refer to [24], [25], [26], [27], and [28] for their detailed technical parameters, including bandwidth, peak rate, modulation type, latency, cost, battery life, etc. Compared to Releases 13 & 14, Releases 15 & 16 for NB-IoT and LTE-M have further improved the spectral and energy efficiencies. Moreover, Release 17 has carried out a study on the possibility and required specification updates to support NB-IoT and LTE-M in nonterrestrial networks. Considering the random access of NB-IoT and LTE-M, Release 15 introduced an early data transmission (EDT) mode, while Release 16 further enhanced the uplink data payload of EDT and introduced a preconfigured uplink resources (PUR) mode [27]. These emerging strategies can reduce the overhead for signaling exchanges, thus improving the system energy efficiency and reducing the random access latency, which will be detailed in the next section.

III. Existing Random Access Solutions and Limitations

Efficient random access protocols and multiple access techniques are the fundamental premises for connecting a massive number of devices. Compared to the classical grant-based four-step random access (FSRA) developed in 4G LTE [29], the industry community has gradually simplified the random access procedure in recent 3GPP releases. Due to the need for the contention or the preconfiguration of orthogonal physical resources for avoiding interdevice interferences, these solutions belong to the grant-based/grant-free orthogonal multiple access (OMA) paradigms. However, the maximum number of accommodated devices is limited by the number of orthogonal resources. To overcome this limitation, various NOMA solutions have also been intensively investigated in the academic community. Nevertheless, most of them have inherent limitations in supporting future massive IoT connectivity.

A. Random Access in Industry Standardization

In unlicensed bands, both Sigfox and LoRa adopt ALOHA-based random access schemes, where the devices exploit a random frequency and time division multiple access technique to transmit their signals [30]. Without the need for direct signaling interactions between the devices and the BS to establish connection, these schemes can be classified into the grant-free OMA category. Although the related access procedure is simple enough, the severe collision is a limiting factor for the realization of low-latency and high-efficiency random access when the number of devices becomes large. In licensed bands, similar to 4G LTE, the early NB-IoT and LTE-M in Release 13 adopt the grant-based FSRA protocol illustrated in Fig. 4(a), which includes Msg1–Msg4 [15]. More specifically, the FSRA procedure is summarized as follows.

1) *Step 1:* According to the system information periodically broadcasted by the BS, the requesting devices transmit a contention-based orthogonal preamble, i.e., Msg1, on the uplink physical random access channel (PRACH).
2) *Step 2:* After successfully receiving the preamble, the BS broadcasts a random access response (RAR), i.e., Msg2, which encompasses the detected preamble identification, time-alignment instructions for uplink synchronization, temporary cell radio network temporary identifier (TC-RNTI), and the uplink grant for Msg3.
3) *Step 3:* Upon receiving the RAR within a given time window, the device, whose Msg1 is successfully detected by the BS, transmits its connection request, i.e., Msg3, via a physical uplink shared channel (PUSCH) indicated in Msg2. If the RAR is not received within the given time window, this round random access predicates failure.
4) *Step 4:* After receiving Msg3, the BS replies to the devices with a contention resolution message, i.e., Msg4. If a device finds its contention resolution identification in Msg4, an acknowledgment is fed back to the BS, the FSRA procedure is completed and the granted device moves to the connected mode. Otherwise, a new access schedule is attempted.

The grant-based FSRA procedure necessitates two round-trip interactions between the devices and the BS to establish connection, which results in significant signaling overhead and long access latency. Moreover, the number of available orthogonal preambles is generally limited due to the limited number of physical resources. Typically, a fraction of the total 64 orthogonal preambles are reserved for contention-free access

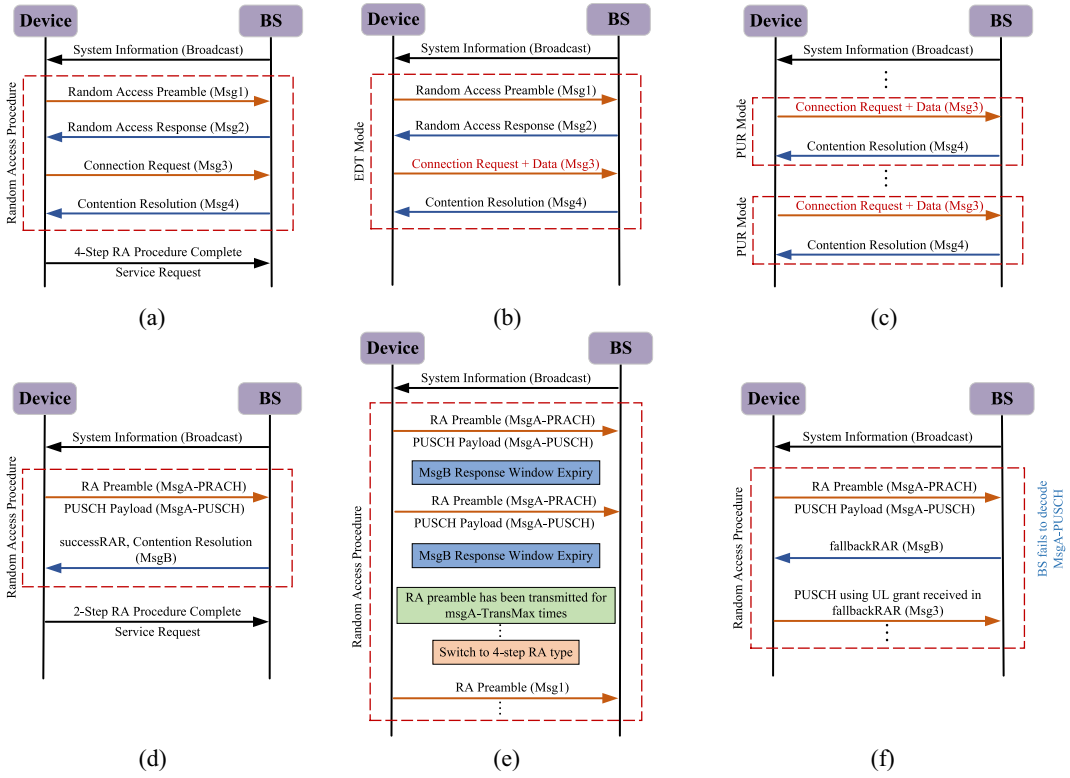


Fig. 4. (a) Standard FSRA in 4G LTE, (b) FSRA with EDT mode, (c) PUR-based random access, and (d)–(f) respectively, correspond to three cases of TSRA in 5G NR.

(e.g., handover) and only the remaining orthogonal preambles can be adopted for contention-based random access [27]. Therefore, as the number of simultaneously served devices becomes large, the access efficiency is significantly degraded due to high collision probability. To improve the access efficiency for NB-IoT and LTE-M, the EDT and PUR modes were introduced in Release 15 and Release 16, respectively [31].

As depicted in Fig. 4(b), for the EDT mode, a device in the idle state may directly transmit a maximum of 1000 data bits embedded in Msg3 of FSRA. After a successful reception at the BS, the device may remain in the idle state or switch to the connected state for further data transmission. In other words, if a device has short data ($\leq$ 1000 bits), the data can be transmitted in a single successful FSRA without the need to go to the connected mode. On the other hand, for long data transmission, a device may rely on the grant-based FSRA to establish the connection.

In contrast, PUR is tailored for quasistatic devices conveying periodic or pseudo-varying short data traffic. Specifically, the devices adopting the PUR mode are preconfigured with uplink transmission resources via dedicated radio resource control signaling. Thus, Msg1 and Msg2 are omitted, and data can be directly delivered in Msg3, as illustrated in Fig. 4(c). In practice, the PUR mode has two categories: 1) dedicated PUR and 2) shared PUR. The former is designed for devices conveying periodic traffic, where the uplink time–frequency resources are exclusive for each device, while for the latter, the same uplink time–frequency resources are shared by up to two devices, where the superimposed signals can be distinguished by mutually orthogonal demodulation reference signals (DMRS). Clearly, the PUR-based random access is a grant-free OMA solution, which cannot accommodate a large number of devices due to the limited number of orthogonal resources and DMRS.

In contrast to FSRA, the recent Release 16 introduces a grant-based two-step random access (TSRA) protocol for the 5G new radio (NR) [31]. By resorting to a simplified single round-trip interaction between the devices and the BS, the access latency and control signaling overhead can be reduced. In particular, the TSRA combines the PRACH preamble and PUSCH payload as a single MsgA transmitted by the devices, and then merges RAR and contention resolution message into a single MsgB. Fig. 4(d)–(f) portrays three typical cases of the TSRA. Specifically, they are as follows.

1) *Case 1:* Upon the device receiving an MsgB with successful RAR and contention resolution, the grant-based TSRA is completed and the device moves to the connected mode, as illustrated in Fig. 4(d).
2) *Case 2:* If the device fails to receive an MsgB after a maximum number of MsgA trials, the device switches to the grant-based FSRA mode by transmitting Msg1 for a new access attempt, as shown in Fig. 4(e).
3) *Case 3:* If the MsgA cannot be decoded correctly, the BS broadcasts an MsgB with fallback RAR, and the device falls back to the grant-based FSRA mode by transmitting Msg3 for connection request, see Fig. 4(f).

Note that only 5G NR supports the grant-based TSRA, while NB-IoT and LTE-M currently do not have this mode. At the time of writing, all the standardized random access solutions for cellular LPWANs adopt OMA techniques,

i.e., exploiting orthogonal radio resources to distinguish different devices and, thus, they belong to the grant-based/grant-free OMA paradigms. Here, almost all these solutions require a grant-based random access procedure for orthogonal resource contention. The only exception is the PUR-based random access, where the orthogonal resources are preconfigured. Specifically, for Msg1 or MsgA, the PRACH preamble of each active device is anonymously selected from a predefined orthogonal sequence pool, while for shared PUR, two devices are distinguished via orthogonal DMRS.

However, the number of orthogonal radio resources is limited but the number of requesting devices has been exponentially increasing, since the rise of the IoT. Therefore, all these OMA solutions suffer from unavoidable preamble collision with degraded access efficiency, which eventually leads to the network congestion [15], [27], [31].

B. Review of Nonstandardized NOMA

Compared with OMA, NOMA has the potential to accommodate a larger number of devices, which can be simultaneously served over a small amount of time–frequency resource elements by exploiting the devices' unique but nonorthogonal signatures. A NOMA scheme is generally a grant-free solution, offering advantage of reducing access signaling overhead and latency. There are potentially many NOMA techniques, but the most popular ones are classified into two categories: 1) power-domain NOMA and 2) code-domain NOMA [13]. More recently, spatial-domain NOMA has attracted much attention, which relies on MIMO technology to support multiple users on the same time–frequency resources via uplink multiuser detection (MUD) and downlink multiuser transmit (MUT) precoding/beamforming [32], [33], [34], [35], [36], [37], [38], [39].

1) Power-Domain NOMA: The signals of multiple devices are superimposed on the same time–frequency resources with different power levels, which requires dedicated power allocation strategies at the transmitter and the successive interference cancelation (SIC)-based data detection at the receiver for efficient decoding [40], [41], [42], [43]. However, power-domain NOMA is usually limited to serving a small number of human-type devices. Specifically, it is challenging to ensure distinguishable power levels for massive devices, particularly in grant-free random access. Furthermore, it is also impractical to design a reliable SIC-based receiver, since the power levels are generally not distinctive while the resolutions of analog-to-digital converters (ADCs) are limited. In practice, the users have to be divided into groups, each containing only a small number of users. The users in the same group can adopt a power-domain NOMA scheme for transmission, while the users in different groups have to employ OMA schemes for avoiding intergroup interferences. Therefore, a grant-based scheduling procedure is required, which will result in extra signaling overhead and latency.

2) Code-Domain NOMA: On the other hand, code-domain NOMA realizes multiplexing in the code domain, where multiple devices share the same time–frequency resources but adopt nonorthogonal low-cross-correlation spreading sequences as their unique signatures [44], [45], [46], [47], [48], [49]. The superimposed signals of different devices are distinguished by leveraging the uniqueness of spreading sequences. This category of NOMA schemes is inspired by the classical code-domain multiple access (CDMA) utilizing orthogonal spreading sequences to distinguish devices. The key difference lies in that the spreading sequences of code-domain NOMA are nonorthogonal low-cross-correlation sequences and, thus, more devices can be accommodated given the same amount of physical resources. Compared with power-domain NOMA, code-domain NOMA is superior in supporting massive IoT connectivity with grant-free random access, since the number of available spreading sequences is much larger than the resolution of the received power levels. In general, various code-domain NOMA schemes can be further categorized into low-density spreading (LDS) NOMA class and dense spreading (DS) NOMA class, depending on whether the adopted spreading sequences are sparse or not.

In practice, LDS NOMA schemes, such as the intensively investigated LDS-CDMA, LDS-OFDM, and sparse code multiple access (SCMA), adopt sparse spreading codes, which can be transmitted either in the time or frequency domain. Compared with the DS codes of conventional CDMA, the sparse codes of LDS NOMA can still offer certain spreading gains to suppress the undesired interdevice interferences, while facilitating the application of low-complexity message passing algorithms for data detection [50]. LDS-CDMA and LDS-OFDM are two initial code-domain NOMA schemes that are directly extended from the traditional CDMA and OFDM cellular systems. In particular, for LDS-CDMA, the symbol to be transmitted is spread over the time domain, while for LDS-OFDM, the chips are transmitted in the frequency domain. Developed from basic LDS-CDMA, the state-of-the-art SCMA scheme directly maps the transmitted bit streams onto a set of sparse codewords, which results in shaping gains, leading to an improved detection performance. Specifically, let us consider the SCMA schematic diagram of Fig. 3 in [51] as an example, where each device has a unique sparse codebook of $L_{\mathrm{SCMA}} = 4$ sparse codewords such that $(\log_2 L_{\mathrm{SCMA}})$-bit information per channel use can be delivered. In this context, $K_{\mathrm{SCMA}} = 6$ devices share $N_{\mathrm{SCMA}} = 4$ resource elements and this is termed as a subcarrier block, where the overloading rate is defined as $\gamma = K_{\mathrm{SCMA}}/N_{\mathrm{SCMA}} = 150\%$. Moreover, the K devices' sparse codebooks can be reused in multiple subcarrier blocks for joint transmission and decoding. Note that sparse codes of the same device share the same sparse pattern (i.e., the positions of nonzero elements in a vector), which is unique for each device.

On the other hand, multiuser shared access (MUSA) is a typical DS NOMA scheme, where a set of nonorthogonal DS sequences constitutes a pool and each device anonymously chooses a sequence to spread its transmit symbol [49]. Compared with the LDS-CDMA and LDS-OFDM where multiple resource blocks reuse the same LDS sequence, the devices in MUSA may pick different sequences for spreading different symbols, attaining an enhanced performance with the aid of interference averaging. Another difference between the two classes lies in their spreading sequence configuration

mode. Devices in LDS-CDMA and LDS-OFDM are preconfigured to employ unique spreading sequence, while MUSA adopts the so-called contention-based random access, where devices share the same pool of sequences. More details on SCMA and MUSA can be found in [48], [49], and [50].

Now, let us consider the sporadic traffic generated by massive communication. Although the number K of devices to be served can be hundreds even thousands, the number of simultaneously active devices K_a is relatively small [10]. For example, $K = 480$ and $K_a = 48$ result in an activity ratio of $\rho = 0.1$. Unfortunately, most existing NOMA schemes are incapable to cope with this sparse traffic. Taking SCMA as an example, designing high-dimensional sparse codebooks for simultaneously supporting devices for a large K is quite challenging, since the sparse teletraffic of massive communication cannot be readily exploited. Considering the sparse codebooks designed in [50] for $K_{\text{SCMA}} = 48$ and $N_{\text{SCMA}} = 24$, SCMA has to divide $K = 480$ total devices into ten SCMA groups having a total of $10N_{\text{SCMA}} = 240$ resource elements. The devices in the same group adopt SCMA scheme for transmission, while the devices in different groups employ OMA schemes for avoiding intergroup interferences. In this context, a grant-based scheduling procedure may be required, which results in extra signaling overhead and latency.

