Massive Access in Space-based Internet of Things: Challenges, Opportunities, and Future Directions

Jian Jiao, Shaohua Wu, Rongxing Lu, and Qinyu Zhang

Abstract

The integration of fifth generation (5G) mobile network and satellite communication promises a promotion on the interconnection of everything, and has become one of the key development directions of beyond 5G and even 6G. As an inevitable development trend of mobile communication, the space-based Internet of Things (S-IoT) is expected to be commercially available in building an integrated information infrastructure network of the space, air, ground and oceans, and so on. However, the tremendous growth in the number of connected user equipments poses a fundamental rethinking of conventional multiple access technologies, which is considered one of the open challenges in future S-IoT networks. In this article, we first review the main challenges of massive access in S-IoT. Then, we present and discuss how to develop a scalable massive access S-IoT network by leveraging introduced technologies, including intelligent hybrid non-orthogonal multiple access (NOMA) transmission, grant-free superimposed pilot NOMA random access and multi-slot pilot allocation random access protocol. Finally, some exciting future directions of massive access in S-IoT are revealed.

Introduction

Fifth generation (5G) mobile communication can be regarded as the beginning from *Internet of Human* to *Internet of Everything*. 5G not only upgrades the existing 4G network by enhancing mobile broadband (eMBB), but also draws forth two typical Internet of Things (IoT) scenarios: ultra-reliable low-latency communications (uRLLC) and massive machine type communications (mMTC), which are devoted to promoting the ability for broadband multiple access at *anywhere and anytime* [1]. However, the terrestrial network only covers about 20 percent of the surface of the Earth, yet resulting in the difficulty to provide ubiquitous connectivity and on-demand multiple access services, such as smart agriculture, natural disaster prevention, climate monitoring and intelligent transportation system (ITS), and so on.

With the rapid development of Ka/Q/V millimeter-wave (mmWave) band high throughput satellites (HTS), the space-based IoT (S-IoT) has been recognized as one of the technology driven paradigm shifts and the continuous evolution of wireless networks for beyond 5G and even 6G. The envisioned S-IoT can provide global coverage and time-critical broadband access in a cost-efficient manner, including intelligent traffic, smart grid, telemedicine and industrial automation, and so on, especially in the extreme topographies with limited terrestrial Internet infrastructures, such as environmental monitoring, and disaster relief due to the agile deployment capability [2]. Moreover, the multibeam low earth orbit (LEO) HTS can provide much lower latency and propagation loss in comparison with the geostationary earth orbit (GEO) satellite, which is regarded as a key component in the forthcoming S-IoT and has aroused universal attention in both industry and academia.

Further, the aforementioned emergent IoT applications are usually equipped with a large-scale of ground user equipments (UEs), which continuously monitor, interpret and transmit the status information to the access HTS. It is worth noting that the practical implementation of massive access wireless communication is one of the open challenges in the upcoming S-IoT network [3]. Considering that the non-trivial propagation delay due to the huge communication distances in S-IoT and the limited resources and computing capacity of the HTS [4], it is a huge challenge for S-IoT both enabling the massive access of large-scale UEs efficiently and ensuring the timeliness of updated information. To address these issues, this article will elaborate the dynamic topology with intelligent adaptation transmission, hybrid pilot design and on-demand multi-slot random access, which will potentially support massive access and provide insights into the practical implementation of future S-IoT network.

The remainder of this article is organized as follows. The challenges of massive access in S-IoT are given in the following section. We then elaborate the promising technologies for massive access in S-IoT. After that, future research directions and conclusion are presented.

The Demand and Challenges of Massive Access in S-IoT

The latency and reliability requirements of S-IoT services are drawn in Fig. 1, where the time-critical and large-scale access services that can be supported by S-IoT are shown in the right part [1,

Digital Object Identifier:
10.1109/MWC.001.2000456

Jian Jiao, Shaohua Wu (corresponding author), and Qinyu Zhang (corresponding author) are with Harbin Institute of Technology (Shenzhen) and also with Peng Cheng Laboratory; Rongxing Lu is with the Faculty of Computer Science, University of New Brunswick.

1536-1284/21/$25.00 © 2021 IEEE

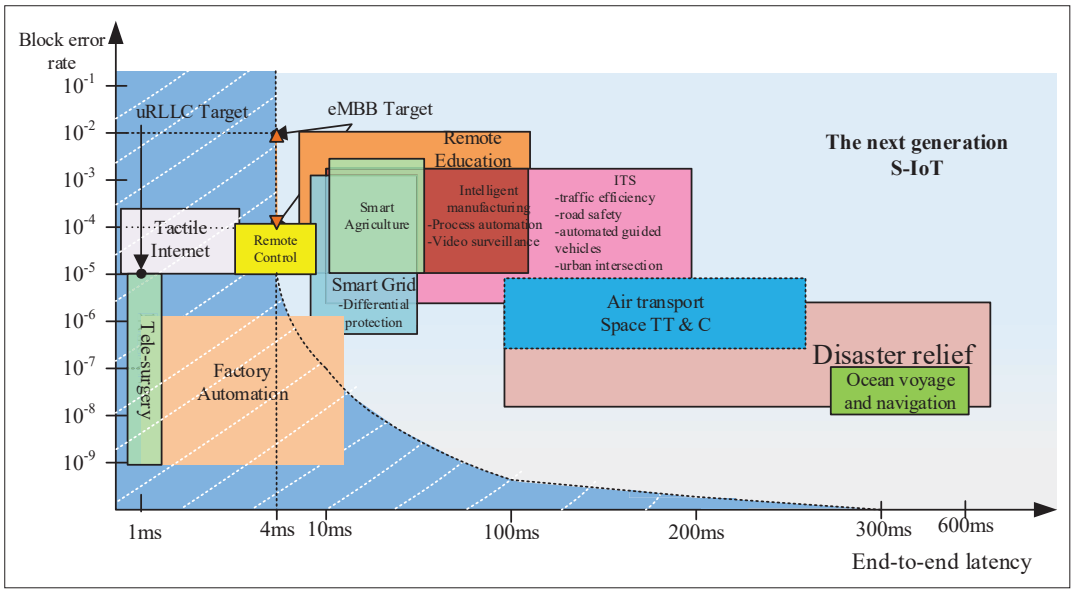

FIGURE 1. The end-to-end latency and reliability requirements for future S-IoT network.