3) Spatial-Domain NOMA: Spatial-domain NOMA, also known as space-division multiple access in some earlier literature [52], exploits the extra spatial degrees-of-freedom (DoF) offered by MIMO techniques to realize multiplexing. Specifically, the signals of multiple devices conveyed on the same time–frequency resources are distinguished by exploiting unique user-specific channel impulse responses (CIRs). Intuitively, the user-specific CIR plays a role similar to the nonorthogonal spreading sequence in code-domain NOMA. Given the users' CIRs, therefore, the signals of multiple devices conveyed on the same time–frequency resources can be recovered using MUD at the uplink BS receiver [32], [33], [34], [35] or distinguished through MUT precoding at the downlink BS transmitter [36], [37], [38], [39]. Hence, spatial-domain NOMA requires accurate estimate of users' CIRs or MIMO channel state information (CSI). Acquisition of accurate MIMO CSI however imposes considerable pilot resource overhead, which may be unaffordable in practice. In order to attain near-optimal performance based on the limited pilot resources, joint channel estimation and turbo MUD/decoding solutions have attracted substantial research interests [33], [34]. However, these joint channel estimation and data detection solutions are generally computationally expensive, and they can only support limited number of users.

By deploying large-scale or massive antenna array at BS, favorable massive MIMO (mMIMO) environment is created for implementing spatial-domain NOMA. In particular, the CIRs associated with different users become nearly orthogonal, and the signals of multiple devices can be separated with low-complexity conjugate beamforming [53]. But acquisition of the mMIMO CSI becomes even more challenging. In order to achieve affordable-complexity mMIMO CSI estimation, orthogonal pilot sequences must be adopted. However, the number of orthogonal pilot sequences available is limited, and these pilot resources will have to be reused in neighboring cells. This causes severe pilot contamination which results in the BS being unable to reliably differentiate the signals of different cells. Sophisticated pilot designs [54], [55], [56], [57] have been developed to mitigate pilot contamination. In the past decade, the emerging of mMIMO techniques has accelerated the development of spatial-domain NOMA. However, the number of devices that can be supported by the existing spatial-domain NOMA solutions is still limited, and it is very challenging to apply these existing spatial-domain NOMA techniques to support a massive number of devices in future massive communication scenarios.

IV. Compressive Sensing-Based Grant-Free Massive Access Paradigm

The discussions in the previous sections reveal that neither the current standardized OMA solutions nor the existing nonstandardized NOMA solutions can well accommodate future IoT applications with massive communication. To tackle this issue, a CS-based GFMA paradigm was recently developed, where the active devices directly transmit their uplink access signals over the same time–frequency resources without the need for any scheduling in advance [58], [59]. Meanwhile, by leveraging the sporadic traffic of devices, the multidevice detection at the BS can be formulated as a CS problem that can be effectively resolved by various sparse signal recovery algorithms [60]. Therefore, the complicated signaling interactions for access scheduling, including resource granting and contention resolution, associated with grant-based random access protocols are circumvented. Furthermore, compared with conventional NOMA schemes, the designs of distinguishable access signatures and the multidevice detection algorithm are significantly simplified by further taking into account the sporadic traffic.

To begin with, the essence of the CS theory can be well captured in the following discussion starting with the mathematical expression of:

$$\mathbf{Y} = \Phi\mathbf{X} + \mathbf{N} \tag{1}$$

where $\mathbf{Y} \in \mathbb{C}^{M\times Q}$ represents the low-dimensional measurements, $\Phi \in \mathbb{C}^{M\times N}$ is the sensing matrix with $M \ll N$, $\mathbf{X} \in \mathbb{C}^{N\times Q}$ is the original high-dimensional sparse signal, and $\mathbf{N}$ is the additive white Gaussian noise (AWGN). The CS theory indicates that given $\mathbf{Y}$ and Φ, the sparse matrix $\mathbf{X}$ can be exactly recovered as long as $M \geq N_a \log_2(N)$ is satisfied [51]. Here, $N_a \ll N$ is the maximum number of nonzero elements in the columns of $\mathbf{X}$. As such, the classic model in (1) becomes a standard single-measurement vector (SMV) CS problem for $Q = 1$ [61] and a multiple-measurement vector (MMV) CS problem for $Q > 1$ [62], [63], [64].

The key challenge in solving the CS recovery problem is how to design a computationally efficient sparse signal recovery algorithm. Various CS recovery algorithms have been proposed, which can be classified into three categories, namely, convex relaxation algorithms, greedy-based algorithms, and Bayesian inference algorithms. Specifically, convex relaxation algorithms, such as basis pursuit [65] and

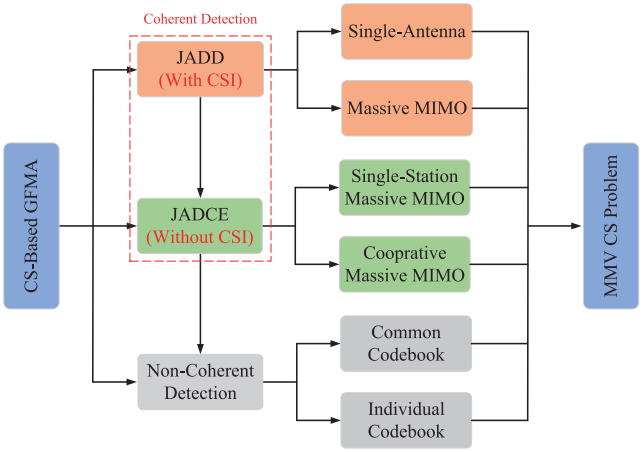


Fig. 5. Classification of existing CS-based GFMA schemes. The evolutions from single-antenna to large-scale antenna array-based BSs, from single-station to cooperative mMIMO systems, and from unsourced to sourced access scenarios are illustrated.

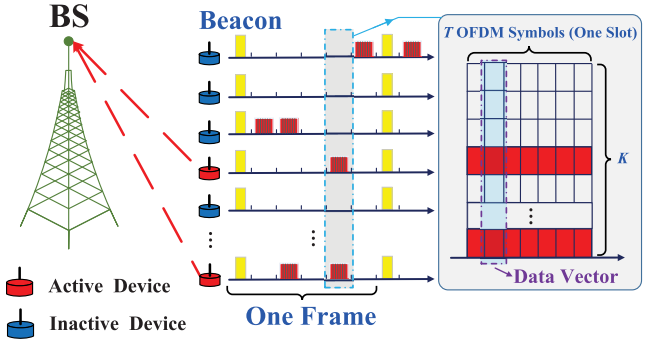


Fig. 6. System model of spreading-based GFMA scheme in single-antenna systems, where the temporal structured sparsity of the signal matrix is illustrated [75] ©IEEE.

least absolute shrinkage and selection operator (LASSO) [66], relax the nonconvex CS recovery problem as a conventional convex optimization problem and employ linear programming methods to acquire the solution. These algorithms generally enjoy an excellent recovery performance but the related computational complexity is extremely high, especially for large problem dimensions. In contrast, greedy-based algorithms, such as orthogonal matching pursuit (OMP) [67], subspace pursuit (SP) [68], and CoSaMP [69], identify the nonzero indices of $\mathbf{X}$ and estimate their corresponding coefficients in a greedy iterative manner. In general, they have low algorithmic complexity but suffer from significant performance losses when the number of measurements is relatively small or the signal-to-noise ratio (SNR) is low. Both convex relaxation and greedy-based algorithms only consider the sparsity of $\mathbf{X}$ but fail to leverage its statistical information for effectively improving recovery accuracy. To overcome this limitation, Bayesian inference algorithms, such as belief propagation [70], expectation propagation (EP) [71], and sparse Bayesian learning (SBL) [72], were developed under the Bayesian framework, by establishing various flexible a priori distributions to capture the sparsity properties and the statistical information of $\mathbf{X}$, thus reaping a better recovery performance. Moreover, the tradeoff between recovery performance and computational complexity can be effectively achieved by a low-complexity approximation to the standard Bayesian inference framework, known as the approximate message passing (AMP) framework [73].

This section presents the roadmap of the promising CS-based GFMA paradigm, as illustrated in Fig. 5, where various specific GFMA schemes are introduced to fulfill the heterogeneous massive communication requirements of practical IoT applications. In particular, the evolutions from single-antenna to large-scale antenna array-based BSs, from single-station to cooperative mMIMO systems, and from unsourced to sourced access scenarios are detailed in this section. A dominated common characteristic of these schemes is that the multidevice detection at the BS, i.e., activity detection and channel estimation (or data detection), can be formulated as a CS problem of (1). Moreover, the sparse matrix $\mathbf{X}$ generally exhibits different structured sparsity properties in different access scenarios, which can be exploited to further improve recovery performance with the aid of the bespoke algorithms.

A. Joint Activity and Data Detection

For the CS-based GFMA paradigm, the wireless transceiver can be flexibly designed to accommodate the practical heterogeneous massive communication requirements, resulting in various specific GFMA schemes. Assume that the CSI is available at receiver. Note that this assumption is valid in the scenarios where the channels can be efficiently estimated with very low pilot overhead or the CSI remains unchanged for a long time, such as single-antenna systems and fixed sensor networks, respectively [74]. With the CSI, BS can jointly detect the active devices and their payload data from the overlapped received signal. Focusing on this joint activity and data detection (JADD) problem, this section first investigates CS-based GFMA schemes in single-antenna systems. Then the problem is extended to mMIMO systems, where the additional spatial DoF is exploited to enhance uplink throughput and improve detection performance.

1) JADD for GFMA in Single-Antenna Systems: The GFMA in single-antenna systems generally adopts spreading-based transmission scheme. Consider a massive IoT connectivity scenario with one single-antenna BS serving K single-antenna devices, where K is usually large and OFDM is adopted to combat the time dispersion effect, as illustrated in Fig. 6. Due to the sporadic uplink traffic, only K_a ($K_a \ll K$) out of the total K devices are active during each time slot with T OFDM symbols, in which the activity and CSI remain unchanged. The BS periodically broadcasts its beacon signals to facilitate synchronization, power control, and channel estimation at the devices. Since only one single antenna is considered at both the devices and the BS, the beacon signal overhead is very small for downlink channel estimation. To distinguish the kth device from others at the BS, its access signal, $x_{k,t} \in \mathbb{C}$, in the tth OFDM symbol duration ($1 \le t \le T$) is spread across L subcarriers by a unique spreading sequence $\mathbf{s}_k \in \mathbb{C}^{L\times 1}$. Moreover, benefiting from the channel reciprocity, the downlink CSI estimates are exploited for preequalizing in the uplink transmission to precompensate the impact of uplink channels at the devices. The active devices

can directly transmit their spread access signals on the exactly same time–frequency resources, without any scheduling in advance. This avoids the complicated signaling interactions in FSRA or TSRA, and hence significantly reduces the access latency.

Adopting the aforementioned spreading-based GFMA, the signals of all the active devices are overlapped at the BS, which results in severe interdevice interferences. Therefore, it is essential to design a reliable JADD scheme at the BS, which can be formulated as an MMV CS problem of (1). Specifically, the sensing matrix is expressed as $\Phi = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K] \in \mathbb{C}^{M\times N}$ with $M = L$ and $N = K$, and the sparse signal matrix is expressed as $\mathbf{X} = [\alpha_1\mathbf{x}_1, \alpha_2\mathbf{x}_2, \ldots, \alpha_K\mathbf{x}_K]^T \in \mathbb{C}^{N\times Q}$ with $Q = T$. Here, the binary variable $\alpha_k \in \{0, 1\}$ denotes the activity indicator with 1 being active and 0 otherwise, and $\mathbf{x}_k \in \mathbb{C}^{T\times 1}$ is the spread access signal of the kth device.

Note that the device activity and payload data are embedded in the access signal matrix $\mathbf{X}$, i.e., the indices of nonzero rows indicates the identities of active devices and the corresponding coefficients are the transmitted signals. Hence, the JADD problem is equivalent to reconstructing $\mathbf{X}$ based on the overlapped received signal $\mathbf{Y}$ and the spreading matrix Φ. Based on the estimate of $\mathbf{X}$, the payload data can be further detected. To this end, Wang et al. [74] developed a CS-message passing algorithm (MPA) detector, where the CoSaMP algorithm and MPA are employed for CS-based active device detection and payload data detection, respectively, but only a single OFDM symbol ($T = 1$) is considered in [74]. In practice, the data packet of active devices usually occupies several consecutive OFDM symbols [76]. Therefore, the system is synchronized in a slot structure[2] and the device activity remains constant during each slot, which leads to the *temporal common sparsity* pattern, as illustrated in Fig. 6. Furthermore, although the device activity may change across different time slots, the variation is gradual, i.e., the active devices generally transmit their data in consecutive time slots (i.e., burst transmission) with a high probability. This leads to the temporal correlation over several consecutive time slots, which is referred to as *temporal dynamic sparsity*. Under this context, various JADD algorithms have been proposed to leverage these two temporal structured sparsity properties for improved detection performance. For instance, Wang et al. [78] proposed a structured iterative support detection (SISD) algorithm to leverage the temporal common sparsity, which follows the idea of greedy-based CS recovery algorithms. By resorting to the Bayesian inference framework, a joint expectation maximization AMP (EM-AMP) algorithm was developed to further exploit the a priori statistical information of the transmitted discrete symbols [79]. However, both SISD algorithm and EM-AMP algorithm reconstruct the access signals of different symbol durations separately, which fails to take the full advantages of the temporal common sparsity. Hence, Du et al. [80] proposed a block sparsity-based detection algorithm, which vectorizes $\mathbf{X}$ into a block-sparse vector for a better use of the temporal common sparsity based on the block CS theory [81]. In addition, an orthogonal AMP with accurate structure learning (OAMP-ASL) algorithm was also proposed in [75], where the temporal common sparsity is incorporated in the a priori distribution of $\mathbf{X}$ for further improving performance.

[2]In the frame structure of 5G NR, each OFDM frame consists ten subframes with each subframe having several time slots. The number of time slots within each subframe depends on the subcarrier spacing and each slot consists of 7 OFDM symbols [77].

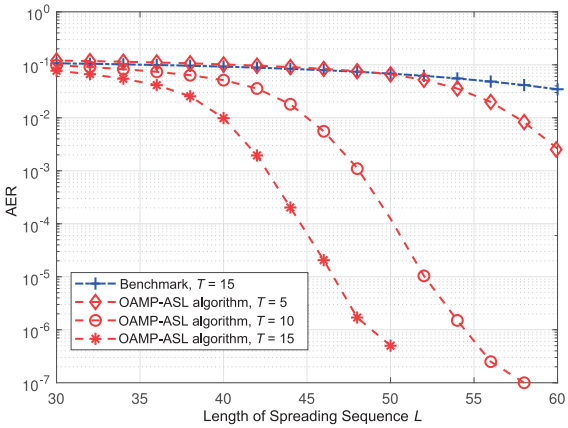


Fig. 7. AER performance comparison of the advanced OAMP-ASL-based JADD scheme [75] and the benchmark [82], where $K = 500$ and $K_a = 50$, and K_a is unknown at the BS.

Fig. 7 provides an example to verify the superiority of leveraging the temporal common sparsity, where the activity error rate (AER) of the OAMP-ASL algorithm [75], which incorporates the temporal common sparsity, is compared with that of the traditional greedy CS recovery algorithm [82] without considering the temporal common sparsity. As can be observed, the OAMP-ASL algorithm considerably outperforms the baseline scheme without incorporating the temporal common sparsity. Also as expected, the performance of the OAMP-ASL algorithm improves as the number of symbols within a slot T increases, since a larger T indicates an enhanced temporal common sparsity.

On the other hand, focusing on the temporal dynamic sparsity, Wang et al. [83] proposed a dynamic CS-based JADD approach, where the active device set estimated in the current slot is adopted to initialize the estimate of the active device set in the next slot. However, the solution of [83] assumes the availability of the sparsity level, i.e., the number of active devices, which is unrealistic in practical scenarios. Moreover, the prior information is exploited blindly where the reliability of the estimate from the previous slot is not evaluated. To overcome this limitation, a prior-information aided adaptive subspace pursuit (PIA-ASP) algorithm [84] is developed, which reaps a better symbol error rate (SER) performance, as illustrated in Fig. 8.

For the benefit of the reader, a brief summary of the aforementioned JADD schemes is provided in Table I.

2) JADD for GFMA in Massive MIMO Systems: mMIMO has been identified as a pivotal technique for current 5G NR and future beyond 5G (B5G)/6G cellular systems, providing game-changing improvements in the spectral and energy efficiencies [85], [86], [87]. Moreover, the transmission reliability of massive IoT connectivity can be considerably improved by leveraging the extra spatial DoF [88]. To reap these

TABLE I
SUMMARY OF JADD ALGORITHMS FOR SPREADING-BASED GFMA SCHEME IN SINGLE-ANTENNA SYSTEMS

JADD Scheme	CS Model	CS Algorithm	Sparsity Structure	Advances	Complexity
CS-MPA algorithm [74]	SMV	Greedy & Bayesian	None	Exploit the sporadic uplink traffic of devices	$\mathcal{O}(KLK_a + q^w)$
SISD algorithm [78]	MMV	Greedy	Temporal common sparsity	Exploit the temporal common sparsity	$\mathcal{O}(TK^3)$
EM-AMP algorithm [79]	MMV	Bayesian	Temporal common sparsity	Further exploit the *a priori* statistical information of the transmitted discrete symbols	$\mathcal{O}(TKL)$
Block sparsity-based algorithm [80]	SMV	Greedy	Block sparsity	Make better use of the temporal common sparsity	$\mathcal{O}(T^2KL + T^2Ls^2 + T^3s^3)$
OAMP-ASL algorithm [75]	MMV	Bayesian	Temporal common sparsity	Incorporate the temporal common sparsity in the *a priori* distribution of the access signal matrix	$\mathcal{O}(TKL + T^2K + TKq)$
Dynamic CS-based algorithm [83]	MMV	Greedy	Temporal dynamic sparsity	Exploit the temporal dynamic sparsity	$\mathcal{O}(TKLK_a + TLK_a^3)$
PIA-ASP algorithm [84]	MMV	Greedy	Temporal dynamic sparsity	Make better use of the temporal dynamic sparsity	$\mathcal{O}(TKL + TK + (s_p + j)^3)$

Notes: q is the modulation order, w is the maximum number of symbols spreading over the same subcarrier, s_p is the quality information, and j is the update index of sparsity level.

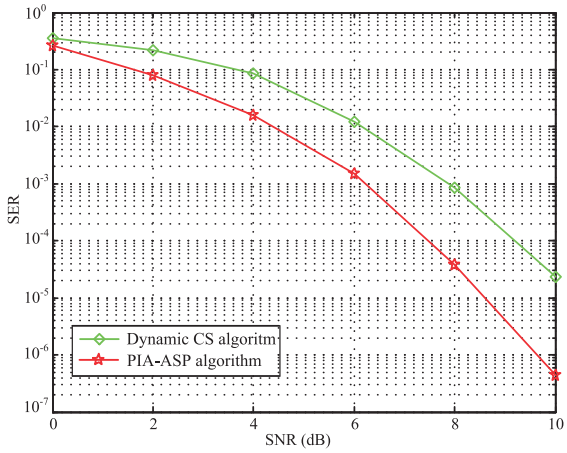


Fig. 8. SER performance comparison of the dynamic CS-based JADD scheme and PIA-ASP-based JADD scheme, where $K = 200$, $L = 100$, and K_a varies from 14 to 20 in seven different time slots [84] ©IEEE.

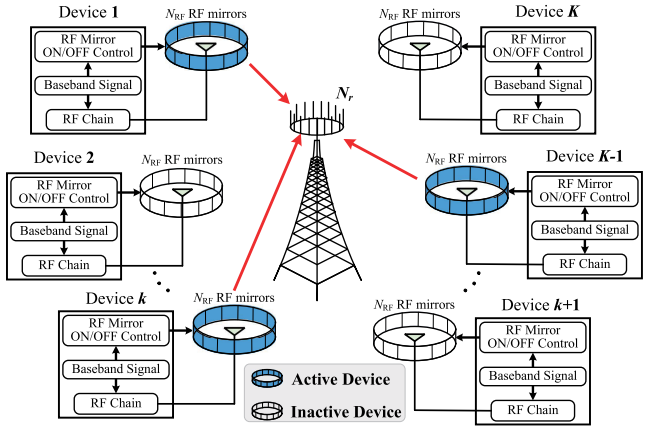


Fig. 9. System model and transmitter structure of MBM-based GFMA scheme in mMIMO systems [96] ©IEEE.

benefits, one effective way is to intuitively extend the problem formulation in the previous section to mMIMO systems. Specifically, for mMIMO systems with N_r BS antennas, the channel vector between the BS and the kth device, denoted by $\mathbf{h}_k \in \mathbb{C}^{N_r \times 1}$, is unique for the device, which can be regarded as a device-specific signature spreading in the spatial domain. Naturally, if the channels associated with all the K devices are available at the BS, the JADD for GFMA can be formulated as a MMV CS problem of (1). Here, $\Phi = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_K] \in \mathbb{C}^{N_r \times K}$ is the known massive access channel matrix and $\mathbf{X} = [\alpha_1\mathbf{x}_1, \alpha_2\mathbf{x}_2, \ldots, \alpha_K\mathbf{x}_K]^T \in \mathbb{C}^{K \times T}$ is the sparse signal matrix as explained in Section IV-A1. In this context, the CS recovery algorithms introduced in Section IV-A1 can be directly applied to leverage the temporal common sparsity or temporal dynamic sparsity of $\mathbf{X}$.

On the other hand, by equipping multiple antennas at the devices, the spatial modulation (SM) can be incorporated to boost the spectral efficiency for GFMA, without increasing the hardware complexity and energy consumption of the devices [89], [90]. In the SM scheme, each active device activates only one transmit antenna, based on which the additional $\log_2(N_t)$-bit information can be conveyed through the active antenna index [91], [92], where N_t is the number of transmit antennas. SM is a so-called index modulation scheme that exploits the transmit antenna index to convey additional information bits [92]. In this context, only one radio-frequency (RF) chain is required at the devices, but the number of transmit antennas scales exponentially with the number of additionally conveyed information bits. Following a similar index modulation idea, the more efficient media-based modulation (MBM) [93], [94], [95] has been widely investigated to overcome the aforementioned limitation of SM. Specifically, each device is equipped with one RF chain, one transmit antenna, and N_{RF} low-cost RF mirrors, where each RF mirror has a controllable binary ON/OFF status, as illustrated in Fig. 9. Therefore, each device has $N_t = 2^{N_{\text{RF}}}$ different mirror activation patterns, i.e., N_t different channel realizations, which can be exploited to encode N_{RF} extra information bits.

For both the SM-based and MBM-based GFMA schemes, the related JADD at the BS can also be formulated as a MMV CS problem as expressed in (1). Specifically, the sensing matrix is $\Phi = [\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_K] \in \mathbb{C}^{N_r \times KN_t}$ with $\mathbf{H}_k \in \mathbb{C}^{N_r \times N_t}$ denoting the MIMO channel matrix between the BS and the kth device for all the channel realizations. Again assume that the full CSI is available at the BS and the device activity remains constant within a slot having T successive symbols. The tth column of the sparse signal matrix $\mathbf{X}$ is expressed as $[\mathbf{X}]_{:,t} = [(\mathbf{x}_{1,t})^T, (\mathbf{x}_{1,t})^T, \ldots, (\mathbf{x}_{K,t})^T]^T \in \mathbb{C}^{KN_t \times 1}$, where $\mathbf{x}_{k,t} = \alpha_k s_{k,t}\mathbf{d}_{k,t} \in \mathbb{C}^{N_t \times 1}$ is the uplink access signal of the kth device transmitted in the tth symbol duration. Here, $\alpha_k \in \{0, 1\}$ $\forall k \in \{1, 2, \ldots, K\}$, denotes the activity indicator, $s_{k,t} \in \mathbb{C}$ is the conventional modulated symbol, and $\mathbf{d}_{k,t} \in \mathbb{C}^{N_t \times 1}$ is the MBM vector which has unity on the index corresponding to the activated RF mirror and zeros elsewhere. Note that the total information bits are encoded in both the modulated symbol $s_{k,t}$ and the nonzero index of $\mathbf{d}_{k,t}$. Due to

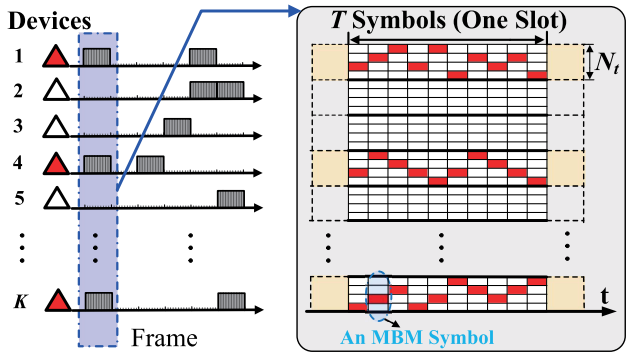


Fig. 10. Doubly structured sparsity of the access signal matrix in MBM-based GFMA scheme [96] ©IEEE.

the sporadic uplink traffic of devices and the characteristics of MBM, the signal matrix $\mathbf{X}$ exhibits *doubly structured sparsity*, as illustrated in Fig. 10. Specifically, in each column of $\mathbf{X}$, only the access signals of K_a active devices are nonzero, and all columns share the same device level sparsity. Moreover, the access signal of a specific active device $\mathbf{x}_{k,t}$ is also sparse, where only one entry of the MBM vector $\mathbf{d}_{k,t}$ is unity and the others are zero.

To exploit the doubly structured sparsity for improving JADD performance, a variety of CS recovery algorithms have been developed. Inspired by the idea of subspace matching pursuit, Ma et al. [91] proposed a greedy two-level structured sparsity (TLSS)-based detector for SM-based GFMA. Subsequently, a two-stage detection scheme was further developed, where a structured OMP (StrOMP) algorithm was proposed for activity detection and an SIC-based structured SP (SIC-SSP) algorithm was designed for the demodulation of the detected active devices [97], which will be denoted as StrOMP+SIC-SSP. The aforementioned works focus on only a single time slot in which the device activity remains constant over multiple successive symbol durations. Furthermore, Ma et al. [98] proposed a prior-information aided adaptive media modulation subspace matching pursuit (PIA-MSMP) algorithm to accommodate the dynamic device activity across different time slots within a frame. Different from the greedy CS recovery algorithms proposed in [91], [97], and [98], a doubly structured AMP (DS-AMP) algorithm was developed under the Bayesian framework, which further takes the a priori information of the finite constellations of $\mathbf{X}$ into account [96]. Moreover, the theoretical state evolution (SE) of the DS-AMP algorithm was derived to analyze its performance in [96]. Compared to single-antenna systems, the channel estimation in MBM-based mMIMO systems is much more challenging. Therefore, the GFMA schemes introduced in this section are mainly tailored for the IoT applications where the devices are fixed or have very low mobility, thus the CSI can be estimated accurately and it does not have to be updated frequently. In particular, the channel estimation issue for media modulation-based GFMA was also investigated in [96] and [98]. A brief summary of the aforementioned representative JADD algorithms for MBM-based GFMA is provided in Table II.

Fig. 11 provides an example to compare the JADD performance of the discussed algorithms as well as the traditional AMP algorithm [73], in terms of both AER and bit error rate (BER), where "SE of DS-AMP" represents the theoretical SE of the DS-AMP. Obviously, by fully exploiting the doubly structured sparsity and the a priori statistical information of $\mathbf{X}$, the DS-AMP algorithm significantly outperforms its counterparts that do not leverage the structured sparsity or do not leverage the a priori statistical information. Moreover, both AER and BER performance becomes better as the number of BS antennas increases, which verifies the superiority of mMIMO in MBM-based GFMA. Besides, the derived SE can accurately predict the performance of the DS-AMP algorithm, providing insightful guidance for optimizing practical system designs.

B. Joint Activity Detection and Channel Estimation

It should be noted that the GFMA cannot always be formulated as a JADD problem. This is because the JADD is based on the condition that the full CSI is available at the BS. In many IoT applications, such as smart traffic and wearable IoT, the channels between the devices and the BS may change frequently. In this context, it is unrealistic to assume that the full CSI is available at the BS, especially for mMIMO systems with a massive number of devices. Therefore, the frame structure with the format of "pilot + data" has recently been proposed, where each device is assigned with a unique nonorthogonal pilot sequence for joint activity detection and channel estimation (JADCE) at the BS. With the estimated active device set and the corresponding channels, the conventional coherent data detection is then executed based on the received data signal [10]. Specifically, the JADCE problem can be formulated as

$$\mathbf{Y} = \mathbf{PH} + \mathbf{N} \tag{2}$$

where $\mathbf{Y} \in \mathbb{C}^{P\times N_r}$ is the received pilot signal, $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_K] \in \mathbb{C}^{P\times K}$ is the pilot matrix, and $\mathbf{p}_k \in \mathbb{C}^{P\times 1}$ is the nonorthogonal pilot sequence of the kth device with the pilot length P, while $\mathbf{H} = [\alpha_1\mathbf{h}_1, \alpha_2\mathbf{h}_2, \ldots, \alpha_K\mathbf{h}_K]^T \in \mathbb{C}^{K\times N_r}$ is the massive access channel matrix, α_k is again the activity indicator of the kth device, and $\mathbf{N}$ is the AWGN. Considering the sporadic uplink traffic of devices, the JADCE problem (2) becomes an SMV CS problem in single-antenna systems, i.e., $N_r = 1$, and an MMV CS problem in MIMO systems, i.e., $N_r > 1$.

Schepker et al. [99] and Ahn et al. [100] proposed two CS-based JADCE schemes, respectively, for GFMA in single-antenna systems, where the OMP-based and EP-based CS recovery algorithms are developed, respectively. Furthermore, Liu and Yu [101] revealed that the error probability of device activity detection can be made arbitrary small by increasing the number of BS antennas. Based on this attractive finding, a large number of JADCE schemes have been proposed in single-station mMIMO systems [101], [102], [103], [104], [105], [106], [107], [108], [109], [110], and then naturally extended to more complicated cooperative mMIMO systems [111], [112], [113], [114], [115]. Following this line, we will first discuss the JADCE problem in single-station mMIMO systems and then extend it to cooperative mMIMO systems.

TABLE II
SUMMARY OF JADD ALGORITHMS FOR MBM-BASED GFMA SCHEME IN MMIMO SYSTEMS

JADD Scheme	CS Model	CS Algorithm	Sparsity Structure	Advances	Complexity
TLSS-based algorithm [91]	MMV	Greedy	Doubly structured sparsity	Exploit the doubly structured sparsity	$\mathcal{O}\left(K_a^3+N_r^2+N_t^3\right)$
StrOMP SIC-SSP algorithm [97]	MMV	Greedy	Doubly structured sparsity	Propose a two-stage detection scheme and employ SIC for enhanced performance	$\mathcal{O}\left(K_a^3\right)$
PIA-MSMP algorithm [98]	MMV	Greedy	Doubly structured sparsity and dynamic device activity	Further consider the dynamic device activity	$\mathcal{O}\left(TK_aN_rN_t+N_r^2+N_t^3\right)$
DS-AMP algorithm [96]	MMV	Bayesian	Doubly structured sparsity	Further exploit the statistical information of the access signal matrix	$\mathcal{O}\left(TKN_rN_t\right)$

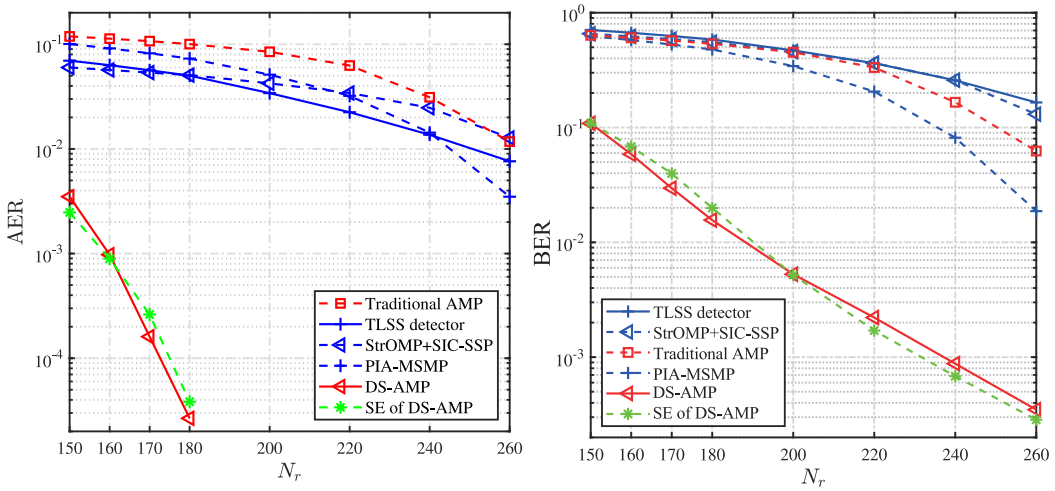


Fig. 11. Performance comparison of different JADD algorithms for MBM-based GFMA schemes, where $K = 500$, $K_a = 50$, $N_{\mathrm{RF}} = 2$, and $T = 12$ are considered: (a) AER performance and (b) BER performance [96] ©IEEE.