2], and the block error rate (BLER) of uRLLC envisioned by 5G is less than 10^{-5} and the end-to-end latency is less than 1 ms, which is shown in the left part of Fig. 1 [5].

Dynamic Topology of S-IoT Networks

In recent years, the upcoming giant LEO constellation projects such as Starlink, OneWeb and Telesat, and so on, have launched thousands of LEO mmWave HTS, which can cooperate with terrestrial infrastructures to realize S-IoT networks, and provide global geographic coverage and time-critical broadband access in a cost-efficient manner. To fully exploit the ubiquitous broadband access capability of HTS to simultaneously support multiple UEs, non-orthogonal multiple access (NOMA) can be utilized in S-IoT networks [4, 6]. In fact, although the UEs within a beam coverage have similar distances to the HTS, it still can observe significant differences in their channel gains due to the various path losses, which makes it feasible to apply NOMA in S-IoT networks. Moreover, compared to orthogonal multiple access (OMA) schemes, the S-IoT with NOMA can reduce at least half of the transmission phases at the cost of a little more power consumption [6], which can significantly reduce the end-to-end transmission delay in comparison with OMA due to the non-trivial propagation delay in S-IoT. However, the performance of S-IoT with NOMA is constrained due to the mobility and the number of grouped UEs, which should be carefully optimized in the dynamic topology S-IoT network.

Short Packet Communications in MMTC

mMTC is dominated by uplink-oriented short packet transmissions, where a HTS should provide uplink mMTC for massive UEs simultaneously. Thus, the S-IoT could perform random access to reduce the control overhead. The most widely used random access protocols in satellite communications are the Aloha protocols, which can partially alleviate the conflict of massive access [7]. A coded slotted Aloha scheme for non-collaborative random access is proposed in [8], where the bipartite graph with erasure coding is utilized for asymptotical analysis. Further, the frameless Aloha protocol for time-varying channels is introduced in [9] by leveraging rateless coding, where the theoretical performance analysis and sub-optimal solution between the access failure probability (AFP) and activate probability of UEs are presented under limited access time slots. However, the Aloha protocols essentially utilize channel resources in an orthogonal manner, limiting the number of UEs that can simultaneously be accessed in the system [3]. Thus, the HTS can simultaneously serve multiple UEs at the same frequency via NOMA, which improves the performance of massive access in S-IoT networks. Furthermore, the pilot sequence and data payload follow successive transmission in a conventional regular orthogonal pilot (RP) scheme, which is sub-optimal in short packet communications because the length of the pilot is comparable to the data payload.

Massive Access in a Crowded Scenario

In addition, various applications for the future S-IoT are foreseen to increase exponentially UEs during the years ahead, which impose critical challenges for the S-IoT network to provide massive connectivity with short packets. Since the number of IoT UEs is growing at an impressive rate and much larger than that of the available pilot sequences, pilot collision is unavoidable. To alleviate this issue, some relevant works have increased the access time slots to provide secondary access opportunity for collision UEs. A strongest-user collision resolution (SUCR) random access protocol is proposed in [10], where the access point allows the collision UEs with the strongest channel gain to perform the second access. Moreover, the extension versions named SURC combined idle pilots access (SUCR-IPA) and SURC combined graph-based pilots access (SUCR-GBPA) are proposed in [11]. However, these protocols still could not completely solve the congestion problem from a massive number of bursty UEs transmission.

The HTS can simultaneously serve multiple UEs at the same frequency via NOMA, which improves the performance of massive access in S-IoT networks. Furthermore, the pilot sequence and data payload follow successive transmission in a conventional regular orthogonal pilot (RP) scheme, which is sub-optimal in short packet communications because the length of the pilot is comparable to the data payload.

Two algorithms are proposed to minimize the OP and maximize the system throughput, named the iteration power allocation (IPA) scheme for the quasi-static scenario, and power reallocation method based on the expectation maximization (PREM) scheme for intra-cell and inter-cell scenarios. The proposed intelligent hybrid NOMA transmission schemes are optimized due to the constrains of the complexity and OP requirement for UEs.

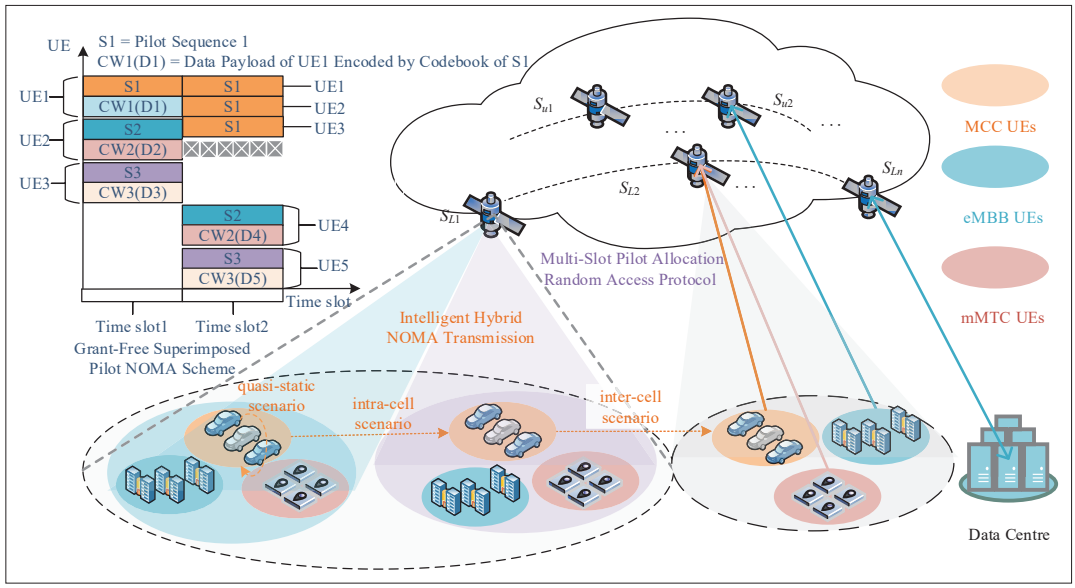

FIGURE 2. The key technologies for the massive access in S-IoT network.