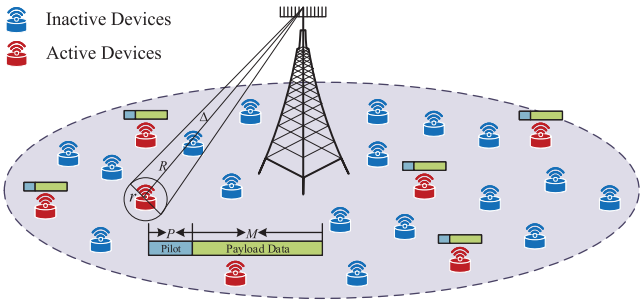


Fig. 12. System model of JADCE for GFMA in single-station mMIMO systems.

1) JADCE in Single-Station Massive MIMO Systems: A BS equipped with an N_r-elements uniform linear array (ULA) serves K single-antenna devices distributed in its coverage, where only K_a devices are activated in each frame duration, as illustrated in Fig. 12. A two-phase transmission scheme is adopted, with each frame consisting of a pilot phase and the subsequent payload data phase. Each device is assigned with a unique nonorthogonal pilot sequence that will be transmitted in the pilot phase. When accessing the network, the active devices directly transmit their uplink access signals with the format of pilot + data on the same time–frequency resources. In the pilot phase, given the received pilot signal $\mathbf{Y}$ and the preallocated pilot matrix $\mathbf{P}$, the JADCE problem is equivalent to estimating the sparse channel matrix $\mathbf{H}$ based on the MMV CS model of (2). Following this formulation, an OMP-based JADCE scheme was developed for GFMA in single-station mMIMO systems [102], and the sparsity of the delay-domain CIR was further leveraged to improve the channel estimation accuracy [103].

On the other hand, equipping a large number of antennas at the BS results in additional sparsity properties of the massive access channel matrix, which can be leveraged to further enhance JADCE performance. Specifically, the sporadic traffic of devices leads to the sparsity of the channel vector associated with each receive antenna, i.e., every column of $\mathbf{H}$ is sparse. Moreover, all the BS antennas observe a common sparsity pattern, which leads to the spatial-domain common sparsity of the channel matrix and facilitates the activity detection through nonzero row detection [10]. Based on the spatial-domain signal model (2), several efficient JADCE schemes were proposed [101], [104], [105], [106], [107], [108]. Specifically, in [101], a vector AMP algorithm was developed to exploit the common sparsity across different BS antennas, and the related probabilities of false alarm and miss detection were analyzed exploiting the SE. Adopting this vector AMP algorithm, each active device's uplink achievable rate was further characterized, based on which the length of the nonorthogonal pilot sequence was optimized in [104]. Afterward, Shao et al. [105] designed a three-phase transmission protocol for GFMA, which also employed the vector AMP algorithm and further considered the downlink transmission phase. The works [101], [104], [105] assume that the large-scale component of the channel fading coefficients are known to the BS. Considering a more practical scenario, an updated vector AMP algorithm was derived in [106], which takes the unknown large-scale fading parameters into account.

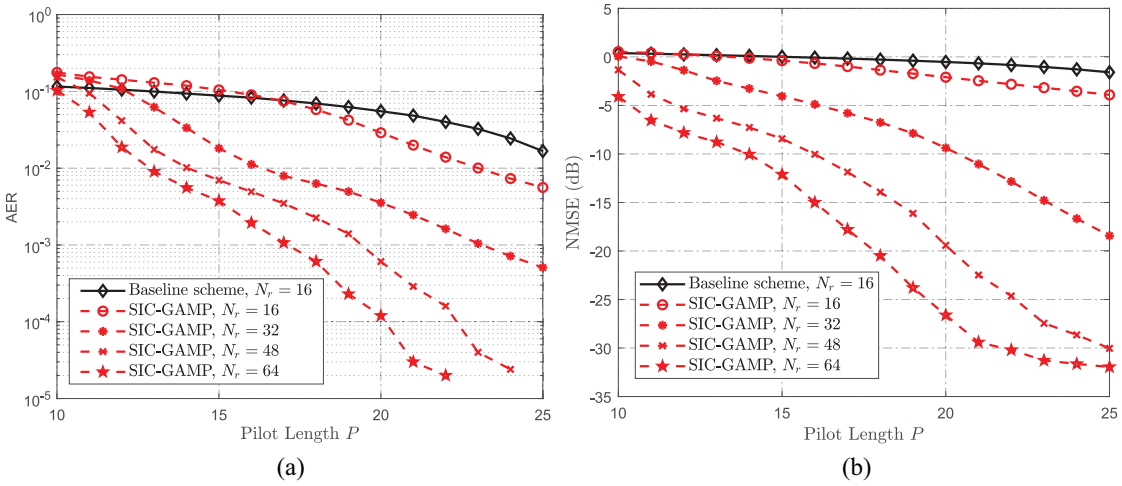


Fig. 13. Comparison JADCE performance of the SIC-GAMP [109] and the baseline scheme [101] for GFMA in single-station mMIMO systems: (a) AER performance and (b) NMSE performance.

Due to the large numbers of devices and BS antennas, the JADCE generally imposes a high computational complexity. To mitigate this problem, a dimension reduction-based JADCE scheme was proposed in [107], which projects the original channel matrix onto a low-dimensional space by jointly exploiting its sparse and low-rank structures. In addition, considering the time-varying device activity and CSI, the work [108] claimed that the inherent temporal correlation between adjacent time slots can be exploited to enhance the JADCE performance.

In [101], [104], [105], [106], [107], and [108], the reduction of the pilot overhead is limited to the number of active devices, i.e., $P \geq K_a$ is required, which becomes a severe obstacle for the implementation of these spatial-domain JADCE schemes. Further considering the typical one-ring channel model, the mMIMO channel vectors exhibit clustered sparsity in the angular domain [51]. Hence, the JADCE problem (2) can be reformulated as

$$\mathbf{R} = \mathbf{P}\widetilde{\mathbf{H}} + \widetilde{\mathbf{N}} \tag{3}$$

where $\mathbf{R} = \mathbf{Y}\mathbf{A}_R$ is the angular-domain received signal, $\widetilde{\mathbf{H}} = \mathbf{H}\mathbf{A}_R$ is the angular-domain channel matrix, $\widetilde{\mathbf{N}} = \mathbf{N}\mathbf{A}_R$, and $\mathbf{A}_R$ is the spatial-to-angular domain transformation matrix determined by the geometrical structure of the BS array. In contrast to the spatial-domain channel model (2), the angular-domain channel model (3) is more favorable to improve the accuracy of the CSI estimates of the identified active devices. Motivated by this angular-domain channel model, an SIC-based generalized AMP (GAMP) algorithm was proposed to jointly exploit the spatial-domain and angular-domain structured sparsities, where the pilot overhead can be far smaller than the number of active devices [109]. Considering that some devices may experience common local scattering clusters, a grouping-based JADCE scheme was proposed in [110].

A brief summary of the aforementioned JADCE schemes is provided in Table III. Fig. 13 verifies the effectiveness of the SIC-GAMP algorithm in the case of a single BS mMIMO. Here, $K = 500$, $K_a = 50$, and SNR $= 20$ dB are considered, and the numbers of BS antennas is set to $N_r = 16$, 32, 48, or 64. The state-of-the-art JADCE scheme [101] that only considers the spatial-domain model (2) is adopted as the baseline scheme. It is observed that by exploiting the angular-domain clustered sparsity of mMIMO channels, the SIC-GAMP scheme attains a significant performance improvement over the baseline scheme. Furthermore, the achievable performance of the SIC-GAMP scheme improves with the increase of the number of BS antennas. This is because increasing the number of BS antennas can simultaneously enhance the spatial-domain common sparsity of $\mathbf{H}$ and the angular-domain clustered sparsity of $\widetilde{\mathbf{H}}$, which improves the CS recovery performance. In particular, the SIC-GAMP scheme can reliably support GFMA even at an overloading ratio of 250 % (i.e., $P = 20$ and $K_a = 50$). With $N_r = 64$, the scheme achieves an AER of 10^{-4} and a normalized mean square error (NMSE) of -27 dB.

2) JADCE in Cooperative Massive MIMO Systems: For typical IoT applications, the power-limited devices are generally distributed in a vast area and, thus, multiple BSs should be densely deployed to offer an adequate coverage and save the transmit power of the devices. Adopting the traditional small-cell mMIMO networks, the reduced BS spacing however would inevitably introduce severe uplink intercell interferences, which is a limiting factor for reliable GFMA [111]. To overcome this limitation, various cooperative mMIMO networks have been intensively investigated for GFMA. Xu et al. [112] extended the JADCE problem of GFMA to the cloud radio access network (C-RAN), where the received signals from all the BSs are jointly processed at a central unit. Utkovski et al. [113] further considered the limited capacity of the backhaul links between the BSs and the central unit. Moreover, Chen et al. [114] studied the JADCE of GFMA in multicell systems, and compared the conventional noncooperative mMIMO network and the cooperative mMIMO network in terms of their effectiveness in overcoming intercell interferences.

TABLE III
SUMMARY OF JADCE SCHEMES FOR GFMA IN SINGLE-STATION MMIMO SYSTEMS

Reference	Channel Model	Advances	Complexity
L. Liu, *et al.* [101]	Spatial domain	Leverage the spatial-domain common sparsity across different BS antennas	$\mathcal{O}(PKN_r)$
L. Liu, *et al.* [104]	Spatial domain	Analyze the uplink achievable rate and optimize the pilot length	N/A
X. Shao, *et al.* [105]	Spatial domain	Propose a three-phase unified transmission design for GFMA	$\mathcal{O}(PKN_r)$
Z. Chen, *et al.* [106]	Spatial domain	Consider the unknown large-scale fading parameter	$\mathcal{O}(PKN_r)$
X. Shao, *et al.* [107]	Spatial domain	Propose a dimension-reduced algorithm to reduce the computational complexity	$\mathcal{O}(PN_r r_e + r_e^3)$
J. Jiang, *et al.* [108]	Spatial domain	Consider the time-varying device activity and CSI	$\mathcal{O}(PKN_r)$
M. Ke, *et al.* [109]	Spatial domain & Angular domain	Leverage both the spatial-domain common sparsity and the angular-domain clustered sparsity	$\mathcal{O}(PKN_rN_c)$
J. Jiang, *et al.* [110]	Angular domain	Leverage the fact that some devices may have common local scattering cluster	$\mathcal{O}(PK_g u_g)$

Notes: r_e is the rank for dimension reduction, N_c is the number of subcarriers, K_g is he number of devices in group g, and u_g is the sparsity of sparsity level of angular-domain channel matrix.

TABLE IV
SUMMARY OF JADCE SCHEMES FOR GFMA IN COOPERATIVE MMIMO SYSTEMS

Reference	Network Architecture	Advances	Complexity
X. Xu, *et al.* [112]	C-RAN	Extend the JADCE problem to the C-RAN	$\mathcal{O}(P^2 + N_r^2)$
Z. Utkovski, *et al.* [113]	C-RAN	Consider the limited capacity of the backhaul links	$\mathcal{O}(PKN_rN_p)$
Z. Chen, *et al.* [114]	Multi-cell massive MIMO and cooperative massive MIMO	Compare the JADCE performance of multi-cell massive MIMO and cooperative massive MIMO	$\mathcal{O}(PKN_rN_p)$
M. Ke, *et al.* [115]	Cell-free massive MIMO	Propose cloud computing and edge computing paradigms for signal processing	$\mathcal{O}(PKN_rN_cN_p)$

Notes: N_c is the number of subcarriers and N_p is the number of APs.

Among various cooperative mMIMO networks, cell-free mMIMO networks are the most popular ones and have attracted ever increasing attention from both the academic and industrial communities [111]. In fact, cell-free mMIMO is an incarnation of the general idea of distributed MIMO, network MIMO, C-RAN, and distributed antenna systems, where a large number of access points (APs) cooperate with each other in the network to serve a large area. The APs equipped with a large number of antennas are connected to the related processing units via fronthaul links for joint signal processing, as illustrated in Fig. 14. In this context, the intercell interference can be effectively avoided, as cells and cell boundaries do not exist. Moreover, by performing coherent signal processing across geographically distributed APs' antennas, cell-free mMIMO can provide a uniformly good service for all devices. In contrast, for centralized mMIMO systems with the receive antennas locating at BSs, the devices in the cell center generally reaps a better service quality than the devices in the cell edge due to the heterogeneous path-loss effect [112]. Besides, equipping massive antennas at the APs further combines the distributed MIMO and mMIMO concepts, which is expected to reap all the benefits from these two systems.

Based on the notion of cell-free mMIMO networks, two different signal processing paradigms, namely, cloud computing of Fig. 14(a) and edge computing of Fig. 14(b), have been proposed for supporting centralized and distributed AP cooperation, respectively [115]. For cloud computing, the signals received at all the APs are centrally processed in a central processing unit (CPU). Since the APs are only designed to work as relays with simple signal processing capabilities only, the required cost for large scale deployment of APs is significantly reduced. However, cloud computing requires the information to pass through several subnetworks including the radio access network, backhaul network, and core network, where traffic control, routing, and other network-management operations can contribute to excessive delays. As for edge computing, the central processing is offloaded to some of the APs equipped with distributed processing units (DPUs) such that the corresponding computations can be performed in a distributed manner. Compared to cloud computing, edge computing can alleviate the burden on the fronthaul links and the CPU, facilitate a faster access response as well as support more efficient AP cooperation, at the expense of the increased cost in network deployment [115].

For GFMA in cell-free mMIMO systems, the APs transfer the pilot signals received from all the active devices to the related processing unit. Based on the MMV CS models in (2) and (3), the processing unit can perform JADCE by jointly processing the received signals from multiple APs using for example the SIC-GAMP algorithm [109]. Table IV summarizes the JADCE schemes in cooperative mMIMO systems. In Fig. 15, to verify the superiority of cell-free mMIMO-based IoT networks, a conventional noncooperative multicell mMIMO architecture is compared as the benchmark, where each BS only serves its own cell's devices and treats the intercell interference as noise. Here, we assume that $K = 2800$ devices are uniformly distributed in the network having a radius of $R_{\text{net}} = 2.65$ km and $B = 7$ APs are geographically distributed to serve these devices. The AP-to-AP distance is $d = \sqrt{3}$ km, the number of active devices is $K_a = 140$, the number of AP antennas is $N_r = 16$, and the number of cooperating APs at each DPU is N_{co}. The SIC-GAMP-base JADCE scheme [109] is employed by the both systems. As shown in Fig. 15, the cell-free mMIMO network achieves much better AER and NMSE performance than the traditional noncooperative multicell mMIMO network architecture. It can also be seen that by increasing the number of APs

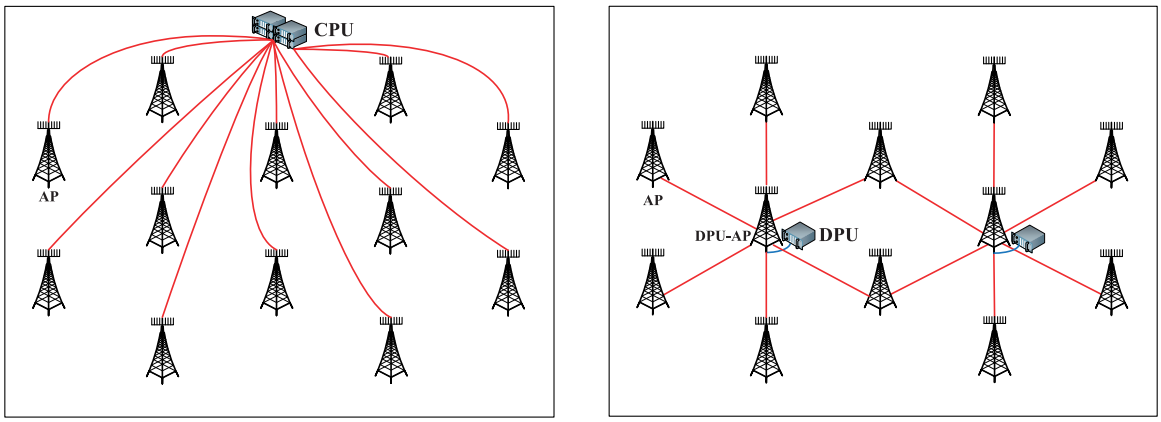


Fig. 14. Two processing paradigms in cell-free mMIMO systems: (a) cloud computing and (b) edge computing [115] ©IEEE.

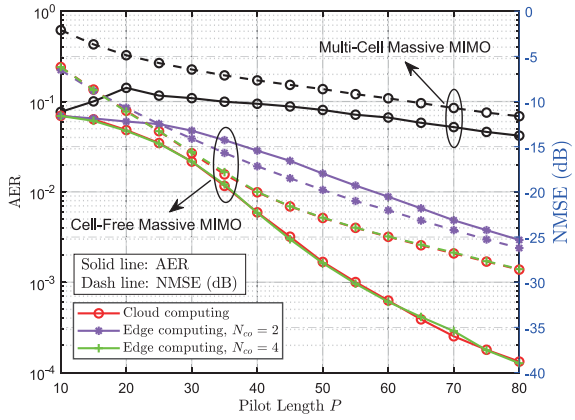


Fig. 15. JADCE performance of GFMA in multicell noncooperative mMIMO and cell-free mMIMO systems. The both systems adopt the SIC-GAMP-based JADCE scheme [109].

for cooperation N_{co}, the performance of edge computing approaches that of cloud computing. In particular, we observe that only $N_{co} = 4$ APs are required for edge computing to obtain almost the same performance as cloud computing. This is because the signals received at the APs far away from a device are approximately zero due to the severe path loss, and incorporating them can hardly improve the JADCE performance further.

C. Noncoherent Data Detection

The aforementioned GFMA schemes adopt a coherent data detection framework, where the detection performance is highly dependent on the accuracy of the CSI estimate. These solutions become inefficient or even impractical in high-mobility communications scenarios with small data packets, since the devices have to frequently transmit nonorthogonal pilot sequences for the CSI update. To address this limitation, two noncoherent detection frameworks were introduced, where the payload data of active devices is directly detected from the overlapped received signal, without any knowledge of the full CSI.

1) Common Codebook-Based Noncoherent Detection: We first consider the unsourced random access scenarios, where the BS is solely interested in the list of transmitted messages without regard to their individual sources. In practice, the unsourced random access is motivated by the content-oriented IoT applications [116]. For example, in the quality inspection process of smart factories, many sensors are distributed at different positions on the production line to acquire the quality of products. The server only concerns about the weighted average of these sensors' quality information and does not have to know the identities of sensors that generate it. Focusing on unsourced random access, an efficient common codebook-based noncoherent detection (CCND) framework was developed in [117], [118], [119], [120], [121], [122], [123], [124], and [125].

In the CCND framework, all the potential devices share a common codebook hardwired at the moment of production. When accessing the network, a specific active device k first maps its B information bits into an integer $b \in \{1, 2, \ldots, 2^B\}$, then the bth codeword $\mathbf{c}_b \in \mathbb{C}^{L\times 1}$ of the common codebook $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_{2^B}] \in \mathbb{C}^{L\times 2^B}$ is directly transmitted to the BS, where L is the length of codewords. Define a set of $K2^B$ Bernoulli random variables $\{\delta_{k,b} | k = 1, \ldots, K; b = 1, \ldots, 2^B\}$ to model the device activity and codeword selection behavior. Specifically, $\delta_{k,b} = 1$ if the kth device is active and it selects the bth codeword to transmit; and $\delta_{k,b} = 0$ otherwise. Therefore, the overlapped received signal at the BS is expressed as

$$\widehat{\mathbf{Y}} = \sum_{k=1}^{K}\sum_{b=1}^{2^B} \mathbf{c}_b \delta_{k,b} \mathbf{h}_k^T + \widehat{\mathbf{N}} = \mathbf{C}\boldsymbol{\Delta}\mathbf{H} + \widehat{\mathbf{N}} \tag{4}$$

where $\boldsymbol{\Delta} = [\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_K] \in \mathbb{B}^{2^B\times K}$ is the codeword selection matrix with $\boldsymbol{\delta}_k = [\delta_{k,1}, \ldots, \delta_{k,2^B}]^T \in \mathbb{B}^{2^B\times 1}$, and $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_K]^T \in \mathbb{C}^{K\times N_r}$ is the massive-access channel matrix. Note that the transmitted information is encoded in the nonzero indices of the codeword selection matrix $\boldsymbol{\Delta}$. Besides, $\boldsymbol{\Delta}$ contains only K_a nonzero rows each of which having a single nonzero entry. Therefore, the data detection is equivalent to detecting the nonzero row indices of $\boldsymbol{\Delta}$ based on $\widehat{\mathbf{Y}}$ and the known $\mathbf{C}$, which can be formulated as a MMV CS problem of (1) by combining the codeword

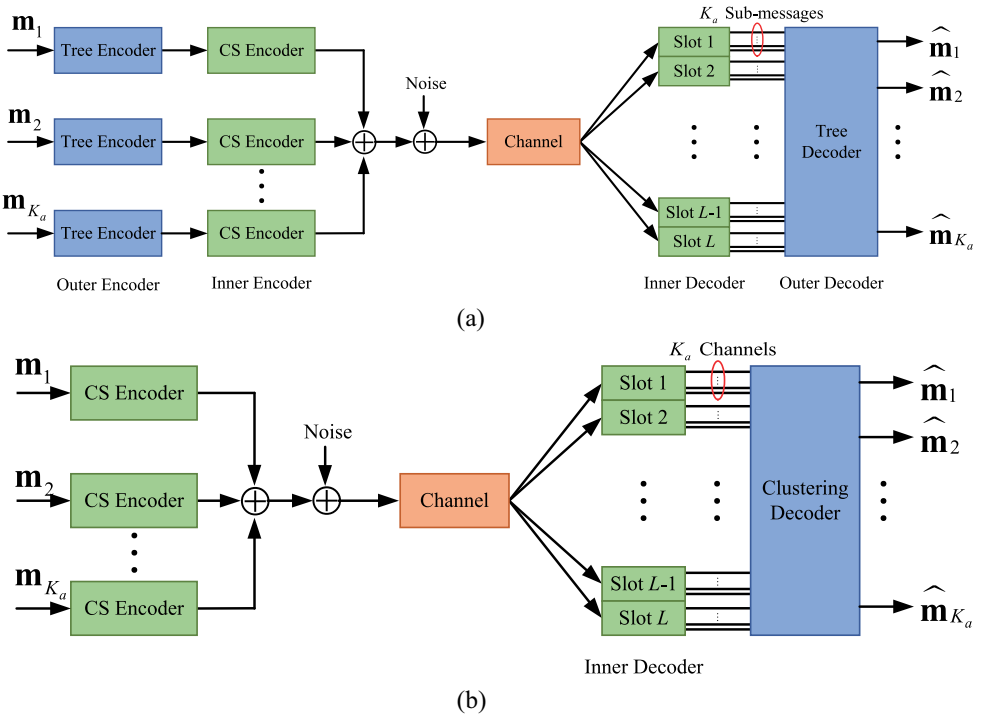


Fig. 16. CCND for GFMA: (a) coupled CS-based transmission scheme and (b) uncoupled CS-based transmission scheme for mMIMO systems [132] ©IEEE.

selection matrix and the massive-access channel matrix as a sparse matrix to be estimated, i.e., $\mathbf{X} = \mathbf{\Delta H} \in \mathbb{C}^{2^B \times N_r}$ with $N = 2^B$ and $Q = N_r$, while defining the sensing matrix as $\Psi = \mathbf{C}$ with $M = L$. In particular, each active device contributes a single nonzero coefficient in a specific column of $\mathbf{X}$, hereby resulting in a K_a-sparse 2^B-dimensional vector. Moreover, different columns, i.e., different receive antennas, have a common sparsity pattern. This structured sparsity can be exploited for enhancing CS recovery performance [116].

However, since the number of codewords scales exponentially with the message length, the computational complexity of the CS recovery is prohibitive even for the case with dozens of information bits. This becomes a major obstacle for practical implementation of CCND. To overcome this obstacle, Amalladinne et al. [126] introduced a coupled CS transmission scheme for reduced complexity, as shown in Fig. 16(a). The scheme consists of outer and inner encoders/decoders. For the outer encoder, each active device's message is nonuniformly divided into multiple submessages and the redundancy check bits are added in these submessages to form the subblocks having uniform length. Then, the inner encoder is employed to map the subblocks to the codewords in a common codebook, which are transmitted in their corresponding slots. At the BS, the inner decoder first recovers the lists of submessages for all the slots and the submessages are then stitched together by the outer decoder using the redundancy check bits. Afterward, the detection performance is further enhanced by passing information between the CS-based inner decoder and the outer decoder dynamically [127]. Besides, the sparse regression code was introduced to reduce the size of codebook, where the submessages are encoded by the structured linear combination of the columns of the common codebook [128].

The prior works [126], [127], [128] are limited to single-antenna systems. As a remedy, Fengler et al. [129] investigated the CCND problem in mMIMO systems and revealed that the transmit power per bit can be made arbitrary small if the number of BS antennas is sufficiently large. By equipping massive antennas at the BS, a low-complexity covariance-based CS recovery algorithm was developed for the implementation of an inner decoder [130]. Besides, a tensor-based transmission scheme was proposed for block fading channels in mMIMO systems [131]. In addition, the problem was extended to the cell-free mMIMO systems, where the detection performance can be further improved by considering the cooperation of geographically distributed APs [116].

Although the coupled CS-based schemes significantly reduce the decoding complexity, the spectral efficiency is scarified due to the employment of redundancy check bits. To avoid the loss, Shyianov et al. [132] proposed that the strong correlation between the reconstructed MIMO channels in different slots provided enough information to stitch the submessages and devised an uncoupled CS-based scheme by removing the outer encoder and decoder. Specifically, the message is uniformly divided into multiple submessages, which are directly transmitted in slot-wise using the common codebook-based encoder, as illustrated in Fig. 16(b). At the BS, with the recovered submessages, an expectation maximization-based clustering algorithm is designed to obtain the original message. Since no redundancy is introduced, the spectral efficiency is dramatically improved. For uncouple CS-based schemes, however, due to the small number of information bits of submessages and the absence of check bits, the collision, i.e., multiple active devices select the same codeword, becomes a new challenge. To address this issue, the

TABLE V
SUMMARY OF CCND SCHEMES FOR UNSOURCED GFMA

Reference	BS Architecture	Advances	Complexity
Y. Polyanshiy [117]	Single-antenna	The first work of a CCND framework for unsourced random access	$\mathcal{O}\left(LK2^B\right)$
O. Ordentlich, *et al.* [118]	Single-antenna	The first low-complexity CCND scheme	$\mathcal{O}\left(2^B\right)$
V. K. Amalladinne, *et al.* [126]	Single-antenna	A coupled CS-based CCND scheme for reduced complexity	$\mathcal{O}\left(L2^{B_{\text{sub}}}\right)$
A. Fengler, *et al.* [128]	Single-antenna	Leverage sparse regression code to reduce the codebook size	$\mathcal{O}\left(L2^{B_{\text{sub}}}\right)$
A. Fengler, *et al.* [129]	Massive MIMO	Extend the problem to the massive MIMO systems	$\mathcal{O}\left(LN_r2^{B_{\text{sub}}}\right)$
S. Haghighatshoar, *et al.* [130]	Massive MIMO	A covariance-based CS recovery algorithm to reduce the complexity of inner decoder	$\mathcal{O}\left(2^{B_{\text{sub}}}\right)$
A. Decurninge, *et al.* [131]	Massive MIMO	A tensor-based transmission scheme for CCND in massive MIMO systems	$\mathcal{O}\left(2^{K_a B_{\text{sub}}}\right)$
V. Shyianov, *et al.* [132]	Massive MIMO	An uncoupled CS-based CCND scheme for massive MIMO systems	$\mathcal{O}\left(K_a^3\right)$
X. Xie, *et al.* [133]	Massive MIMO	Leverage the diversity of devices' AoAs to resolve the collisions	$\mathcal{O}\left(LN_r2^{B_{\text{sub}}}\right)$

Notes: B_{sub} is the length of sub-blocks.

work [133] exploited the diversity of different devices' Angles of Arrival (AoA) to resolve the collision and leveraged the angular-domain sparsity of mMIMO channels to improve the detection performance.

A brief summary of the aforementioned CCND schemes are provided in Table V.

2) Individual Codebook-Based Noncoherent Detection: Despite of its many advantages, CCND is limited to unsourced random access scenarios. In practice, most IoT applications rely on sourced random access, where the server concerns both the transmitted messages and the identities (IDs) of active devices that generate them. By making several minor modifications to the packet structure, a straightforward solution is to embed the device ID in the transmitted information bits, and then to map the combined device ID and payload data onto a codeword in the common codebook [134]. To identify K devices, a device ID sequence with at least $\lceil \log_2(K) \rceil$ bits is embedded, where the operator $\lceil \cdot \rceil$ rounds a real number to the nearest integer larger or equal to it. Hence, the payload efficiency is significantly degraded, especially for the scenarios with massive number of devices and small data packets.

Recently, an individual CCND scheme was proposed to support sourced GFMA more efficiently [135]. In this scheme, each device is allocated with an individual codebook, i.e., $\widetilde{\mathbf{C}}_k = [\tilde{\mathbf{c}}_{k,1}, \tilde{\mathbf{c}}_{k,2}, \ldots, \tilde{\mathbf{c}}_{k,2^B}] \in \mathbb{C}^{L\times 2^B}$, to convey B-bit information, where L is the length of codewords. Based on the B information bits to be conveyed, each active device first select a codeword from its individual codebook, and then transmit the codeword on one subcarrier of L successive OFDM symbols. In this context, a total of $\widetilde{N}2^B$ bits can be transmitted adopting OFDM with $\widetilde{N}$ subcarriers. Defining L successive OFDM symbols as a subframe, the device activity during a frame of J subframes and the CSI within a subframe are assumed to be invariant. At the BS, the signal received at the $\tilde{n}$th subcarrier in the jth subframe is expressed as

$$\widetilde{\mathbf{Y}}_{\tilde{n}}^j = \sum_{k=1}^{K} \widetilde{\mathbf{C}}_k \widetilde{\mathbf{X}}_{k,\tilde{n}}^j + \mathbf{N}_{\tilde{n}}^j = \widetilde{\mathbf{C}}\widetilde{\mathbf{X}}_{\tilde{n}}^j + \mathbf{N}_{\tilde{n}}^j \tag{5}$$

where $\widetilde{\mathbf{X}}_{k,\tilde{n}}^j = \alpha_k \mathbf{e}_{k,\tilde{n}}^j (\mathbf{h}_{k,\tilde{n}}^j)^T \in \mathbb{C}^{2^B\times N_r}$ is the equivalent channel matrix, α_k, $\mathbf{e}_{k,\tilde{n}}^j \in \mathbb{B}^{2^B\times 1}$, and $\mathbf{h}_{k,\tilde{n}}^j \in \mathbb{C}^{N_r\times 1}$ are the activity indicator, codeword selection vector, and MIMO channel vector, respectively. Furthermore, $\widetilde{\mathbf{C}} = [\widetilde{\mathbf{C}}_1, \ldots, \widetilde{\mathbf{C}}_K] \in \mathbb{C}^{L\times K2^B}$ and $\widetilde{\mathbf{X}}_{\tilde{n}}^j = [(\widetilde{\mathbf{X}}_{1,\tilde{n}}^j)^T, \ldots, (\widetilde{\mathbf{X}}_{K,\tilde{n}}^j)^T]^T \in \mathbb{C}^{K2^B\times N_r}$. Combining $\widetilde{N}$ subcarriers and J subframes, the noncoherent data detection can be formulated as a MMV CS problem of (1), where the measurement matrix is given as $\mathbf{Y} = [\widetilde{\mathbf{Y}}_1^1, \ldots, \widetilde{\mathbf{Y}}_{\widetilde{N}}^J] \in \mathbb{C}^{L\times J\widetilde{N}N_r}$ with $M = L$ and $Q = J\widetilde{N}N_r$, the sensing matrix is expressed as $\Psi = \widetilde{\mathbf{C}}$ with $N = K2^B$, and the sparse channel matrix to be estimated is expressed as $\mathbf{X} = [\widetilde{\mathbf{X}}_1^j, \ldots, \widetilde{\mathbf{X}}_{\widetilde{N}}^J] \in \mathbb{C}^{K2^B\times J\widetilde{N}N_r}$. Here, both the device activity and transmitted information are encoded in the nonzero row indexes of $\mathbf{X}$. Moreover, $\mathbf{X}$ exhibits structured sparsity in the space, time, and frequency domains, as illustrated in Fig. 17, where the common sparsity can be observed at different receive antennas and subcarriers within a subframe, and different subframes have an approximate common sparsity pattern.