The Key Technologies for Massive Access in S-IoT

In this section, we introduce three NOMA-related technologies to address the aforementioned challenges for massive access in S-IoT, including intelligent hybrid NOMA transmission, superimposed pilot NOMA (SP-NOMA), and multi-slot pilot allocation (MSPA) random access protocol, which have the potential to break through the bottleneck of massive access in S-IoT, as shown in Fig. 2.

Intelligent Hybrid NOMA Transmission for Dynamic Topology

Note that the end-to-end delay from the LEO HTS to the end UEs can be limited in 10 ms, the S-IoT network has the potential to support the mission critical communications (MCC) for the Internet of Vehicles, and also provide the ubiquitous connectivity inherited from the HTS. The performance of dynamic topology S-IoT with NOMA is affected to the mobility and number of grouped UEs. Moreover, the onboard power and storage resources of HTS are limited, which raises several challenging issues. Therefore, the first key technology introduced in this article is to design an intelligent hybrid NOMA transmission scheme for the dynamic topology S-IoT network to minimize the outage probability (OP) and maximize the system throughput.

Downlink Transmission Scheme: Moreover, how to group the UEs and allocate the downlink resources is crucial due to the mobility of UEs, which should be carefully optimized in the proposed intelligent hybrid NOMA transmission. Note that the high directivity and sparsity inherent in the mmWave band channel, we establish a sparse geometric-based channel model for the satellite-terrestrial channel in S-IoT networks. Then, we study three typical scenarios according to the mobility of UEs during the transmission, and formulate optimization problems for the intelligent hybrid NOMA transmission, including the quasi-static, intra-cell and inter-cell scenarios, where the UEs remain in the same beam, move into a different beam of the same HTS, and move into a different HTS, respectively. Two algorithms are proposed to minimize the OP and maximize the system throughput, named the iteration power allocation (IPA) scheme for the quasi-static scenario, and power reallocation method based on the expectation maximization (PREM) scheme for intra-cell and inter-cell scenarios. The proposed intelligent hybrid NOMA transmission schemes are optimized due to the constrains of the complexity and OP requirement for UEs.

Simulation results validate that our proposed intelligent hybrid NOMA transmission schemes outperform the benchmark schemes [12] as shown in Fig. 3, where the NOMA power allocation coefficient is 0.75, the number of groups in each beam is set as 2, and the number of UEs in each group is initialized as 5, and the UEs are randomly distributed in the beam following the Poisson distribution. The number of antennas in HTS is 32, and the mmWave band channel is assumed at 30 GHz with 2 GHz bandwidth. It can be observed that the proposed IPA and PREM schemes can achieve almost the same throughput as the global exhausted search schemes (complete power allocation (CPA) and complete power reallocation (CPR) schemes) at a much lower calculation cost, and the throughput is significantly higher than the original power allocation (OPA) scheme and the k-means based power reallocation (KPR) scheme in the three scenarios, especially in the high signal-to-noise ratio (SNR) region. Moreover, the throughput of the quasi-static scenario is lower than that of intra-cell and inter-cell scenarios, when the SNR is less than 20 dB. It is worth noting that the number of grouped UEs in intra-cell and inter-cell scenarios has significant impact on system performance, which should be further studied to provide practical guidelines to design an intelligent hybrid NOMA transmission for the dynamic topology S-IoT network.

Uplink Transmission Scheme: In addition, for the uplink transmission in S-IoT networks, we can classify the UEs according to their quality of service (QoS) requirements, that is, the eMBB UEs, mMTC

UEs and MCC UEs. Consider the ongoing giant multi-layer LEO mmWave HTS constellation such as Starlink, where the high mobility of LEO HTS would lead to only several minutes line of sight (LoS) link duration to access each LEO HTS. Therefore, we assume that the MCC UEs and mMTC UEs only access to the lower level HTS SL due to the better channel conditions and the lower propagation delay, and let the eMBB UEs connect to higher layer HTS Su with longer LoS link duration and also can connect to the lower layer HTS if necessary. Then, by taking account of the satellite handoff cost and the LoS duration of each layer HTS, we can formulate three optimization problems for the intelligent hybrid NOMA uplink transmission to guarantee the QoS requirements of the above mentioned three types of UEs. Preliminary studies indicate that different types of UEs can be grouped in the intelligent hybrid NOMA transmission under a well-designed combination scheme, and different QoS requirements can be simultaneously satisfied. Thus, the intelligent hybrid NOMA uplink transmission for the dynamic topology S-IoT should be further studied.

Grant-Free SP-NOMA Random Access for Short Packet Communications

Consider the tremendous growth of UEs in future mMTC applications with S-IoT, the dedicated pilot allocation in the conventional grant-based scheme is infeasible. The decentralized assignment of pilot sequences and grant-free random access have become a natural choice for the S-IoT. Moreover, the LEO HTS is still hundreds or thousands of miles away from UEs, the grant-free random access can reduce the propagation delay and signaling overhead caused by conventional grant-based schemes. Thus, in grant-free random access, each activated UE selects a pilot sequence at random from a predefined pilot sequence set, and then sends it to the HTS followed by data payload, which could alleviate the overload and pilot collision.