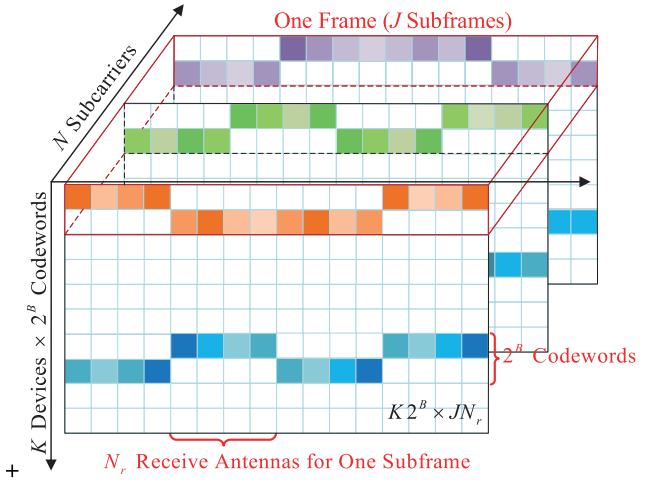


Fig. 17. Space–time–frequency structured sparsity of individual CCND for GFMA in mMIMO-OFDM systems [135] ©IEEE.

Qiao et al. [135] proposed an AMP-based space–time–frequency joint activity and blind information detection (STF-JABID) algorithm to leverage the space–time–frequency structured sparsity of $\mathbf{X}$ for improving AER and BER performance. Fig. 18 demonstrates the superiority of the proposed algorithm [135], where the state-of-the-art SOMP algorithm [82] and generalized MMV-AMP (GMMV-AMP) algorithm [109] without taking the space–time–frequency structured sparsity into account are used as the benchmarks for comparison. Here, it is assumed that the number of devices is $K = 100$ with $K_a = 10$ active devices, $\widetilde{N} = 512$, $B = 1$, $J = 2$, and $N_r = 2$. It is clear that the proposed

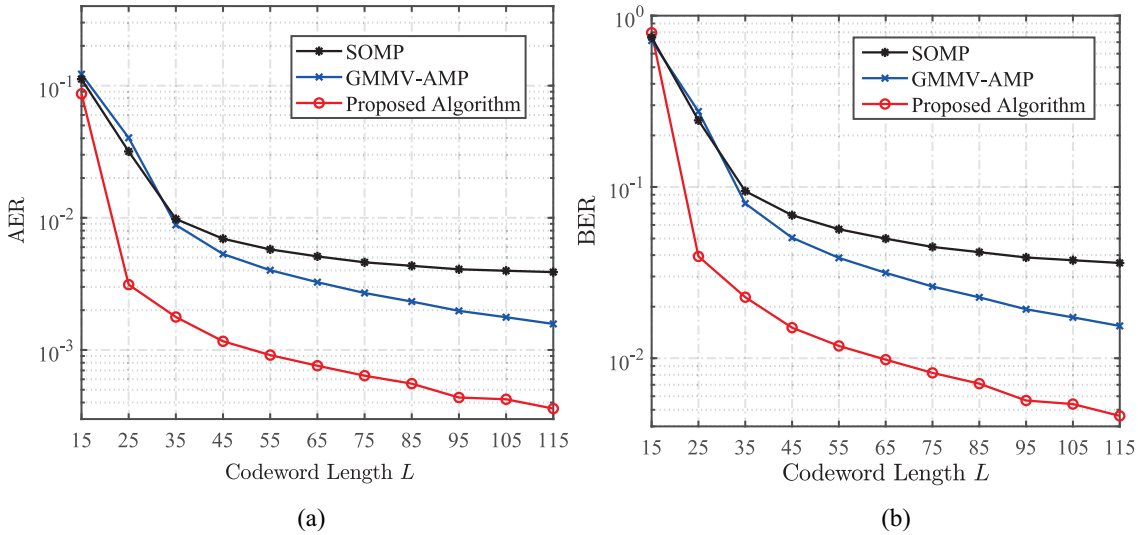


Fig. 18. (a) AER and (b) BER performance comparison of individual CCND schemes for GFMA in mMIMO-OFDM systems [135] ©IEEE.

algorithm [135] significantly outperforms the two benchmarks by fully exploiting the structured sparsity of the equivalent channel matrix.

V. Future Research Directions

Although extensive research efforts have been made to accelerate the development of the CS-based GFMA paradigm, numerous practical challenging issues still remain open to be resolved. In this section, we discuss some future research directions to address the key challenges in implementing the CS-based GFMA paradigm for the 6G massive communication.

A. Practical Hardware Constraints

As reviewed in the previous sections, most GFMA schemes consider ideal hardware assumptions, such as fully digital MIMO architecture at the BS, infinite-resolution ADCs, perfect synchronization between the devices and the BS, equal amplitudes and phases between in-phase (I) and quadrature (Q) branches, i.e., I/Q balance, etc. Due to the resulting high hardware cost and huge power consumption, the GFMA in mMIMO systems should be investigated under the more practical hybrid MIMO architecture [136], [137] and low-resolution ADCs [138]. Moreover, due to the imperfect hardware, the carrier frequency offset caused by the asynchronization of the oscillators between the devices and the BS [139], [140] as well as the I/Q imbalance [141] should be further incorporated into the CS-based GFMA paradigm. Considering these practical hardware constraints, the problem formulations and receive algorithms presented in the previous sections are not directly applicable and, thus, the corresponding GFMA schemes have to be redesigned.

B. GFMA in Space-Air-Ground-Sea Integrated Networks

Most existing works usually implement various IoT applications in the terrestrial cellular networks. However, the 6G ubiquitous connectivity is expected to rely on the space-air-ground-sea integrated networks (SAGEINs). Due to the inherent limitations of terrestrial infrastructures, it is impractical or uneconomic for deploying terrestrial BSs to seamlessly integrate the devices distributed across the ground, ocean, and air [142]. As supplements to terrestrial networks, nonterrestrial networks (NTNs), including satellite constellations at different Earth orbits, high-altitude platform (HAP) networks, and unmanned aerial vehicle (UAV) networks, can offer effective coverage to remote areas where terrestrial BSs are unavailable. To connect the different layers in hierarchical SAGEINs, an efficient GFMA scheme design is required, though it involves numerous challenging issues, such as the seamless integration of heterogeneous networks [142], efficient cooperation between different networks [143], channel modeling for aerial communication links [144], high ground-to-space path loss [145], etc. Therefore, considerable research efforts should be directed to address these challenging issues in the deployment of GFMA in SAGEINs.

C. Deep-Learning-Enhanced Design

Due to the remarkable accomplishments demonstrated by deep learning in various fields, such as computer vision, natural language processing, and image recognition, it is expected to serve as one of the primal driving forces to propel the advancement of 6G. Recently, deep learning has shown its huge potentials in resource allocation, signal processing, channel estimation, and transceiver design for wireless communication systems [146], [147], [148]. In particular, deep learning can fully leverage the implicit information in the available data and the benefits of well-developed wireless communication models, to reap a better performance or a lower complexity compared with the traditional design approaches. For GFMA in mMIMO systems, the massive numbers of devices and BS antennas make the problem dimension extremely large, where the implementation of the aforementioned CS-based GFMA schemes would result in a high complexity. The deep learning-enhanced design is expected to provide a low-complexity and better-performance alternative. For example, Zhang et al. [149] proposed a deep

neural network (DNN)-aided message passing-based block SBL algorithm to solve the JADCE problem of GFMA, which could approach the lower NMSE bound. Meanwhile, Kim et al. [150] proposed a DNN-based activity detection scheme for GFMA. Considering the imperfect CSI, a DNN-based on variational autoencoder is further developed for activity detection in GFMA [151]. Yu et al. [152] proposed a JADCE neural network, which fully exploits the information contained in the received preamble and data signals for improved performance. Moreover, Bai et al. [153] developed a machine learning framework, where the information distilled from the initial data recovery phase are utilized to further enhance channel estimation, which in turn improves data recovery performance.

However, the analytical framework and the generalization of the deep learning-based approaches pose the new challenges in their practical implementation to GFMA in mMIMO systems. To address these issues, the model-driven deep learning framework can be adopted, where the well-developed wireless communication models and signal processing techniques are exploited to design the training network, leaving only a few key parameters that need to be trained [146], [154].

D. GFMA for High-Speed IoT Applications

Most existing works assume that the device activity and CSI remain unchanged during the considered time interval, e.g., [104], [105], [106], [107], and [108]. In addition, the devices are assumed to be perfectly synchronized. To support mobile IoT applications, such as smart traffic and UAV communications in 6G, a more complicated massive connectivity scenario should be further investigated, where the devices are moving at a high speed and, thus, result in fast time-varying channels. Furthermore, the devices can randomly access or leave the network, which leads to the time-varying device activity and the asynchronous transmission between the devices [155], [156]. In this context, the existing CS-based GFMA schemes will fail to work, and the transceiver should be redesigned to capture the variation of the channels and to handle the asynchronous transmission problem. Here, the temporal correlation of the channels can be exploited for improving multidevice detection performance [156].

E. Joint Activity Detection,Channel Estimation, and Data Decoding

Most previous works assumed a two-phase processing, i.e., JADCE and data decoding. There is another line of research, i.e., joint activity detection, channel estimation, and data decoding (JADCEDD), where partially detected data can be used as soft pilots to enhance the channel estimation accuracy in an iterative fashion. In the field, Li et al. [157] proposed a belief propagation-based-joint device detection, channel estimation and data decoding algorithm for unsourced massive access, in which the CE results can be enhanced by regarding the corrected decoded data in the LDPC phase as extra pilots to execute the second CE. Similarly, Bian et al. [158] achieved the JADCEDD by BiG-AMP algorithm in a turbo receiver that can effectively exploit the common sparsity pattern in the received pilot and data signal, and improve the data detection performance by incorporating with channel decoder. Furthermore, Di Renna and de Lamare proposed a bilinear message-scheduling GAMP algorithm for JADCEDD in a grant-free mMIMO scenario. By applying the activity detection results or the residual for the message to determine the update of messages, they introduced two message-scheduling techniques to reduce the computational cost while maintaining the detection performance. Besides, Zhou et al. [160] extended this topic to the NOMA-OTFS system in LEO satellite

TABLE VI
LIST OF MAJOR ABBREVIATIONS

Abbreviations	Meanings
ADC	Analog-to-digital converter
AER	Activity error rate
AMP	Approximate message passing
AoA	Angle of arrival
AP	Access point
CCND	Common Codebook-based non-coherent detection
CDMA	Code-domain multiple access
CIR	Channel impulse response
CPU	Central processing unit
C-RAN	Cloud radio access network
CS	Compressive sensing
CSI	Channel state information
DMRS	Demodulation reference signal
DPU	Distributed processing unit
DS	Dense spreading
DS-AMP	Doubly structured AMP
EC-GSM	Extended coverage global system for mobile communications
EDT	Early data transmission
eDRX	Extended discontinuous reception
EM	Expectation maximization
FSRA	Four-step random access
GAMP	Generalized AMP
GFMA	Grant-free massive access
GMMV-AMP	Generalized multiple measurement vector AMP
GPRS	General packet radio service
JADCE	Joint activity detection and channel estimation
JADD	Joint activity and data detection
LASSO	Least absolute shrinkage and selection operator
LDS	Low-density spreading
LPWAN	Low-power wide-area network
MBM	Media-based modulation
MIMO	Multi-input multi-output
mMTC	Massive machine-type communications
MPA	Message passing algorithm
MUD	Multiuser detection
MUSA	Multiuser shared access
MUT	Multiuser transmit
NB-IoT	Narrow-band Internet-of-Things
NOMA	Non-orthogonal multiple access
OAMP-ASL	Orthogonal AMP with accurate structure learning
OMA	Orthogonal multiple access
OMP	Orthogonal matching pursuit
PIA-ASP	Prior-information aided adaptive subspace pursuit
PIA-MSMP	Prior-information aided adaptive media modulation subspace matching pursuit
PRACH	Physical random access channel
PSM	Power-saving mode
PUR	Pre-configured uplink resources
PUSCH	Physical uplink shared channel
RFID	Radio frequency identification
GAGEIN	Space-air-ground-sea integrated network
SCMA	Sparse code multiple access
SE	State evolution
SISD	Structured iterative support detection
SIC-SSP	Successive inference cancellation based structured SP
SM	Spatial modulation
SMV	Single measurement vector
SP	Subspace pursuit
TLSS	Two-level structured sparsity
TSRA	Two-step random access

IoT, achieving good detection performance by overcoming the long round-trip latency and severe Doppler effect. In, general, the JADCEDD is more practical than the JADD and usually exhibits better performance than the JADCE. However, this advantage is obtained at the expense of computational complexity, such as introducing more iterations. How to strike a tradeoff is needed to be considered in the future research.

VI. Conclusion

The future 6G massive communication is expected to require instant and seamless wireless connectivity for extremely large numbers of devices, which is a key enabler of the digital transformation of many aspects of society. Thanks to the recently completed infrastructures and well-developed technologies, the existing cellular networks can serve as a solid foundation for implementing massive connectivity in practice. This review has explored various typical IoT use cases and their service requirements. Moreover, the state-of-the-art IoT standards and the random access solutions from the both industry and academic communities have been reviewed. In particular, we have pointed out the limitations of the existing random access solutions, which do not take into account the inherent sparse communication behavior of massive communication. Against this background, a CS-based GFMA paradigm has been introduced, where the active devices directly access the network without any scheduling, and the activity detection, channel estimation, and/or data detection at the BS can be formulated as an SMV/MMV CS problem. Under the CS-based GFMA paradigm, various network architectures, transmission schemes, data detection frameworks, and receive algorithms can be flexibly incorporated to meet different service requirements of heterogeneous IoT applications. In this respect, we have detailed the roadmap with evolutions from single-antenna to large-scale antenna array-based BSs, from single-station to cooperative mMIMO systems, and from unsourced to sourced random access scenarios. Finally, we have discussed the key challenges and open issues to provide enlightening guidance for future research directions.

Appendix

A list of major abbreviations used in this article is provided in Table VI.

References

[1] X.-H. You et al., "Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts," *Sci. China Inf. Sci.*, vol. 64, no. 1, pp. 1–74, Jan. 2021.

[2] O. Vermesan and J. Bacquet, *Next Generation Internet of Things: Distributed Intelligence at the Edge and Human Machine-to-Machine Cooperation*. Aalborg, Denmark: River Publ., 2019.

[3] B. Guo, Y. Liu, S. Liu, Z. Yu, and X. Zhou, "CrowdHMT: Crowd intelligence with the deep fusion of human, machine, and IoT," *IEEE Internet Things J.*, vol. 9, no. 24, pp. 24822–24842, Dec. 2022.

[4] Z. Yang, B. Liang, and W. Ji, "An intelligent end–edge–cloud architecture for visual IoT-assisted healthcare systems," *IEEE Internet Things J.*, vol. 8, no. 23, pp. 16779–16786, Dec. 2021.

[5] N. H. Mahmood et al., "White paper on critical and massive machine-type communication towards 6G," 6G Res. Vis. no. 11, Univ. Oulu, Oulu, Finland, White Paper, Jun. 2020.

[6] A. Ikpehai et al., "Low-power wide area network technologies for Internet-of-Things: A comparative review," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2225–2240, Apr. 2019.

[7] "The ITU-R framework for IMT-2030," Int. Telecommun. Union, Geneva, Switzerland, document ITU-R WP 5D, Jul. 2023.

[8] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward massive machine-type cellular communications," *IEEE Wireless Commun.*, vol. 24, no. 1, pp. 120–128, Feb. 2017.

[9] C. Bockelmann et al., "Towards massive connectivity support for scalable mMTC communications in 5G networks," *IEEE Access*, vol. 6, pp. 28969–28992, 2018.

[10] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the Internet-of-Things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88–99, Sep. 2018.

[11] F. Guo, F. R. Yu, H. Zhang, X. Li, H. Ji, and V. C. M. Leung, "Enabling massive IoT toward 6G: A compressive survey," *IEEE Internet Things J.*, vol. 8, no. 15, pp. 11891–11915, Aug. 2021.

[12] B. S. Khan, S. Jangsher, A. Ahmed, and A. Al-Dweik, "URLLC and eMBB in 5G industrial IoT: A survey," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 1134–1163, 2022.

[13] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, 3rd Quart., 2018.

[14] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 615–637, Mar. 2021.