SP-NOMA Random Access Scheme: Furthermore, the most common traffic in massive access is uplink-oriented mMTC, which is dominated by short packet communications. Note that the pilot sequence and data payload are successively transmitting in conventional RP scheme. However, the length of the pilot sequence is non-negligible compared to the data payload in short packet communications. Therefore, inspired by the widely used spread spectrum technology in satellite communications, we propose a SP-NOMA random access scheme for massive access in S-IoT networks, in which the pilot and data sequences are transmitted simultaneously in a superimposing manner, and do not require additional time-frequency resources for pilots. In grant-free SP-NOMA random access, each activated UE randomly selects a pilot and superimposes on its data payload, then performs access to HTS. Assume that if multiple activated UEs select the same pilot, the HTS can identify the collision by utilizing the statistical channel state information (CSI) of received signals and a real domain codebook of pilots [3]. Note that the collided signals will obstruct the recovery of single UEs, it is essential to estimate the total interference power of collided signals. Furthermore, the SP-NOMA scheme would lead to the ill-posed problem with the conventional least square channel estimation (CE) algorithm, because the matrix of pilot sequences is irreversible. Thus, we propose an iteration CE based on ridge regression (ICER) algorithm to calculate the optimal power allocation coefficients between pilot and data payload.

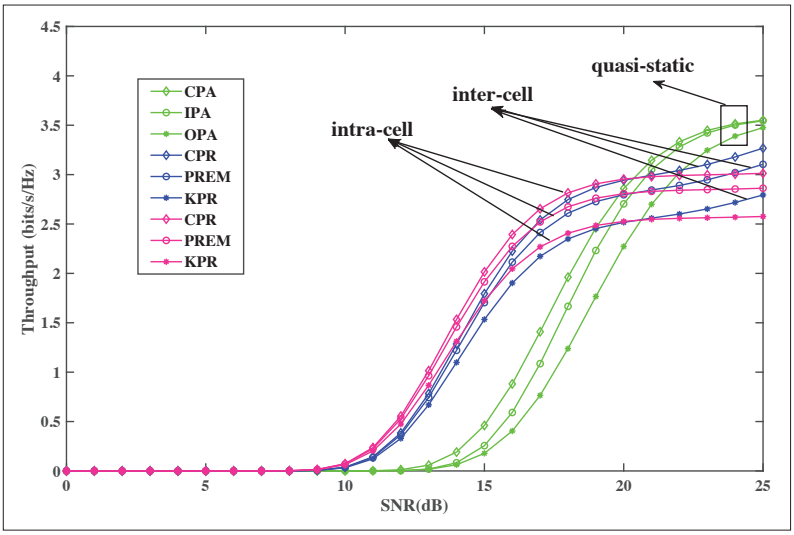

FIGURE 3. The system sum throughput performance of intelligent hybrid NOMA transmission schemes.

Decoding Methods: The decoding procedure of the SP-NOMA random access scheme is summarized as follows. Assume that the system has A activated UEs; each activated UE randomly chooses a pilot and then superimposes on its data to perform access to HTS. By utilizing the statistical information of received signal power and the pilot codebook, the HTS identifies $(A - a)$ collision UEs and estimates the total interference power. Then, the HTS optimizes the power allocation coefficient ε between pilot and data via the ICER algorithm. Next, the received power of singleton UEs are sorted in descending order as $P_1 \geq P_2 \geq \ldots P_a$. Then, the HTS utilizes the successive interference cancellation (SIC) hard decoding method, or successive joint decoding (SJD) soft decoding method, to recover the single pilots, that is, the pilot is chosen by only one UE. Finally, the HTS broadcasts the access status of activated UEs and the optimal power allocation coefficients.

The HTS can select the SIC decoding method with less complexity, or the SJD decoding method with superior decoding performance. The SJD decoder decodes all a single UEs' signals conjointly. Let R_c denote the code rate of activated UE, if the signal to interference plus noise ratio (SINR) of all a single UEs are less than R_c or the decoding process of SJD fails, the HTS regards the weakest single UE (i.e., the a-th UE) as interference, and then compares the SINR of residual $a - 1$ single UEs with R_c. This procedure repeats until the decoding process is successful or there are no more single UEs to be recovered. In contrast, in each step of SIC, the strongest signal with the largest received power P_1 is decoded by regarding the rest of the signals as interference. If the strongest single UE is decoded successfully, the decoded signal would be subtracted from the total received signals. Then, the HTS decodes the second strongest signal with P_2. The SIC decoding stops when a single UE fails to be decoded or all single UEs are decoded successfully.

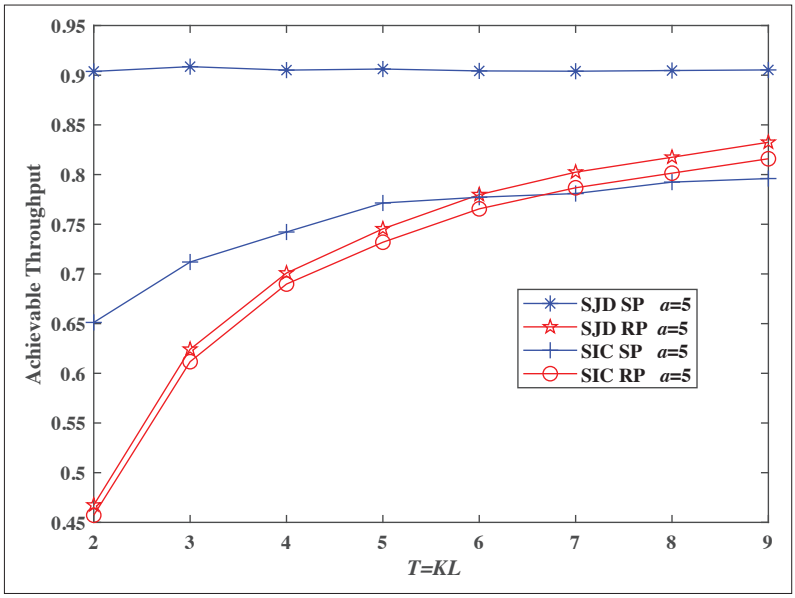

FIGURE 4. The corresponding achievable throughput versus different packet length under SP-NOMA and RP-NOMA when SJD and SIC methods are used, where R_c = 0.2. It can be seen from the simulation results that SJD achieves better achievable throughput performance than that of SIC at the cost of higher decoding complexity; The achievable throughput of SP-NOMA is higher than that of RP-NOMA scheme when $L \leq 5$.

Since the total interference power of collision UEs is important in the decoding performance of SIC and SJD, it is essential to derive the corresponding expression of total interference power in closed form. Then, we can further derive the theoretical performance of our SP-NOMA scheme.

For performance comparison, we employ the grant-free random access and SIC/SJD decoding methods in the RP-NOMA scheme. As shown in Fig. 4, we investigate the performance of the system achievable throughput versus the normalized data packet length T and pilot length L = 1 under the RP-NOMA and SP-NOMA schemes with the SJD and SIC methods, respectively. The abscissa in Fig. 4 represents the ratio K of packet length T to pilot length L. We can observe that the system achievable throughput of SP-NOMA is about twice higher than that of RP-NOMA, when the data payload is equal to the pilot sequence length (i.e., L = 2), and validates that the SP-NOMA scheme is suitable for short packet communications.

Hybrid Pilot NOMA Scheme: In short packet communications, the transmission of pilot sequence and data payload in the RP-NOMA scheme is in an orthogonal manner, and it may result in rate loss. In comparison with the RP-NOMA scheme, the SP-NOMA scheme requires no additional time resource reserved for pilots, and thereby can achieve a higher spectral efficiency. However, the SP-NOMA scheme introduces the interference between pilot and data, which leads hard to get perfect CSI. Therefore, the CE algorithm plays a key role in the SP-NOMA scheme. Furthermore, the power allocation between pilot and data is another important issue, which affects the estimation quality of the CE algorithm. Moreover, when the number of single UEs increases, the throughput performance of the SP-NOMA scheme degrades quickly, especially in the low SNR region. Therefore, a hybrid pilot NOMA (HP-NOMA) scheme as an optimal combination of the RP-NOMA and SP-NOMA schemes, could be proposed to further improve the OP and throughput, and the detailed derivation of optimal parameters for the HP-NOMA scheme is left in future work.

MSPA Random Access Protocol for a Crowded Scenario

Although grant-free random access can partially alleviate pilot collision in massive access, since the number of idle pilots decreases dramatically in a crowded scenario, the pilot contamination becomes the bottleneck in massive access. Note that the SURC and its improved versions have increased one access time slot to provide a secondary access opportunity for collision UEs [10, 11], there is no significant performance improvement on throughput and AFP in an overload case. Moreover, considering that the typical space communications can be characterized by delay tolerant but reliability-critical, there exists an inherent trade-off between the minimum access time slots and achievable AFP under limited pilot sequences, which allows for guaranteeing the desired AFP of UEs at the price of slightly increasing the access time slots. Thus, we propose a MSPA random access protocol, which can adjust an optimal number of access time slots in the overload case, where the activated UEs jointly transmit randomly chosen pilot sequences along with their data payload in multi-slot, which can satisfy the AFP requirement and raise the system throughput.

By introducing the graph representation for the MSPA random access protocol, the activated UEs are denoted as variable nodes, and the selected pilots in certain time slots are check nodes. Then, assume that the HTS utilizes SIC to resolve the received UEs and refer to the belief propagation (BP) decoding for an erasure code over a bipartite graph, we can derive the relationship to the AFP and system throughput under the number of activated UEs, pilots and access time slots in a finite length regime, which can be used to design a satisfying MSPA random access protocol. Thus, the jointly optimized MSPA random access protocol can significantly improve both the AFP and system throughput when extending the access time slots properly.

Grant-Free MSPA Random Access Protocol: An implementation of the MSPA random access protocol is introduced as follows. Initially, each activated UE will receive a control signal broadcasted by the HTS, allowing for estimating the average channel gain and synchronizing with HTS. Then, each activated UE randomly selects a pilot from a mutually orthogonal pilot set, and sends its data to the HTS in an appropriate way (SP or RP scheme). The HTS detects the power of received signals and utilizes the pilot codebook to determine whether a pilot is in collision, that is, the pilot is selected by more than one UE. Then, the HTS broadcasts the pre-coded random access response and reselection pilots (i.e., the pilots that can be used for reselection transmission in the subsequent time slots) information over the downlink channel to all activated UEs, which enables each UE to identify whether the collision is generated at the first transmission round. The collision UEs are permitted to reselect a pilot from the reselection pilot set in each subsequent time slot for accessing, while the

single UEs no longer participate in the pilot reselection phases. Maximum ratio combining is utilized by the HTS to merge all packet copies which belong to the same UEs, and then recover the data from superposition signals. At the end of the access time window, by utilizing the information cached in previous steps, the HTS performs SIC or SJD to recover the accessed UEs, and broadcasts the decoding result to all accessed UEs.

Figure 5 illustrates the AFP of MSPA and other existing random access schemes versus the system load, denoted by g. As can be observed, increasing the number of access time slots, denoted by ΔN, can significantly decrease the AFP. Moreover, in a transmission frame, expanding the access time slot appropriately can resolve more UEs, and raise the system throughput. Moreover, the MSPA scheme outperforms the SUCR-GBPA and SUCR-IPA schemes because both the idle pilots and collision pilots are utilized by the collision UEs for accessing in the MSPA scheme [11]. Specifically, when $\Delta N = 3$, the system throughput of the MSPA random access protocol reaches the peak for all considered values of system load g, and then decreases as $\Delta N > 3$. This indicates that there exists an optimum solution for achieving the maximum throughput in mMTC, which can be obtained by deriving the expressions of these key parameters according to the QoS requirements.

Unequal Access Latency Protection MSPA Random Access Protocol: Considering that the UEs may have diversity QoS requirements and coexist in a hybrid access scenario, to design and optimize the MSPA random access protocol with unequal access latency protection (UAL-MSPA) is important and should be further investigated in the massive access scenario of S-IoT. Inspired by the motivation, we propose two types of UAL-MSPA schemes in [13]. One is the independent UAL-MSPA (I UAL-MSPA) scheme, where the UEs are grouped by the access latency requirement, and different priority groups transmit in successive stages. The other is called the expand UAL-MSPA (E UAL-MSPA) scheme, in which the UEs with loose access latency and AFP requirements are scheduled to offload in the transmission stage of UEs with stringent access latency and AFP requirements.

In the I UAL-MSPA random access scheme, each UE group performs random access in different duration to avoid pilot collision. Without loss of generality, we assume that there are two priority groups, L_1 and L_2, which contain K_1 and K_2 activated UEs, respectively. The access latency threshold of L_1 is less than that of L_2, denoted by $(t_1 = \Delta N_1) < (t_2 = \Delta N_1 + \Delta N_2)$. Specifically, in stage 1, L_1 performs the MSPA random access protocol with pilot set $\mathbf{D_{11}} = \{\xi_1, \xi_2, \cdots, \xi_{\tau_p}\}$ allocated by the HTS, and L_2 remains silent and is scheduled to transmit in stage 2. The HTS recovers the data by utilizing received pilots and data signals in the previous ΔN_1 time slots, and sends an ACK to the resolved UEs in L_1. In stage 2, L_1 is not allowed to access, and L_2 performs the MSPA random access protocol with allocated pilot set $\mathbf{D_{22}} = \{\xi_1, \xi_2, \cdots, \xi_{\tau_p}\}$. At the end of stage 2, the HTS decodes data by utilizing the received pilots and data signals in the previous ΔN_2 time slots, and sends an ACK to the resolved UEs in L_2.

Based on the preliminary quantitative analysis for system throughput and access resource indicators (i.e., the number of time slots, UEs and pilots, and so on), we find that under the constraints of given pilot set $\mathbf{D} = \{\xi_1, \xi_2, \cdots, \xi_{\tau_p}\}$ and AFP, a maximum number of tolerable UEs, denoted by K^*, can be estimated. This means that when $K_1 < K^*$, a part of L_2 can be allowed to access in the first stage, which makes full use of the access resources of L_1, and effectively reduces the AFP of L_2. In stage 1 of the E UAL-MSPA scheme, the HTS allocates pilot sets $\beta\mathbf{D} = \mathbf{D_{11}} = \{\xi_1, \xi_2, \cdots, \xi_v\}$ and $(1 - \beta\mathbf{D} = \mathbf{D_{12}} = \{\xi_{v+1}, \xi_{v+2}, \cdots, \xi_{\tau_p}\}$ for L_1 and L_2, respectively, where $1 \le v \le \tau_p$ and $\mathbf{D_{11}} \cup \mathbf{D_{12}} = \mathbf{D}$. At the end of stage 1, the HTS sends an ACK to the resolved UEs in L_1 and L_2. In stage 2 of the E UAL-MSPA scheme, the HTS allocates the pilot set $\mathbf{D_{22}} = \{\xi_1, \xi_2, \cdots, \xi_{\tau_p}\}$ for the unresolved UEs in L_2, and decodes data by utilizing the received pilots and data information in the previous $\Delta N_1 + \Delta N_2$ time slots.

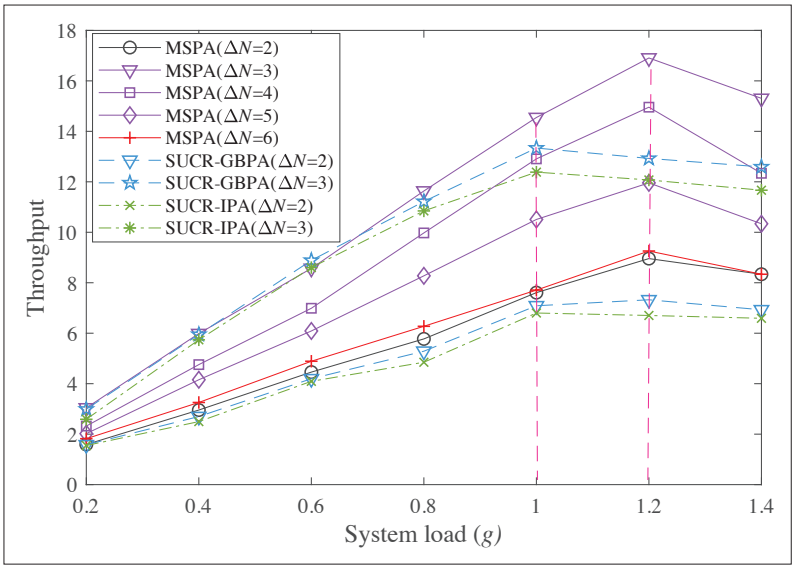

FIGURE 5. The access failure probability and throughput of MSPA random access protocol versus the system load.

Figure 6 illustrates the sum throughput as a function of the number of UEs K, including the I UAL-MSPA, E UAL-MSPA schemes and other relevant schemes without UAL protection, where $K = K_1 + K_2$ and $K_1 = K_2$. In comparison with the SUCR-IPA and SUCR-GBPA schemes, the I UAL-MSPA and E UAL-MSPA schemes show significant gains in sum throughput when $K > 100$, as the offloading design can alleviate the pilot collision of L_2. This validates the potential of offloading transmission to improve the system throughput. Moreover, the sum throughput of the E UAL-MSPA scheme converges to that of the I UAL-MSPA scheme as $\alpha \to 0$ and $\beta \to 1$, indicating that the E UAL-MSPA scheme has the flexibility to cope with a higher overload case. Finally, investigating the optimal parameters of the UAL MSPA random access scheme and the intermittent activations with more than two groups are important matters to be addressed in future work.

Future Directions for Massive Access in S-IoT

The timeliness of information is of paramount importance for the emerging massive access services in the upcoming S-IoT network. To evaluate the timeliness of information, a metric named the age of information (AoI) is introduced to model

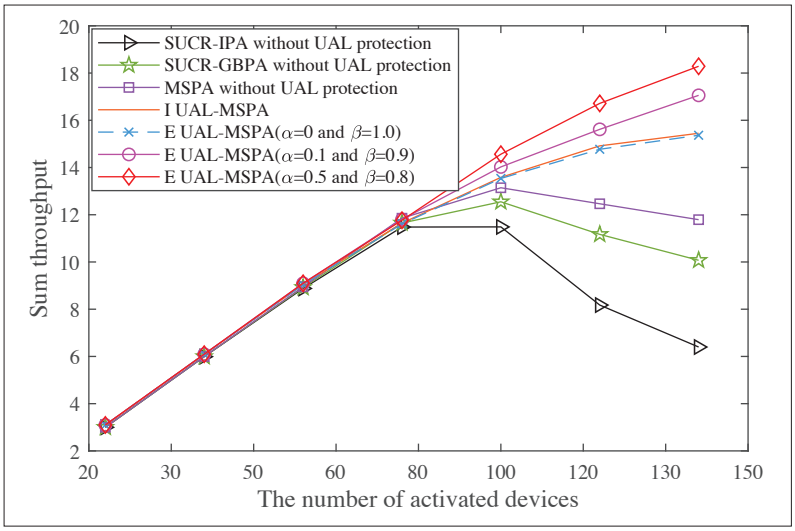

FIGURE 6. The sum throughput of the I UAL-MSPA, E UAL-MSPA and other relevant with no UAL protection designed random access schemes versus the number of activated UEs K, where $\Delta N_1 = \Delta N_2 = 2$, $K = K1 + K_2$ and $K_1 = K_2$.

the freshness of mMTC and MCC in S-IoT, which is defined as the difference between the current time and the time instant when the status update is generated. Moreover, short code design for short packet communications is an urgent need. In addition, consider the storage and power resources are constrained in HTS, how to design an age-optimal resource allocation policy to make full use of the multi-dimensional limited resources in the S-IoT network is well worth studying.

AoI Modeling and Optimization for Reliable Transmission

Information freshness is a stringent requirement for the above mentioned timeliness services in S-IoT networks. Nevertheless, the S-IoT channel can be characterized by high bit error rate and long propagation delay due to the huge communication distances, and frequent link disruptions caused by the dynamic topology. Thus, the bit-level forward error correction (FEC) approaching the Shannon limit still cannot guarantee the reliable transmission of data packets, and the erroneously decoded packet is treated as lost and needs retransmission. However, the conventional reliable transmission protocol, such as hybrid automatic repeat request (HARQ), is inefficient in S-IoT, since the retransmission of lost packets would significantly increase the end-to-end delay due to the long propagation delay as well as the AoI in S-IoT. The consultative committee for space data systems released the long erasure code (LEC) specification, where a packet-level LEC is proposed at the link layer to help mitigate the packet loss by physical layer bit-level FEC. Thus, the receiver can utilize packet-level LEC to help recover the failure decoded bit-level FEC. Therefore, in order to enhance the freshness and lower the AoI for reliable transmission in S-IoT, it is worthwhile designing a joint age-optimal HARQ transmission scheme to enhance the timeliness and ensure reliability.

Short Code Design for mMTC and MCC

In the future S-IoT, the most common traffic in MCC and uplink mMTC applications are short packet communications, which only involve several hundreds of bits. Although modern channel coding provides a practical approach to meet the Shannon limit with long block-length, the use of long codes for transmitting short packets will incur large overhead and result in long delays. Therefore, achieving high efficient mMTC and MCC mandate the channel codes of short block-length. It has been shown that the existing high performance finite block-length coding schemes can approach the Polyanskiy-Poor-Verdú meta-converse bound under maximum likelihood decoding (MLD) sense decoder [14]. On one hand, the code rates of all those schemes are fixed, and the transmitter needs to wait for CSI estimation and feedback to perform adaptive coding and modulation. If the receiver fails to decode the data block correctly, it will either drop the message or request another transmission. These feedback and retransmissions will significantly increase the end-to-end delay due to the non-trivial propagation delay in S-IoT networks. On the other hand, the MLD sense decoder is too complex for practical implementation and introduces significant processing latency. Decoding latency and feedback latency are directly related to the decoder complexity and code block-length, and the impact of decoder complexity, reliability, and latency should be jointly considered in practical operations. Therefore, to resolve these problems, we need to design powerful rateless channel codes to approach the short code performance limit, and develop near-optimal low complexity decoding for short codes.

Furthermore, the S-IoT wireless channel can be characterized by high propagation loss and time-varying fading, therefore ultra-high reliability can only be achieved at very high SNR. For example, to achieve BLER $\leq 10^{-9}$ for a short code with block-length of hundreds of bits, SNR $\geq$ 90 dB is required [15], which may be unrealistic in most power-constrained S-IoT devices. To address this issue, the short codes need to be augmented with some form of diversity techniques. With diversity orders 8 and 16, the required SNR is significantly reduced to 18 dB and 9 dB, respectively [15]. Note that space, frequency and time diversity can be achieved by using multiple antennas, or multiple frequency subcarriers of independent fading or transmissions in different time slots, respectively. Thus, a joint coding and space/frequency/time diversity design is essential, such as code-domain NOMA, to achieve reliable transmission of time/age-critical mMTC and MCC applications in S-IoT networks.

Multi-Dimensional Resource Optimization for Massive Access

The envisioned S-IoT network promises various time-critical and high throughput massive access scenarios. Since the average/peak power, storage and link duration are limited for most nodes in S-IoT networks, an age-optimal multi-dimensional resource allocation policy urgently needs to be addressed. Note that the Markov Decision Process (MDP) method in the AoI minimization problem encounters the problem of exponentially exploding state space, and huge computation complexity brought by the increasing number of network variables, which is known as the curse of dimensionality. To prevent the MDP system

from stepping into the curse of dimensionality, the Lyapunov optimization framework is recognized as the underlying tool to decouple the long-term optimization problem over multi-slot into a set of single time slot deterministic optimization problems [4]. Therefore, we can derive an age-optimal resource allocation policy via the Lyapunov optimization framework, subject to the long/short-term power, network stability and minimum throughput constraints in S-IoT network.

Conclusion

In this article, we have highlighted the significance of massive access communication as a key enabling technology for future S-IoT networks. However, there are still fundamental challenges ahead for the practical implementation of massive access in S-IoT, for example, when it comes to intelligent hybrid NOMA transmission, grant-free SP-NOMA and MSPA random access protocols, which are detailed. This provides researchers both in academia and industry with a promising research potential, and some future research directions for S-IoT are revealed, including AoI modeling and optimization, short code design for mMTC and MCC, and age-optimal multi-dimensional resource allocation for the massive access.

Acknowledgment

This work was supported in part by the National Natural Sciences Foundation of China (NSFC) under Grant 62071141, Grant 61871147, Grant 61831008, and Grant 62027802; in part by the Shenzhen Basic Research Program under Grant GXWD20201230155427003-20200822165138001; in part by the Natural Science Foundation of Guangdong Province under Grant 2020A1515010505; in part by the Guangdong Science and Technology Planning Project under Grant 2018B030322004; and in part by the project The Verification Platform of Multi-tier Coverage Communication Network for Oceans under Grant LZC0020.

References

[1] T. Ho *et al.*, "Next-Generation Wireless Solutions for the Smart Factory, Smart Vehicles, the Smart Grid and Smart Cities," 2019, arxiv: 1907.10102; available: https://arxiv.org/abs/1907.10102.
[2] J. Liu *et al.*, "Space-Air-Ground Integrated Network: A Survey," *IEEE Commun. Surveys and Tutorials*, vol. 20, no. 4, 2018, pp. 3645–76.
[3] Y. Wu *et al.*, "Massive Access for Future Wireless Communication Systems," *IEEE Wireless Commun.*, vol. 27, no. 4, 2020, pp. 148–56.
[4] J. Jiao *et al.*, "Network Utility Maximization Resource Allocation for NOMA in Satellite-Based Internet of Things," *IEEE Internet of Things J.*, vol. 7, no. 4, 2020, pp. 1391–1404.
[5] H. Chen *et al.*, "Ultra-Reliable Low-Latency Cellular Networks: Use Cases, Challenges and Approaches," *IEEE Commun. Mag.*, vol. 56, no. 12, 2018, pp. 119–25.
[6] J. Jiao *et al.*, "Design and Analysis of Novel Ka Band NOMA Uplink Relay System for Lunar Farside Exploration," *China Commun.*, vol. 17, no. 7, 2020, pp. 1–14.
[7] E. Paolini *et al.*, "Coded Random Access: Applying Codes on Graphs to Design Random Access Protocols," *IEEE Commun. Mag.*, vol. 53, no. 6, 2015, pp. 144–50.
[8] G. Liva, "Graph-Based Analysis and Optimization of Contention Resolution Diversity Slotted ALOHA," *IEEE Trans. Commun.*, vol. 59, no. 2, 2011, pp. 477–87.
[9] Č. Stefanović, P. Popovski, and D. Vukobratovic, "Frameless ALOHA Protocol for Wireless Networks," *IEEE Commun. Lett.*, vol. 16, no. 12, 2012, pp. 2087–90.
[10] E. Björnson *et al.*, "A Random Access Protocol for Pilot Allocation in Crowded Massive MIMO Systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, 2017, pp. 2220–34.
[11] H. Han, Y. Li, and X. Guo, "A Graph-Based Random Access Protocol for Crowded Massive MIMO Systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, 2017, pp. 7348–61.
[12] J. Cui *et al.*, "Unsupervised Machine Learning-Based User Clustering in Millimeter-Wave-NOMA Systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, Nov. 2018, pp. 7425–40.
[13] J. Jiao *et al.*, "Unequal Access Latency Random Access Protocol for Massive Machine-Type Communications," *IEEE Trans. Wireless Commun.*, vol. 19, no. 9, 2020, pp. 5924–37.
[14] M. Shirvanimoghaddam *et al.*, "Short Block-Length Codes for Ultra-Reliable Low-Latency Communications," *IEEE Commun. Mag.*, vol. 57, no. 2, 2019, pp. 130–37.
[15] N. Johansson *et al.*, "Radio Access for Ultrareliable and Low-Latency 5G Communications," *Proc. IEEE ICC2015*, London, June 2015.

Biographies

Jian Jiao received the Ph.D. degrees in communication engineering from the Harbin Institute of Technology (HIT) in 2011. He is an associate professor with the school of Electrical and Information Engineering, Harbin Institute of Technology Shenzhen, and also an associate professor with the Peng Cheng Laboratory, Shenzhen, China. His current interests include error control codes, satellite communications, and massive random access.

Shaohua Wu received the Ph.D. degree in communication engineering from Harbin Institute of Technology, Harbin, China, in 2009. He is a professor with the School of Electrical and Information Engineering, Harbin Institute of Technology Shenzhen, and also an associate professor at the Peng Cheng Laboratory, Shenzhen. His current research interests include wireless image/video transmission, space communications, advanced channel coding techniques, and B5G wireless transmission technologies.

Rongxing Lu is currently an associate professor at the Faculty of Computer Science (FCS), University of New Brunswick (UNB), Canada. His research interests include applied cryptography, privacy enhancing technologies, and IoT-Big Data security and privacy. He has published extensively in his areas of expertise, and is the recipient of eight best (student) paper awards from several reputable journals and conferences. Currently, he serves as the Vice-Chair (Conferences) of IEEE ComSoc CIS-TC (Communications and Information Security Technical Committee). He is the winner of the 2016-17 Excellence in Teaching Award, FCS, UNB.

Qinyu Zhang received the Ph.D. degree in biomedical and electrical engineering from the University of Tokushima, Japan, in 2003. He has been a full professor and the Dean of the School of Electrical and Information Engineering, Harbin Institute of Technology Shenzhen. He received the National Science Fund for Distinguished Young Scholars, Young and Middle-Aged Leading Scientist of China, the Chinese New Century Excellent Talents in University, and three scientific and technological awards from governments. His research interests include aerospace communications and networks, wireless communications and networks, cognitive radios, signal processing, and biomedical engineering.

The envisioned S-IoT network promises various time-critical and high throughput massive access scenarios. Since the average/peak power, storage and link duration are limited for most nodes in S-IoT networks, an age-optimal multi-dimensional resource allocation policy urgently needs to be addressed.