[15] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-free non-orthogonal multiple access for IoT: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1805–1838, 3rd Quart., 2020.

[16] A. Subrahmannian and S. K. Behera, "Chipless RFID sensors for IoT-based healthcare applications: A review of state of the art," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–20, 2022.

[17] K.-H. Chang, "Bluetooth: A viable solution for IoT? [Industry perspectives]," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 6–7, Dec. 2014.

[18] S. G. Varghese et al., "Comparative study of ZigBee topologies for IoT-based lighting automation," *IET Wireless Sens. Syst.*, vol. 9, no. 4, pp. 201–207, Aug. 2019.

[19] S. R. Pokhrel and C. Williamson, "Modeling compound TCP over WiFi for IoT," *IEEE/ACM Trans. Netw.*, vol. 26, no. 2, pp. 864–878, Apr. 2018.

[20] Y. Chen, Y. A. Sambo, O. Onireti, and M. A. Imran, "A survey on LPWAN-5G integration: Main challenges and potential solutions," *IEEE Access*, vol. 10, pp. 32132–32149, 2022.

[21] U. Raza, P. Kulkarni, and M. Sooriyabandara, "Low power wide area networks: An overview," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 855–873, 2nd Quart., 2017.

[22] A. Lavric, A. I. Petrariu, and V. Popa, "Long range SigFox communication protocol scalability analysis under large-scale, high-density conditions," *IEEE Access*, vol. 7, pp. 35816–35825, 2019.

[23] J. P. S. Sundaram, W. Du, and Z. Zhao, "A survey on LoRa networking: Research problems, current solutions, and open issues," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 371–388, 1st Quart., 2020.

[24] S. Lippuner, B. Weber, M. Salomon, M. Korb, and Q. Huang, "EC-GSM-IoT network synchronization with support for large frequency offsets," in *Proc. WCNC*, Barcelona, Spain, Apr. 2018, pp. 1–6.

[25] R. Ratasuk, B. Vejlgaard, N. Mangalvedhe, and A. Ghosh, "NB-IoT system for M2M communication," in *Proc. WCNCW*, Apr. 2016, pp. 1–5.

[26] A. Adhikary, X. Lin, and Y.-P.-E. Wang, "Performance evaluation of NB-IoT coverage," in *Proc. VTC*, Montreal, QC, Canada, Sep. 2016, pp. 1–5.

[27] J. Peisa et al., "5G evolution: 3GPP releases 16 & 17 overview," *Ericsson Technol. Rev.*, vol. 2020, no. 2, pp. 2–13, Mar. 2020.

[28] R. Ratasuk, N. Mangalvedhe, A. Ghosh, and B. Vejlgaard, "Narrowband LTE-M system for M2M communication," in *Proc. VTC*, Vancouver, BC, Canada, Sep. 2014, pp. 1–5.

[29] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86–93, Jun. 2013.

[30] G. Ferré and E. P. Simon, "An introduction to Sigfox and LoRa PHY and MAC layers," working Paper, HAL Open Science, 2018.

[31] A. Hoglund et al., "3GPP release-16 preconfigured uplink resources for LTE-M and NB-IoT," *IEEE Commun. Stand. Mag.*, vol. 4, no. 2, pp. 50–56, Jun. 2020.

[32] S. Chen, A. Livingstone, H.-Q. Du, and L. Hanzo, "Adaptive minimum symbol error rate beamforming assisted detection for quadrature amplitude modulation," *IEEE Trans. Wireless Commun.*, vol. 7, no. 4, pp. 1140–1145, Apr. 2008.

[33] J. Zhang, S. Chen, X. Mu, and L. Hanzo, "Joint channel estimation and multi-user detection for SDMA/OFDM based on dual repeated weighted boosting search," *IEEE Trans. Veh. Technol.*, vol. 60, no. 7, pp. 3265–3275, Sep. 2011.

[34] J. Zhang, S. Chen, X. Mu, and L. Hanzo, "Evolutionary algorithm assisted joint channel estimation and turbo multi-user detection/decoding for OFDM/SDMA," *IEEE Trans. Veh. Technol.*, vol. 63, no. 3, pp. 1204–1222, Mar. 2014.

[35] S. Chen, S. X. Ng, E. F. Khalaf, A. Morfeq, and N. D. Alotaibi, "Multiuser detection for nonlinear MIMO uplink," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 207–219, Jan. 2020.

[36] W. Yao, S. Chen, S. Tan, and L. Hanzo, "Minimum bit error rate multiuser transmission designs using particle swarm optimisation," *IEEE Trans. Wireless Commun.*, vol. 8, no. 10, pp. 5012–5017, Oct. 2009.

[37] Y. Zhang, J. Gao, and Y. Liu, "MRT precoding in downlink multi-user MIMO systems," *EURASIP J. Wireless Commun. Netw.*, vol. 2016, pp. 1–7, Oct. 2016.

[38] H. Zhang, M. Ma, and Z. Shao, "Multi-user linear precoding for downlink generalized spatial modulation systems," *IEEE Commun. Lett.*, vol. 24, no. 1, pp. 212–216, Jan. 2020.

[39] S. Chen, S. X. Ng, E. Khalaf, A. Morfeq, and N. Alotaibi, "Particle swarm optimization assisted B-spline neural network based predistorter design to enable transmit precoding for nonlinear MIMO downlink," *Neurocomputing*, vol. 458, pp. 336–348, Oct. 2021.

[40] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura, "Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access," in *Proc. Int. Symp. ISPACS*, Naha, Japan, Nov. 2013, pp. 770–774.

[41] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for future radio access," in *Proc. VTC*, Dresden, Germany, Jun. 2013, pp. 1–5.

[42] K. Higuchi and Y. Kishiyama, "Non-orthogonal access with random beamforming and intra-beam SIC for cellular MIMO downlink," in *Proc. VTC*, Las Vegas, NV, USA, Sep. 2013, pp. 1–5.

[43] N. Nonaka, Y. Kishiyama, and K. Higuchi, "Non-orthogonal multiple access using intra-beam superposition coding and SIC in base station cooperative MIMO cellular downlink," in *Proc. VTC*, Vancouver, BC, Canada, Sep. 2014, pp. 1–5.

[44] R. Hoshyar, F. P. Wathan, and R. Tafazolli, "Novel low-density signature for synchronous CDMA systems over AWGN channel," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1616–1626, Apr. 2008.

[45] D. Guo and C.-C. Wang, "Multiuser detection of sparsely spread CDMA," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 3, pp. 421–431, Apr. 2008.

[46] J. V. D. Beek and B. M. Popovic, "Multiple access with low-density signatures," in *Proc. GLOBECOM*, Honolulu, HI, USA, Nov./Dec. 2009, pp. 1–6.

[47] R. Razavi, R. Hoshyar, M. A. Imran, and Y. Wang, "Information theoretic analysis of LDS scheme," *IEEE Commun. Lett.*, vol. 15, no. 8, pp. 798–800, Aug. 2011.

[48] L. Lu, Y. Chen, W. Guo, H. Yang, Y. Wu, and S. Xing, "Prototype for 5G new air interface technology SCMA and performance evaluation," *China Commun.*, vol. 12, no. 1, pp. 38–48, Dec. 2015.

[49] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang, and J. Xu, "Multi-user shared access for Internet of Things," in *Proc. VTC*, Nanjing, China, May. 2016, pp. 1–5.

[50] F. Wei, W. Chen, Y. Wu, J. Ma, and T. A. Tsiftsis, "Message-passing receiver design for joint channel estimation and data decoding in uplink grant-free SCMA systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 167–181, Jan. 2019.

[51] Z. Gao, L. Dai, S. Han, I. Chih-Lin, Z. Wang, and L. Hanzo, "Compressive sensing techniques for next-generation wireless communications," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 144–153, Jun. 2018.

[52] P. Vandenameele, L. Van Der Perre, and M. Engels, *Space Division Multiple Access For Wireless Local Area Networks*. Boston, MA, USA: Kluwer, 2001.

[53] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[54] H. Yin, D. Gesbert, M. Filippou, and Y. Liu, "A coordinated approach to channel estimation in large-scale multiple-antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 264–273, Feb. 2013.

[55] J. Zhang, B. Zhang, S. Chen, X. Mu, M. El-Hajjar, and L. Hanzo, "Pilot contamination elimination for large-scale multiple-antenna aided OFDM systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 759–772, Oct. 2014.

[56] P. Zhao, Z. Wang, C. Qian, L. Dai, and S. Chen, "Location-aware pilot assignment for massive MIMO systems in heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6815–6821, Aug. 2016.

[57] X. Guo, S. Chen, J. Zhang, X. Mu, and L. Hanzo, "Optimal pilot design for pilot contamination elimination/reduction in large-scale multiple-antenna aided OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7229–7243, Nov. 2016.

[58] A. Bayesteh, E. Yi, H. Nikopour, and H. Baligh, "Blind detection of SCMA for uplink grant-free multiple-access," in *Proc. ISWCS*, Barcelona, Spain, Aug. 2014, pp. 853–857.

[59] Z. Zhang, X. Wang, Y. Zhang, and Y. Chen, "Grant-free rateless multiple access: A novel massive access scheme for Internet of Things," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2019–2022, Oct. 2016.

[60] K. Senel and E. G. Larsson, "Grant-free massive MTC-enabled massive MIMO: A compressive sensing approach," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6164–6175, Dec. 2018.

[61] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[62] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4634–4643, Dec. 2006.

[63] M. E. Davies and Y. C. Eldar, "Rank awareness in joint sparse recovery," *IEEE Trans. Inf. Theory*, vol. 58. no. 2, pp. 1135–1146, Feb. 2012.

[64] S. F. Cotter, B. D. Rao, E. Kjersti, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, Jul. 2005.

[65] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[66] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Royal Stat. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[67] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 1993, pp. 40–44.

[68] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2230–2249, May 2009.

[69] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, May 2009.

[70] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 489–519, Feb. 2001.

[71] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *Proc. 7th Conf. Uncertainty Artif. Intell.*, San Francisco, CA, USA, Aug. 2001, pp. 362–369.

[72] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Sep. 2001.

[73] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. Motivation and construction," in *Proc. ITW*, Cairo, Egypt, Jan. 2010, pp. 1–5.

[74] B. Wang, L. Dai, Y. Yuan, and Z. Wang, "Compressive sensing based multi-user detection for uplink grant-free non-orthogonal multiple access," in *Proc. VTC*, Boston, MA, USA, Sep. 2015, pp. 1–5.

[75] Y. Mei et al., "Compressive sensing-based joint activity and data detection for grant-free massive IoT access," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1851–1869, Mar. 2022.

[76] A. T. Abebe and C. G. Kang, "Iterative order recursive least square estimation for exploiting frame-wise sparsity in compressive sensing-based MTC," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 1081–1021, May 2016.
[77] "NR; user equipment (UE) radio transmission and reception," 3GPP, Sophia Antipolis, France, document 3GPP, TS 38.101–1, V15.3.0, Sep. 2018.
[78] B. Wang, L. Dai, T. Mir, and Z. Wang, "Joint user activity and data detection based on structured compressive sensing for NOMA," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1473–1476, Jul. 2016.
[79] C. Wei, H. Liu, Z. Zhang, J. Dang, and L. Wu, "Approximate message passing-based joint user activity and data detection for NOMA," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 640–643, Mar. 2017.
[80] Y. Du et al., "Block-sparsity-based multiuser detection for uplink grant-free NOMA," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 7894–7909, Dec. 2018.
[81] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Trans. Signal Proccess.*, vol. 57, no. 8, pp. 3075–3085, Aug. 2009.
[82] J. Determe, J. Louveaux, L. Jacques, and F. Horlin, "On the noise robustness of simultaneous orthogonal matching pursuit," *IEEE Trans. Signal Process.*, vol. 65, no. 4, pp. 864–875, Feb. 2017.
[83] B. Wang, L. Dai, Y. Zhang, T. Mir, and J. Li, "Dynamic compressive sensing-based multi-user detection for uplink grant-free NOMA," *IEEE Commun. Lett.*, vol. 20, no. 11, pp. 2320–2323, Nov. 2016.
[84] Y. Du et al., "Efficient multi-user detection for uplink grant-free NOMA: Prior-information aided adaptive compressive sensing perspective," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2812–2828, Dec. 2017.
[85] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.
[86] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3292–3304, Sep. 2012.
[87] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.
[88] A.-S. Bana et al., "Massive MIMO for Internet-of-Things (IoT) connectivity," *Phys. Commun.*, vol. 37, pp. 1–17, Dec. 2019.
[89] Z. Gao, L. Dai, Z. Wang, S. Chen, and L. Hanzo, "Compressive-sensing-based multiuser detector for the large-scale SM-MIMO uplink," *IEEE Trans. Veh. Technol.*, vol. 65, no. 10, pp. 8725–8730, Oct. 2016.
[90] X. Meng, S. Wu, L. Kuang, D. Huang, and J. Lu, "Multi-user detection for spatial modulation via structured approximate message passing," *IEEE Commun. Lett.*, vol. 20, no. 8, pp. 1527–1530, Aug. 2016.
[91] X. Ma, J. Kim, D. Yuan, and H. Liu, "Two-level sparse structure-based compressive sensing detector for uplink spatial modulation with massive connectivity," *IEEE Commun. Lett.*, vol. 23, no. 9, pp. 1594–1597, Sep. 2019.
[92] T. Mao, Q. Wang, Z. Wang, and S. Chen, "Novel index modulation techniques: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 315–348, 1st Quart., 2019.
[93] A. K. Khandani, "Media-based modulation: A new approach to wireless transmission," in *Proc. ISIT*, Istanbul, Turkey, Jul. 2013, pp. 3050–3054.
[94] B. Shamasundar, S. Jacob, L. N. Theagarajan, and A. Chockalingam, "Media-based modulation for the uplink in massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8169–8183, Sep. 2018.
[95] L. Zhang, M. Zhao, and L. Li, "Low-complexity multi-user detection for MBM in uplink large-scale MIMO systems," *IEEE Commun. Lett.*, vol. 22, no. 8, pp. 1568–1571, Aug. 2018.
[96] L. Qiao, J. Zhang, Z. Gao, D. W. K. Ng, M. D. Renzo, and M.-S. Alouini, "Massive access in media modulation based massive machine-type communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 339–356, Jan. 2022.
[97] L. Qiao, J. Zhang, Z. Gao, S. Chen, and L. Hanzo, "Compressive sensing based massive access for IoT relying on media modulation aided machine type communications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 10391–10396, Sep. 2020.
[98] X. Ma, S. Guo, and D. Yuan, "Improved compressed sensing-based joint user and symbol detection for media-based modulation-enabled massive machine-type communications," *IEEE Access*, vol. 8, pp. 70058–70070, 2020.
[99] H. F. Schepker, C. Bockelmann, and A. Dekorsy, "Exploiting sparsity in channel and data estimation for sporadic multi-user communication," in *Proc. ISWCS*, Ilmenau, Germany, Aug. 2013, pp. 1–5.
[100] J. Ahn, B. Shim, and K. B. Lee, "EP-based joint active user detection and channel estimation for massive machine-type communications," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5178–5189, Jul. 2019.
[101] L. Liu and W. Yu, "Massive connectivity with massive MIMO—Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, Jun. 2018.
[102] B. Knoop, S. Schmale, D. Peters-Drolshagen, and S. Paul, "Activity and channel estimation in multi-user wireless sensor networks," in *Proc. WSA*, Munich, Germany, Mar. 2016, pp. 1–5.
[103] S. Park, H. Seo, H. Ji, and B. Shim, "Joint active user detection and channel estimation for massive machine-type communications," in *Proc. SPAWC*, Sapporo, Japan, Jul. 2017, pp. 1–5.
[104] L. Liu and W. Yu, "Massive connectivity with massive MIMO—Part II: Achievable rate characterization," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2947–2959, Jun. 2018.
[105] X. Shao, X. Chen, C. Zhong, J. Zhao, and Z. Zhang, "A unified design of massive access for cellular Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 3934–3947, Apr. 2019.
[106] Z. Chen, F. Sohrabi, and W. Yu, "Sparse activity detection for massive connectivity," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1890–1904, Apr. 2018.
[107] X. Shao, X. Chen, and R. Jia, "A dimension reduction-based joint activity detection and channel estimation algorithm for massive access," *IEEE Trans. Signal Process.*, vol. 68, pp. 420–435, Jan. 2020.
[108] J. Jiang and H. Wang, "Massive random access with sporadic short packets: Joint active user detection and channel estimation via sequential message passing," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4541–4555, Jul. 2021.
[109] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Trans. Signal Process.*, vol. 68, pp. 764–779, 2020.
[110] J.-C. Jiang and H.-M. Wang, "Grouping-based joint active user detection and channel estimation with massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2305–2319, Apr. 2022.
[111] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.
[112] X. Xu, X. Rao, and V. K. N. Lau, "Active user detection and channel estimation in uplink C-RAN systems," in *Proc. ICC*, London, U.K., Jun. 2015, pp. 2727–2732.
[113] Z. Utkovski, O. Simeone, T. Dimitrova, and P. Popovski, "Random access in C-RAN for user activity detection with limited-capacity fronthaul," *IEEE Signal Process. Lett.*, vol. 24, no. 1, pp. 17–21, Jan. 2017.
[114] Z. Chen, F. Sohrabi, and W. Yu, "Multi-cell sparse activity detection for massive random access: Massive MIMO versus cooperative MIMO," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4060–4074, Aug. 2019.
[115] M. Ke, Z. Gao, Y. Wu, X. Gao, and K.-K. Wong,, "Massive access in cell-free massive MIMO-based Internet of Things: Cloud computing and edge computing paradigms," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 756–772, Mar. 2021.
[116] X. Shao, X. Chen, D. W. K. Ng, C. Zhong, and Z. Zhang, "Cooperative activity detection: Sourced and unsourced massive random access paradigms," *IEEE Trans. Signal Process.*, vol. 68, pp. 6578–6593, Dec. 2020.
[117] Y. Polyanskiy, "A perspective on massive random-access," in *Proc. ISIT*, Aachen, Germany, Jun. 2017, pp. 2523–2527.
[118] O. Ordentlich and Y. Polyanskiy, "Low complexity schemes for the random access Gaussian channel," in *Proc. ISIT*, Aachen, Germany, Jun. 2017,pp. 2528–2532.
[119] A. Vem, K. R. Narayanan, J. Cheng, and J.-F. Chamberland, "A user-independent serial interference cancellation based coding scheme for the unsourced random access Gaussian channel," in *Proc. ITW*, Feb. 2017, pp. 121–125.
[120] E. Marshakov, G. Balitskiy, K. Andreev, and A. Frolov, "A polar code based unsourced random access for the Gaussian MAC," in *Proc. VTC*, Honolulu, HI, USA, Sep. 2019, pp. 1–5.
[121] A. K. Tanc and T. M. Duman, "Massive random access with trellis-based codes and random signatures," *IEEE Commun. Lett.*, vol. 25, no. 5, pp. 1496–1499, May 2021.

[122] A. K. Pradhan, V. K. Amalladinne, K. R. Narayanan, and J. Chamberland, "Polar coding and random spreading for unsourced multiple access," in *Proc. ICC*, Dublin, Ireland, Jun. 2020, pp. 1–6.
[123] M. J. Ahmadi and T. M. Duman, "Random spreading for unsourced MAC with power diversity," *IEEE Commun. Lett.*, vol. 25, no. 12, pp. 3995–3999, Dec. 2021.
[124] D. Truhachev, M. Bashir, A. Karami, and E. Nassaji, "Low-complexity coding and spreading for unsourced random access," *IEEE Commun. Lett.*, vol. 25, no. 3, pp. 774–778, Mar. 2021.
[125] A. K. Pradhan, V. K. Amalladinne, A. Vem, K. R. Narayanan, and J.-F. Chamberland, "Sparse IDMA: A joint graph-based coding scheme for unsourced random access," *IEEE Trans. Commun.*, vol. 70, no. 11, pp. 7124–7133, Nov. 2022.
[126] V. K. Amalladinne, A. Vem, D. K. Soma, K. R. Narayanan, and J.-F. Chamberland, "A coupled compressive sensing scheme for unsourced multiple access," in *Proc. ICASSP*, Calgary, AB, Canada, Apr. 2018, pp. 6628–6632.
[127] V. K. Amalladinne, A. K. Pradhan, C. Rush, J.-F. Chamberland, and K. R. Narayanan, "On approximate message passing for unsourced access with coded compressed sensing," in *Proc. ISIT*, Los Angeles, CA, USA, Jun. 2020, pp. 2995–3000.
[128] A. Fengler, P. Jung, and G. Caire, "SPARCs for unsourced random access," *IEEE Trans. Inf. Theory*, vol. 67, no. 10, pp. 6894–6915, Oct. 2021.
[129] A. Fengler, G. Caire, P. Jung, and S. Haghighatshoar, "Massive MIMO unsourced random access," Jan. 2019, *arXiv:1901.00828*.
[130] S. Haghighatshoar, P. Jung, and G. Caire, "A new scaling law for activity detection in massive MIMO systems," Jun. 2018, *arXiv:1803.02288*.
[131] A. Decurninge, I. Land, and M. Guillaud, "Tensor-based modulation for unsourced massive random access," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 552–556, Mar. 2021.
[132] V. Shyianov, F. Bellili, A. Mezghani, and E. Hossain, "Massive unsourced random access based on uncoupled compressive sensing: Another blessing of massive MIMO," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 820–834, Mar. 2021.
[133] X. Xie et al., "Massive unsourced random access: Exploiting angular domain sparsity," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2480–2498, Apr. 2022.
[134] A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire, "Non-Bayesian activity detection, large-scale fading coefficient estimation, and unsourced random access with a massive MIMO receiver," *IEEE Trans. Inf. Theory*, vol. 67, no. 5, pp. 2925–2951, May 2021.
[135] L. Qiao et al., "Joint activity and blind information detection for UAV-assisted massive IoT access," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1489–1508, May 2022.
[136] R. Mndez-Rial, C. Rusu, N. González-Prelcic, A. Alkhateeb, and R. W. Heath, "Hybrid MIMO architectures for millimeter-wave communications: Phase shifters or switches?" *IEEE Access*, vol. 4, pp. 247–267, 2016.
[137] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.
[138] L. Fan, S. Jin, C.-K. Wen, and H. Zhang, "Uplink achievable rate for massive MIMO systems with low-resolution ADC," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2186–2189, Dec. 2015.
[139] G. Sun et al., "Massive grant-free OFDMA with timing and frequency offsets," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3365–3380, May 2022.
[140] Y. Li, M. Xia, and Y.-C. Wu, "Activity detection for massive connectivity under frequency offsets via first-order algorithms," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1988–2002, Mar. 2019.
[141] M. Valkama, M. Renfors, and V. Koivunen, "Advanced methods for I/Q imbalance compensation in communication receivers," *IEEE Trans. Signal Process.*, vol. 49, no. 10, pp. 2335–2344, Oct. 2001.
[142] J. Qiu, D. Grace, G. Ding, M. D. Zakaria, and Q. Wu, "Air-ground heterogeneous networks for 5G and beyond via integrating high and low altitude platforms," *IEEE Wireless Commun.*, vol. 26, no. 6, pp. 140–148, Dec. 2019.
[143] N. Cheng et al., "Space/aerial-assisted computing offloading for IoT applications: A learning-based approach," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 5, pp. 1117–1129, May 2019.
[144] W. Khawaja et al., "A survey of air-to-ground propagation channel modeling for unmanned aerial vehicles," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2361–2391, 3rd Quart. 2019.
[145] A. Al-Hourani, S. Chandrasekharan, G. Kaandorp, W. Glenn, A. Jamalipour, and S. Kandeepan, "Coverage and rate analysis of aerial base stations," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 6, pp. 3077–3081, Dec. 2016.
[146] Y. Qing, X. Shao, and X. Chen, "A model-driven deep learning algorithm for joint activity detection and channel estimation," *IEEE Commun. Lett.*, vol. 24, no. 11, pp. 2508–2512, Nov. 2020.
[147] X. Shao, X. Chen, Y. Qiang, C. Zhong, and Z. Zhang, "Feature-aided adaptive-tuning deep learning for massive device detection," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 1899–1914, Jul. 2021.
[148] K. Ma, Z. Wang, W. Tian, S. Chen, and L. Hanzo, "Deep learning for mmWave beam-management: State-of-the-art, opportunities and challenges," *IEEE Wireless Commun.*, vol. 30, no. 4, pp. 108–114, Aug. 2023.
[149] Z. Zhang et al., "DNN-aided block sparse Bayesian learning for user activity detection and channel estimation in grant-free non-orthogonal random access," *IEEE Trans. Veh. Technol.*, vol. 68, no. 12, pp. 12000–12012, Dec. 2019.
[150] W. Kim, Y. Ahn, and B. Shim, "Deep neural network-based active user detection for grant-free NOMA systems," *IEEE Trans. Commun.*,vol. 68, no. 4, pp. 2143–2155, Apr. 2020.
[151] T. Zhao, F. Li, and P. Tian, "A deep-learning method for device activity detection in mMTC under imperfect CSI based on variational autoencoder," *IEEE Trans. Veh. Technol.*, vol. 69, no. 7, pp. 7981–7986, Jul. 2020.
[152] H. Yu et al.,"Deep learning-based user activity detection and channel estimation in grant-free NOMA," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2202–2214, Apr. 2023.
[153] Y. Bai, W. Chen, B. Ai, Z. Zhong, and I. J. Wassell, "Prior information aided deep learning method for grant-free NOMA in mMTC," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 112–126, Jan. 2022.
[154] W. Zhu, M. Tao, X. Yuan, and Y. Guan, "Deep-learned approximate message passing for asynchronous massive connectivity," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5434–5448, Aug. 2021.
[155] X. Lin, L. Kuang, Z. Ni, C. Jiang, and S. Wu, "Approximate message passing-based detection for asynchronous NOMA," *IEEE Commun. Lett.*, vol. 24, no. 3, pp. 534–538, Mar. 2020.
[156] Y. Cheng, L. Liu, and P. Li, "Orthogonal AMP for massive access in channels with spatial and temporal correlations," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 726–740, Mar. 2021.
[157] T. Li et al., "Joint device detection, channel estimation, and data decoding with collision resolution for MIMO massive unsourced random access," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1535–1555, May. 2022.
[158] X. Bian, Y. Mao, and J. Zhang, "Joint activity detection, channel estimation, and data decoding for grant-free massive random access," *IEEE Internet Things J.*, vol. 10, no. 16, pp. 14042–14057, Aug. 2023.
[159] R. B. Di Renna and R. C. de Lamare, "Joint channel estimation, activity detection and data decoding based on dynamic message-scheduling strategies for mMTC," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2464–2479, Apr. 2022.
[160] X. Zhou et al., "Active terminal identification, channel estimation,and signal detection for grant-free NOMA-OTFS in LEO satellite Internet-of-Things," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2847–2866, Apr. 2023.

Zhen Gao received the B.S. degree in information engineering from Beijing Institute of Technology, Beijing, China, in 2011, and the Ph.D. degree in communication and signal processing from the Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, in 2016.

He is currently an Assistant Professor with Beijing Institute of Technology. His research interests are in wireless communications, with a focus on multi-carrier modulations, multiple antenna systems, and sparse signal processing.

Dr. Gao was the recipient of the IEEE Broadcast Technology Society 2016 Scott Helt Memorial Award (Best Paper), the Exemplary Reviewer of the IEEE COMMUNICATION LETTERS in 2016, IET *Electronics Letters* Premium Award (Best Paper) 2016, and the Young Elite Scientists Sponsorship Program from the China Association for Science and Technology from 2018 to 2021.

Malong Ke (Member, IEEE) received the B.S. degree in communication engineering from Shandong University, Jinan, China, in 2017, and the Ph.D. degree in information and communication engineering from Beijing Institute of Technology, Beijing, China, in 2023.

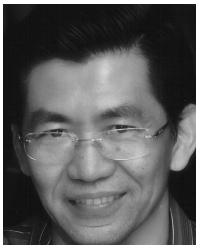
He is currently an Engineer with the Wireless Product Division, Ruijie Network Company Ltd., Fuzhou, China. His research interests include massive access for mMTC, massive MIMO systems, sparse signal processing, integrated sensing and communication, and nonterrestrial network.

Yikun Mei received the B.S. degree from Beijing Institute of Technology, Beijing, China, in 2019, where he is currently pursuing the Ph.D. degree with the School of Information and Electronics.

His research interests include massive access for mMTC and sparse signal processing.

Li Qiao (Graduate Student Member, IEEE) received the B.Sc. degree from Beijing Institute of Technology, Beijing, China, in 2019, where he is currently pursuing the Ph.D. degree with the School of Information an Electronics.

He is a collaborative Ph.D. student with the 5GIC/6GIC, University of Surrey, Guildford, U.K., and was a visiting student with IPC Laboratory, Imperial College London, London, U.K. His current research interests include massive connectivity, federated edge intelligence, and massive MIMO systems.



Sheng Chen (Life Fellow, IEEE) received the B.Eng. degree in control engineering from the East China Petroleum Institute, Dongying, China, in 1982, the Ph.D. degree in control engineering from the City University, London, U.K., in 1986, and the D.Sc. degree from the University of Southampton, Southampton, U.K., in 2005.

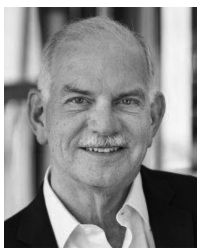
From 1986 to 1999, he held research and academic appointments at The University of Sheffield, Sheffield, U.K.; The University of Edinburgh, Edinburgh, U.K.; and University of Portsmouth, Portsmouth, U.K. Since 1999, he has been with the School of Electronics and Computer Science, University of Southampton, where he holds the post of a Professor of Intelligent Systems and Signal Processing. He has published over 700 research papers. He has more than 19 400 Web of Science citations with h-index 61 and more than 37 900 Google Scholar citations with h-index 82. His research interests include adaptive signal processing, wireless communications, modeling and identification of nonlinear systems, neural network and machine learning, and evolutionary computation methods and optimization.

Prof. Chen is one of the original ISI Highly Cited Researcher in engineering in March 2004. He is a Fellow of the United Kingdom Royal Academy of Engineering, Asia–Pacific Artificial Intelligence Association, and IET.

Derrick Wing Kwan Ng (Fellow, IEEE) received the bachelor's degree (First-Class Hons.) and the M.Phil. degree in electronic engineering from Hong Kong University of Science and Technology, Hong Kong, in 2006 and 2008, respectively, and the Ph.D. degree from The University of British Columbia, Vancouver, BC, Canada, in November 2012.

He was a Senior Postdoctoral Fellow with the Institute for Digital Communications, Friedrich-Alexander-University Erlangen-Nürnberg, Erlangen, Germany. He is currently working as a Scientia Associate Professor with the University of New South Wales, Sydney, NSW, Australia. His research interests include global optimization, physical-layer security, IRS-assisted communication, UAV-assisted communication, wireless information and power transfer, and green (energy-efficient) wireless communications.

Dr. Ng received the Australian Research Council Discovery Early Career Researcher Award 2017, the IEEE Communications Society Leonard G. Abraham Prize 2023, the IEEE Communications Society Stephen O. Rice Prize 2022, the Best Paper Awards at the WCSP 2020, 2021, the IEEE TCGCC Best Journal Paper Award 2018, INISCOM 2018, IEEE International Conference on Communications 2018, 2021, and 2023, IEEE International Conference on Computing, Networking and Communications (ICNC) 2016, IEEE Wireless Communications and Networking Conference 2012, the IEEE Global Telecommunication Conference 2011, 2021, and the IEEE Third International Conference on Communications and Networking in China 2008. He has been listed as a Highly Cited Researcher by Clarivate Analytics (Web of Science) since 2018. He served as an Editorial Assistant for the Editor-in-Chief of the IEEE TRANSACTIONS ON COMMUNICATIONS from January 2012 to December 2019. He is currently serving as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS and an Associate Editor-in-Chief for the IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY.



H. Vincent Poor (Life Fellow, IEEE) received the Ph.D. degree in electrical engineering and computer science from Princeton University, Princeton, NJ, USA, in 1977.

From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana–Champaign, Urbana, IL, USA. Since 1990, he has been on the faculty at Princeton, where he is currently the Michael Henry Strater University Professor. From 2006 to 2016, he served as the Dean of Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests are in the areas of information theory, machine learning and network science, and their applications in wireless networks, energy systems, and related fields. Among his publications in these areas is the recent book *Machine Learning and Wireless Communications* (Cambridge University Press, 2022).

Dr. Poor received the IEEE Alexander Graham Bell Medal in 2017. He is a member of the National Academy of Engineering and the National Academy of Sciences and is a foreign member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies.