OPEN CALL ARTICLE

Short-Packet Communications: Recent Advances and Research Challenges

Thai-Hoc Vu, Ming Zeng, Sunghwan Kim, H. Vincent Poor, and Quoc-Viet Pham

Abstract

Short-packet communications (SPC) is a key enabler for ultra-reliable low latency communications (URLLC). Unlike the classical asymptotic Shannon regime with long blocklengths, SPC renders standard Shannon capacity inadequate for measuring throughput, necessitating new performance metrics for evaluating emerging mission-critical applications. However, engineering SPC systems presents a formidable technical undertaking, primarily attributable to the intricate rate functions prevalent in the short blocklength regime. Moreover, it becomes necessary to integrate SPC with current wireless technologies and the latest advances towards 6G wireless networks to enhance their URLLC capabilities. In this article, we provide a concise review of SPC, encompassing its fundamental principles, notable research studies, and recent advances. Drawing upon this review, we outline the key challenges that SPC faces in the context of future wireless networks and explore promising solutions.

Introduction

Short-packet communications (SPC) has emerged as a key technology to meet stringent requirements of Internet-of-Things (IoT) applications and services in today's 5G networks and future 6G wireless systems, such as tactile Internet, autonomous driving, and telesurgery. These applications raise the challenges of meeting two critical conditions for ultra-reliable low latency communications (URLLC): a latency of less than 1 ms and a packet success rate greater than 99.999% [1]. Unlike analyses of communication systems based on the assumption that the coding blocklength can approach infinity, Shannon's classical asymptotic capacity analysis is no longer accurate for throughput analysis of SPC, which gives rise to a more complex rate function in the finite blocklength regime [2]. As a result, the design of SPC systems is more challenging, especially in large-scale and heterogeneous networks.

SPC has attracted considerable research attention in connection with meeting URLLC requirements for advanced wireless networking applications, with typical concerns including channel estimation, beamforming design, resource allocation, and performance analysis. The pronounced interplay between SPC and wireless technologies has been demonstrated by integrating SPC capabilities with current 5G technologies, such as transceiver transmission with multiple-input multiple-output (MIMO) and full-duplex (FD), enhanced spectrum utilization strategies of cognitive radio (CR) and non-orthogonal multiple access (NOMA), cost-effective solutions involving wireless power transfer (WPT), and physical-layer security (PLS). Additionally, there are considerable research efforts in extending SPC abilities with emerging 6G technologies, such as reconfigurable intelligent surfaces (RISs), rate-splitting multiple access (RSMA), Terahertz (THz) communications, and age-of-information (AoI). Especially, to meet more stringent requirements of envisioned future intelligent wireless systems, advances in deep learning (DL) approaches can establish new ways to design, optimize, and analyze SPC systems, typically challenging with conventional mathematical tools.

This article first presents the preliminaries of SPC, then investigates recent advances, and finally explores major challenges for future research. In particular, the main contributions of this work are summarized as follows.

- We provide an introduction to SPC, including the fundamentals, applications, and comparison with classical communications in the infinite blocklength regime.
- We review state-of-the-art research on SPC and investigate its interplay with current 5G technologies.
- We investigate and explore recent advances within SPC in the context of envisioned future 6G wireless systems.
- Finally, we discuss major research challenges to be further addressed to realize new IoT applications and meet anticipated 6G wireless requirements.

This work's contributions are presented in the following two sections, and the conclusion is provided in the final section. The list of frequently used acronyms is shown in Table 1.

Overview of Short-Packet Communications

This section first overviews fundamentals and applications of SPC, and then concisely presents the state-of-the-art interplay between SPC and 5G enabling technologies.

Thai-Hoc Vu is with the School of Electronic Engineering, Kyonggi University, Suwon 16227, Republic of Korea; Ming Zeng is with the Department of Electrical and Computer Engineering, Laval University, Quebec, QC G1V 0A6, Canada; Sunghwan Kim (corresponding author) was with the School of Electronic Engineering, Kyonggi University, Suwon 16227, Republic of Korea. He is now with the Department of Artificial Intelligence and Information Technology, Sejong University, Seoul 05006, Republic of Korea; H. Vincent Poor is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA; Quoc-Viet Pham is with the School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, D02 PN40 Ireland.

Digital Object Identifier: 10.1109/MNET.2025.3604248
Date of Current Version: XX XXXX XXXX
Date of Publication: XX XXXX XXXX

Abbreviation	Description	Abbreviation	Description
5G	Fifth generation	6G	Sixth generation
AoI	Age-of-information	BLER	Block-error rate
DL	Deep learning	FD	Full-duplex
IoT	Internet-of-Things	MIMO	Multiple input multiple output
NOMA	Non-orthogonal multiple access	PLS	Physical layer security
RIS	Reconfigurable intelligent surface	RSMA	Rate-splitting multiple access
SIC	Successive interference cancellation	SNR	Signal-to-noise ratio
SPC	Short-packet communications	THz	Terahertz
URLLC	Ultra-reliable low-latency communications	WPT	Wireless power transfer

TABLE 1. List of frequently used abbreviations.

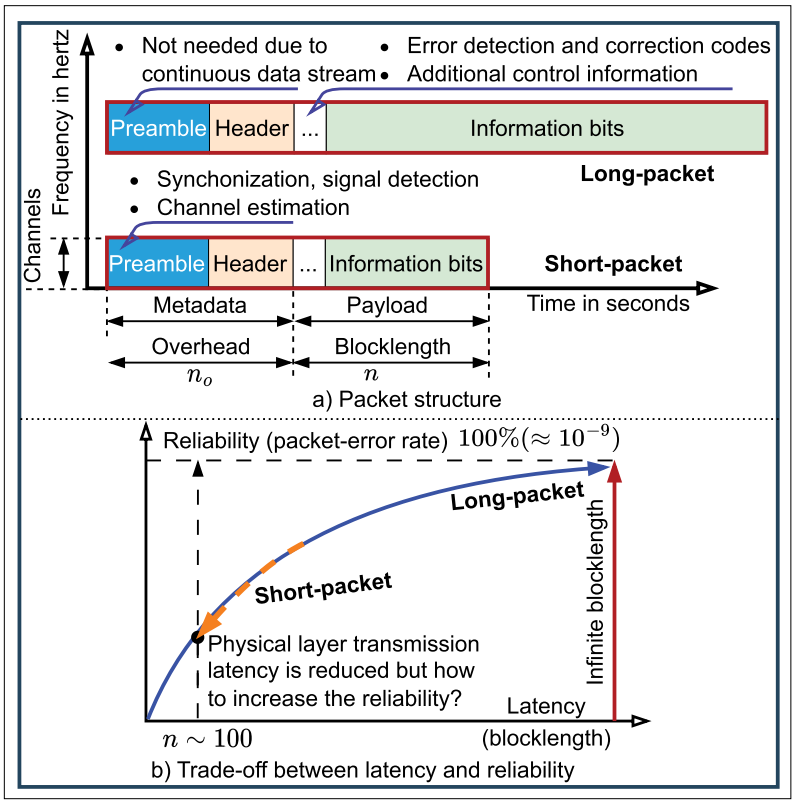

FIGURE 1. a) Illustration of blocklength structure in SPC and b) Trade-off between reliability and latency in SPC.

Short-Packet Communications

In wireless communication systems, how to enhance communication reliability at reduced transmission latency has become a critical issue, with the proliferation of mission-critical machine-type communications that aims to achieve a reliability of 99.999% (or packet error rate of 10^{-5}) and latency of less than one millisecond simultaneously. In this context, SPC has emerged as a key technology from the perspective of physical layer transmission. A fundamental packet structure consists of two main components: metadata and payload, as shown in Fig. 1(a). When metadata is considered overhead and excluded from the core packet, the packet length, commonly referred to as the blocklength in the literature, is defined as the number of channel uses allocated to transmitting the payload (i.e., information bits plus any forward error correction redundancy/control bits). This quantity is measured in channel uses (c.u.) and is approximately equal to the product of the channel bandwidth (in Hz) and the transmission duration (in seconds) [3]. In structures, the size of metadata and payload in short packets are approximately equal, as opposed to long packets. Specifically, the 3GPP standard for 5G, released in 2016 [1], defines a packet as short if it comprises no more than 32 bytes of data; otherwise, it is deemed a long packet. Table 2 summarizes their main differences.

SPC achieves reduced transmission latency due to the finite blocklength used, but encounters more error-prone packets during transmission. This leads to a reduction in communication reliability, measured by the system's packet-error rate compensation, as shown in Fig. 1(b). In the context of strict blocklength constraints (i.e., transmission latency), system reliability can be enhanced in two ways: operating at a lower coding rate for increased error correction redundancy, or using diversity techniques like frequency and antenna diversity to mitigate deep fades across multiple channel realizations. However, the former may violate some data requirements from application conditions, while the latter also has certain limitations. Specifically, in fast-fading environments where the channel changes on the order of, or faster than, the transmitted packet duration, antenna and spatial diversity techniques can fall short since pilots needed for channel estimation eat into the coherence time. This leaves outdated or noisy channel state information that erodes diversity gain and leads to SNR loss and reliability degradation. Furthermore, frequency diversity requires extra spectrum (multiple carriers or wide spreads), which may be limited in constrained frequency bands, especially in IoT networks. In such situations, employing channel coding schemes like polar, turbo, and low-density parity-check codes can enhance system reliability. However, with short blocklengths, they offer limited gains, challenging the design of coding schemes in SPC systems and analysis of theoretical performance limits. For further discussion of this point, the reader is referred to [1].

Based on a tight bound on the maximal coding rate $r \approx k/n$ of mapping k bits of information into a codeword of blocklength $n \geq 100$, Polyanskiy et al. [2] laid an initial foundation to evaluate an upper bound on the performance of SPC over an additive white Gaussian noise (AWGN) channel, showing that the block error rate (BLER) can be written as $\epsilon(n,k) \approx Q\left([C(\gamma) - r]/\sqrt{V(\gamma)/n}\right)$. Herein, $C(\gamma) = \log_2(1+\gamma)$ is the Shannon capacity of complex channels for a given SNR γ, $V(\gamma) = (1-1/(1+\gamma)^2)/\ln^2(2)$ is the channel dispersion, and $Q(\cdot)$ is the Gaussian Q-function. Afterwards, Devassy et al. [4] formulated the average delay for sending a coded packet with parameters (n, k) under AWGN collision channel with unlimited numbers of retransmissions as $d = n/[1-\epsilon(n,k)]$. Thus, if T is the total transmission duration of using blocklength n used to successfully spread k bits over B KHz bandwidth (i.e., $n = TB$), the latency L in millisecond (ms) is measured by $L = dT \approx nd/B = n^2/[1-\epsilon(n,k)]/B$.

State-of-the-Art

Over the past decade, finite blocklength theory [2] has evolved from multiple stages of research before blossoming in 5G and beyond, where URLLC has become a top priority constraint. As a key to achieving this constraint, research on SPC has received considerable interest from various viewpoints.

1) Transceiver Design: In term of transceiver designs, MIMO is a technology that exploits multiple transmitting and receiving antennas either through diversity techniques to improve signal quality for a single user or through spatial-multiplexing strategies to serve multiple users simultaneously. These features enable SPC systems to achieve higher reliability and lower latency. For instance, spatial multiplexing MIMO significantly reduces user latency by multiplexing and transmitting multiple packets across separate antennas. However, this approach has resource limitations, i.e., high power consumption, complex signal processing, and high feedback overhead due to the spatial degrees of freedom demands. Conversely, in low-complexity scenarios, diversity technique is preferred to improve reliability by exploiting transmit antenna selection/maximum-ratio transmission at the transmitter and selection combining/maximal-ratio combining at the receiver. Towards large-scale IoT, massive/cell-free massive MIMO contributes significantly to achieving massive SPC with the zero-forcing and conjugate beamforming techniques [5].

Besides MIMO, another promising choice is FD technology, which uses one or multiple groups of antennas to simultaneously transmit and receive data between nodes. The features of FD enables SPC systems to overcome the barrier of transmission latency and help enhance communication reliability. An evident benefit of this is in transceiver systems with limited antenna interference cancellation capabilities, where increasing blocklength helps improve the BLER of FD transmission while adopting FD modes significantly reduces transmission latency, a unique advantage over half-duplex SPC. However, the interplay between FD technology and SPC still faces challenges related to the structure and signal-processing capabilities of target nodes, where users with limited hardware may struggle to implement FD transmission in practice, resulting in dedicated relay systems with powerful computing capabilities more preferable [6].

2) Enhanced Spectrum Utilization: NOMA addresses spectrum inefficiency inherent in orthogonal multiple access schemes by multiplexing user signals in the power domain rather than separating them by time or frequency. Successive interference cancellation (SIC) is required at the receiver to extract user data in descending order of signal power. In large-scale IoT scenarios, where the SPC systems typically face stringent spectrum constraints and resource management challenges, introducing NOMA into SPC systems can help bridge these gaps. In particular, downlink NOMA enables simultaneous service to multiple users within a single packet by superimposing their information, and this blocklength can be minimized by optimizing power allocation coefficients among users [7]. Further, NOMA also promotes fairness by assigning more power to users with weaker channel conditions in the downlink and to near users in uplink, thereby enhancing channel quality and facilitating demodulation for distant users. However, practical deployments may suffer from increased BLER compared to orthogonal schemes due to hardware impairments and imperfect SIC.

Aspects	Short-Packet Communications	Long-Packet Communications
Objective	Ultra-high reliability ($\geq$ 99.999%) and low-latency ($\leq$ 1 ms).	High transmission data rate and moderate-to-high reliability.
Message scheme	Finite blocklength [2].	Infinite blocklength.
Data amount	$\leq$ 32 bytes (256 bits) of data per packet [3].	> 256 bits of data per packet up to several million bits.
Key metric	BLER, goodput, and AoI	Outage probability, throughput, ergodic rate, and energy efficiency
Application	Massive and mission-critical machine-type communications in 5G (e.g., factory, tactile internet, vehicles-to-everything, metaverses), low-power wireless sensor networks (e.g., LoRa and SigFox).	Internet and wireless services with high throughput demands (i.e., file transfer, video streaming, and cloud gaming), enhanced mobile broadband in 5G, and super-enhanced mobile broadband in 6G.

TABLE 2. Comparison between SPC and Long-Packet Communications.

Aside from NOMA, CR offers an alternative approach to improving spectrum allocation through three dynamic access models: interweave, underlay, and overlay. The interweave model accesses spectrum gaps without disrupting primary users. The underlay model operates at low power within the same frequency bands as primary users to minimize interference. The overlay model enhances reliability and reduces latency by using advanced techniques to coexist with primary users while transmitting secondary data. With such diverse model implementations, introducing CR into SPC in large-scale IoT [8] not only enables multiple SPC systems to coexist within spectrum limits, but also helps mitigate congestion-related transmission delays. This ensures reliable service quality for URLLC applications.

3) Cost-Effective Solution: WPT technology offers a solution that eliminates the need for wired power sources or battery replacements, thereby extending the lifespan of power-limited systems. It utilizes two primary mechanisms: one involves a dedicated power station (the so-called wireless-powered communication network) that supplies energy to users in densely populated areas, while the other, namely simultaneous wireless information and power transfer, allows IoT devices and sensors to harvest energy directly from data signals, making it particularly suitable for scenarios such as small cell coverage areas or tunnels. Due to its cost-effectiveness, WPT has emerged as a powerful solution for IoT development in SPC systems [8], which typically comprise numerous resource-constrained sensors and IoT devices. However, there has been less focus on the theoretical analysis of nonlinear energy harvesting challenges in WPT-enabled SPC systems.

4) Physical Layer Security: PLS is concerned with providing security in wireless communications by leveraging physical channel properties, such as randomness, noise, fading, and interference, to secure information transmission and reduce complexity, alongside traditional cryptographic methods.

However, most existing PLS schemes based on asymptotic analyses are no longer applicable to SPC systems [9]. For example, in PLS-based randomness jamming signal systems, the training stage using finite blocklength may leak key information exchanged between the source and legitimate user, which can be exploited by an eavesdropper to subsequently cancel the jamming signal during communications. Thus, determining the secrecy reliability of SPC is a critical issue and requires further research [10], including developing secure strategies (e.g., artificial noise schemes) and addressing resource allocation issues (i.e., blocklength usage and channel bandwidth).

Recent Advances in SPC

This section discusses recent advances in wireless SPC-enabled systems, where the interplay with advanced technologies plays a crucial role in improving SPC performance.

Age-of-Information (AoI)

AoI measures the time elapsed from the generation of the latest information at a source to its successful reception at the destination [11]. As shown in Fig. 2(a), for real-time monitoring and management of traffic flow in intelligent transportation systems, sensors on vehicles and infrastructure continuously collect data on traffic conditions, vehicle speeds, and congestion levels, and then send these collected data to a central server or edge computing nodes via wireless channels. If sensors generate an update at time t_{i-1} and t_i is the current time for calculating the AoI, the AoI is $(t_i - t_{i-1})$. If there is any delay in transmitting the data to the server t_d and processing data t_p, the AoI is given by $(t_i - t_{i-1} - t_d - t_p)$.

Accordingly, by continuously measuring the AoI with each new packet, the SPC system ensures the freshest possible data for real-time decision-making in advanced URLLC applications, i.e., remote sensing, monitoring, control, and emergency response. Fig. 3 plots an example of AoI in cooperative SPC systems, where a source node uses the automatic repeat request retransmission mechanism to measure the AoI of sending a packet to the destination via aerial vehicles [11]. As observed, when the data size of a packet is significantly small (i.e., $k = 8$ bytes) and perfect channel estimation is ensured for end-to-end transmission, the AoI linearly increases with the blocklength n due to the increased service time for updates. Conversely, for larger packet sizes, the AoI is higher when $n \leq 60$ due to a higher BLER. Notably, when $k > 8$ bytes, increasing n toward its optimal regions (i.e., $n \approx 50$ for $k = 16$ and $n \approx 100$ for $k = 32$) decreases the AoI due to a reduction in BLER. However, further increases in blocklength beyond these optimal points increase the AoI, as the impact of service time on the AoI outweighs the reduced BLER. When the channel estimation process is less accurate, the BLER increases due to the presence of channel errors, necessitating a longer blocklength and resulting in an increased AoI.

Rate-Splitting Multiple-Access

RSMA is an advanced radio technology that splits user messages into common and private components, encoding the common parts into one or more common streams and the private parts into separate streams [12]. The transmitter then precodes these streams, either linearly or non-linearly, using channel state information, superposes them with different power levels, and transmits them to users over multiple antennas, as shown in Fig. 2(b). Users employ single/multiple layer SIC to separate the common stream from the private streams and decode their private stream while treating the others as interference.

Due to its universal principles, RSMA overcomes the limitations of NOMA in user scheduling and SIC procedures, achieving enhanced spectral, energy, and computational efficiency, making it well-suited for SPC. Fig. 4(a) shows the robustness

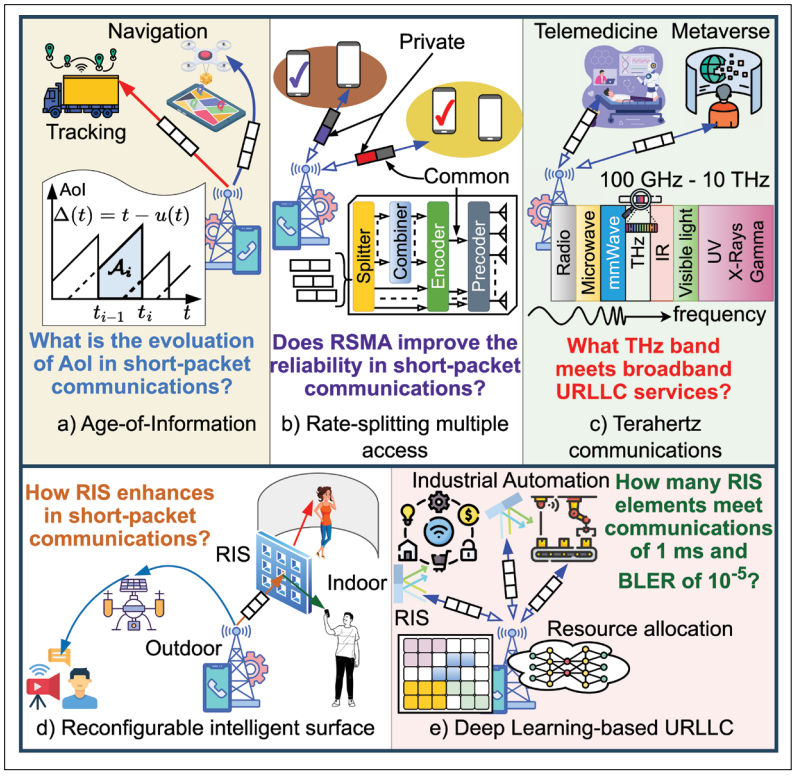

FIGURE 2. Illustration of recent advances in SPC. a) Age-of-Information. b) Rate-splitting multiple access. c) Terahertz communications. d) Reconfigurable intelligent surface. e) Deep Learning-based URLLC.

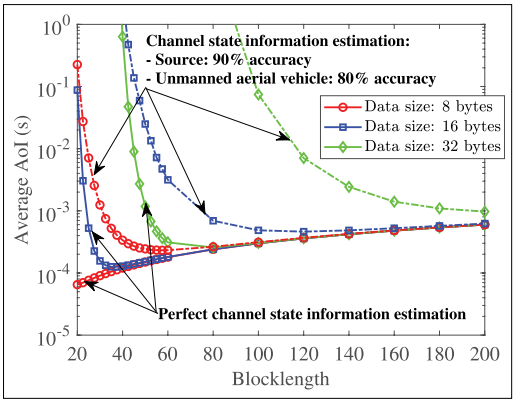

FIGURE 3. Impact of blocklength on the AoI in SPC systems.

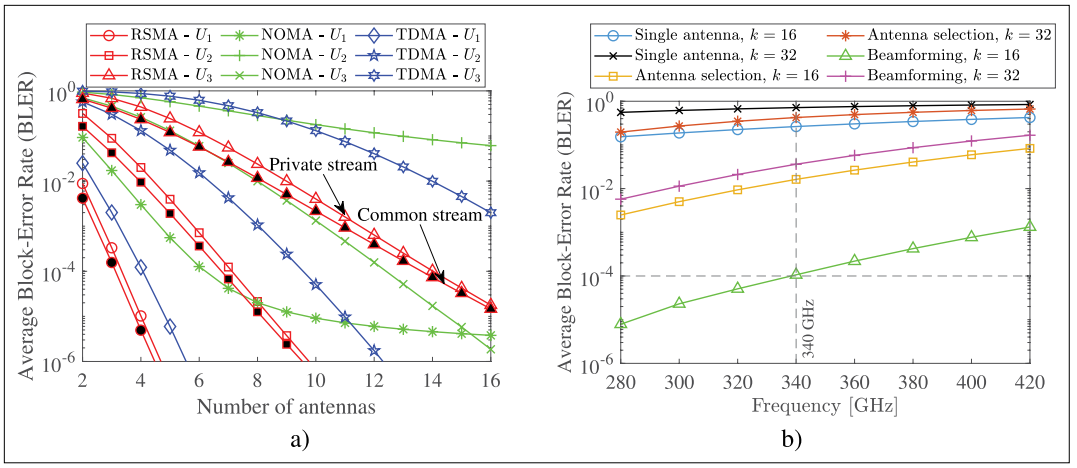

FIGURE 4. BLER performance in SPC systems: a) Impact of the number of antennas in multiple access and b) Impact of the usage frequency in THz communications.

of RSMA in improving BLER compared to NOMA and time-division multiple access (TDMA) when sending a packet of k = 10 bytes and n = 128 using the transmit power P = 10 dBm, where three users, U_1, U_2, and U_3, are located at (−10, 0), (20, 0), and (0, 30), respectively. In RSMA, the common and private streams each convey k = 5 bytes of data, with power allocations of 0.3P and 0.1P, respectively. For NOMA, the power allocations are 0.1P for U_1, 0.2P for U_2, and 0.7P for U_3, while TDMA allocates 1P per time slot. The optimal beamforming design for RSMA and NOMA follows a hybrid zero-forcing and maximal-ratio transmission strategy [12], while that of TDMA is the maximal-ratio transmission. With a 90% interference cancellation capability at users, increasing the number of transmitting antennas L raises the residual inter-channel interference, affecting the BLER for U_1 and U_2 in NOMA with asymmetric SIC. RSMA, with fair SIC among users, reduces the impact of imperfect SIC and improves the overall BLER for U_1 and U_2, but degrades performance for U_3 over NOMA. Notably, RSMA outperforms TDMA in all settings.

Terahertz Communications

THz communications are a promising technology to achieve large contiguous bandwidth and ultra-high-speed data rates while effectively addressing the issue of spectrum scarcity. To do so, this technology utilizes the license-free THz band from 0.1 to 10 THz, as shown in Fig. 2(c). However, transmitting large data packets over THz frequencies can result in high transmission error rates due to molecular absorption and spreading losses. Especially, the very small wavelengths of THz frequencies limit their effectiveness in short-range communications, posing significant barriers to the maturation of THz communications in 6G.

In this context, it is vital to employ SPC to reduce the amount of data transmitted in conjunction with diversity and beamforming techniques to enhance signal strength at the receiver in THz communications. The gains in THz communications are shown in Fig. 4(b), where the source node sends a short packet with n = 128 and P = 10 dBm to a single-antenna user within a 10 m range. As observed, higher THz frequencies degrade signal quality, increasing the BLER for the user. However, when the data amount is reduced (e.g., from 32 bytes to 16 bytes), the BLER decreases significantly. Notably, exploiting multi-antenna transmissions (e.g., 4 antennas) offer better BLER improvement than a single antenna transmission, where beamforming based on maximum-ratio transmission creates an approximate 100-fold performance gain when reducing the data amount from 32 bytes to 16 bytes across all frequencies. Furthermore, the beamforming approach with k = 16 bytes of data meets the BLER requirement of 10^{-4} for frequencies below 340 GHz.

Reconfigurable Intelligent Surfaces

RISs are a revolutionary technology that manipulates electromagnetic waves using two-dimensional surfaces composed of elements capable of passively reflecting, refracting, or simultaneously reflecting and refracting incident waves (the latter is known as a simultaneously transmitting and reflecting RIS [13]), as shown in Fig. 2(d). Since RISs instantly reflect or refract incoming signals to their intended destinations, they can operate similarly to FD communications without causing interference. Additionally, RISs can enhance communication quality by optimizing the phase shifts of their elements or incorporating amplification to counteract the effects of "*multiplicative fading*", which is often referred to as active RISs, in contrast to passive configurations.

These features of RISs are highly beneficial for reducing transmission latency in long-range SPC systems, such as dual-hop and multi-hop configurations. Notably, in multipath fading environments, where communication channels between actuator devices or between a source node and a user are severely impaired by large obstacles, deploying RISs can substantially boost network reliability and quality [14]. Fig. 5(a) illustrates the BLER for transmitting a short block message of k = 8 bytes with a block length of n = 256 over an AWGN channel with zero-mean and unit variance. The transmit SNR is γ = 10 dB, with source node positioned at (0, 0) and the RIS located at (50,50). As the user travels along the x-axis away from the source node while

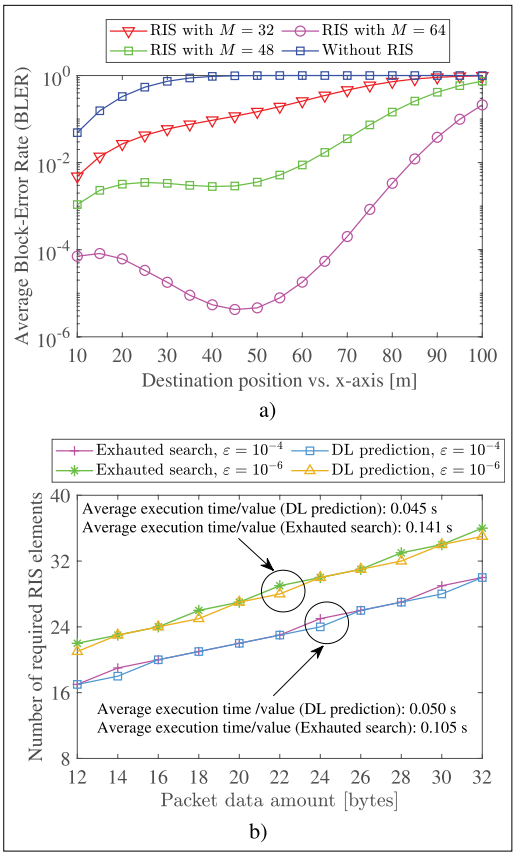

FIGURE 5. SPC in RIS-based systems. a) Average BLER comparison between RIS and non-RIS in SPC. b) Predicted number of RIS elements versus packet data amount.

maintaining a fixed y-coordinate of 10 m, its BLER without the aid of the RIS significantly increases due to wireless path loss. Meanwhile, exploiting RISs helps users achieve not only a lower BLER at certain locations but also an expanded communications coverage with the increment of RIS elements M. Furthermore, increasing the number of RIS elements, M, substantially reduces the BLER. Notably, there is an optimal region on the user's x-axis between (40, 50), where its BLER is minimized when M = 64.

Deep Learning

DL is a branch of machine learning that employ neural networks (NNs) to learn from data and generate accurate estimates [14]. Due to their capabilities to reduce computation and processing time, DL-based models have recently become indispensable in modern wireless systems to readily remedy critical issues, such as congestion control, resource allocation, and queue management. For systems less affected by medium factors, DL-based supervised learning approaches, including deep, convolutional, and recurrent NNs, and long short-term memory networks, can be trained offline using available datasets prior to deployment in online applications. Conversely, for highly dynamic systems, e.g., those involving random user mobility, DL-based unsupervised learning is often employed to adaptively learn from network environments, typically through reinforcement learning techniques.

Given that latency and reliability are paramount in URLLC applications, integrating DL into SPC as shown in Fig. 2(e) can enhance wireless communications by alleviating the burden of resource optimization while achieving desired outcomes [8]. Fig. 5(b) illustrates the prediction of the number of RIS elements required for real-time IoT applications, using a deep-optimized NN trained on a dataset generated by analytical approaches [14]. The figure shows that when a source node transmits a large amount of data to a user located at (40, 20) in meter with P = 10 dBm and a packet size of n = 100, the number of required RIS elements increases. Specifically, to meet the BLER requirement of 10^{-4}, transmitting a packet with 12 bytes of data requires approximately 17 RIS elements, while a packet with 32 bytes of data requires up to 30 RIS elements. This is because compressing a large amount of data into a packet with the same block length leads to a higher BLER during transmission. It is, thus, required to use RIS designs with more elements to compensate for such performance loss, thereby boosting communication reliability to meet stringent URLLC requirements. Also, the figure shows that the DL approach predicts results very close to those obtained from exhaustive search-based analytical methods, with an estimation error of around 1 RIS element, while executing much faster than the exhaustive search-based methods for any BLER target. For instance, the average execution time for a certain packet data amount with a BLER requirement of 10^{-6} is approximately 0.045 seconds for the DL approach, compared to 0.141 seconds for the exhaustive search.

Challenges and Open Research Directions

SPC has shown significant potential to be a key enabler of many critical IoT services and applications in both current and future 6G wireless networks. Nevertheless, major challenges still exist, which severely limit the performance of SPC networks and need to be addressed. This section outlines these issues and highlights potential solutions.

Physical Challenges

SPC can improve latency by reducing data sizes and enhance communication reliability through advanced finite blocklength channel coding paradigms. However, SPC systems face several physical challenges. Firstly, most studies based on finite-blocklength information theory argue that the header length is negligible compared to the packet data unit length, as encoded packets use different-rate channel codes for better header protection. Encapsulation processes are viewed as a consistent mapping from transport-layer bits to physical-layer symbols. Secondly, poor channel estimation quality can increase the demand for feedback channels, reducing the number of channels available for transmission. Thirdly, it is required to find a reasonable solution that can adjust the code rate and training sequence length based on the delay constraints and channel conditions. Finally, since SPC is susceptible to eavesdropping, developing new security schemes/strategies, transmission policies, beamforming designs, and efficient optimization algorithms is necessary.

Open Research Directions

1) Cross-Layer Transmission: Permutation transmission is a novel approach for seamless packet delivery that reduces latency and the rate of sending short blocks, particularly addressing header length issues relative to data unit length. At the transport layer, permutation-based transmission focuses on packet header length and payload design, conveying a portion of application-layer data through permutations with repetitions of varying lengths across a group of packets, rather than embedding them directly into packets. This method uses fewer resource units compared to classical methods [15].

Adopting this approach in SPC offers several benefits: increased goodput, improved resource utilization efficiency, reduced latency, and enhanced security against data tracing. At the physical layer, permutation-coded modulations enhance Euclidean distance properties by constructing random permutation codes and employing maximum-likelihood decoding. However, this approach is still in its early stages, with open issues such as ensuring secure transmission and reducing the complexity of encoder and decoder-based modulations.

2) Enhanced Channel Quality: Integrated sensing and communication (ISAC) is an emerging paradigm that combines sensing and communication functions within a single system, sharing hardware, spectrum, and signal processing frameworks to enhance wireless system efficiency and performance. The sensing component collects and processes information from noisy observations, while the communication component transmits and receives information using tuned signals.

Applying ISAC with SPC can further boost communication performance by leveraging sensing information for channel estimation, beamforming, and interference management. For example, ISAC-based SPC can enable joint radar and frequent small data transmissions for autonomous vehicles, improving situational awareness and safety. Despite its potential, designing efficient and robust ISAC waveforms that support both sensing and communications under strict latency and reliability constraints remains challenging, necessitating the development of unified theoretical frameworks and performance limits.

3) Improved Resource Allocation: NN codes, a type of channel code, use DL algorithms to construct encoding and decoding functions based on the main characteristics of classical channel coding schemes. Specifically, NN codes employ autoencoders to compress high-dimensional data into lower dimensions and then reconstruct it. In addition, mobile edge computing (MEC) is known as a potential solution for reducing end-to-end latency through local data processing, enhancing reliability by minimizing data travel distance and optimizing resource utilization through efficient task offloading.

From the dominant features of these two technologies, it is clear that integrating NN codes into SPC systems achieves higher code rates and lower packet error rates without prior knowledge of channel distribution, allowing flexible adaptation to various channel conditions, non-linear problems, and URLLC scenarios. Meanwhile, integrating MEC into large-scale IoT-enabled SPC systems increases scalability and flexibility in adapting to varying network conditions and user demands in dynamic environments, i.e., industrial automation, transportation, and augmented reality, while ensuring high performance and reliability in critical IoT scenarios.

4) Robust Data Integrity and Security: Covert communications are the practice of disguising communications between legitimate nodes to make it harder for illegal parties to monitor, thus better protecting sensitive information transmission.

Integrating covert communications into SPC makes URLLC systems more resilient to interference and jamming, as hidden transmissions are harder to disrupt. Moreover, it is also required to incorporate NN codes into SPC systems for learning finite blocklength wiretap codes with statistical reliability and secrecy constraints based on training data. This facilitates intelligent SPC systems to achieve vanishing decoding error probability at the legitimate receiver and minimal leakage to eavesdroppers in the asymptotic regime.

Conclusion

In this article, we have provided a concise review of wireless communications in the finite blocklength regime, also known as short-packet communications. SPC's unique features make it a promising technology for various services and applications in current and envisioned future 6G wireless systems. We have explored the amalgamation of SPC and wireless technologies as key to enhancing the performance of URLLC services. Based on our research and recent advances, we have finally discussed not only physical issues but also open research directions of cross-layer transmission, enhanced channel quality, improved resource allocation, and robust data integrity and security to stimulate further research on SPC.

Acknowledgment

This work was supported in part by the Research Program through the National Research Foundation of Korea under Grant NRF-2023R1A2C1003546, in part by NSERC through its Discovery Program, and in part by an Innovation Grant from Princeton NextG.

References

[1] M. Shirvanimoghaddam et al., "Short block-length codes for ultra-reliable low latency communications," *IEEE Commun. Mag.*, vol. 57, no. 2, pp. 130–137, Feb. 2019.

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

[3] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultra-reliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

[4] R. Devassy et al., "Finite-blocklength analysis of the ARQ-protocol throughput over the Gaussian collision channel," in *Proc. 6th Int. Symp. Commun., Control Signal Process. (ISCCSP)*, May 2014, pp. 173–177.

[5] A. A. Nasir et al., "Cell-free massive MIMO in the short blocklength regime for URLLC," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5861–5871, Sep. 2021.

[6] L. Yuan, Q. Du, and F. Fang, "Performance analysis of full-duplex cooperative NOMA short-packet communications," *IEEE Trans. Veh. Technol.*, vol. 71, no. 12, pp. 13409–13414, Dec. 2022.

[7] X. Lai et al., "Average secure BLER analysis of NOMA downlink short-packet communication systems in flat Rayleigh fading channels," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 2948–2960, May 2021.

[8] T.-H. Vu et al., "Performance analysis and deep learning design of wireless powered cognitive NOMA IoT short-packet communications with imperfect CSI and SIC," *IEEE Internet Things J.*, vol. 9, no. 13, pp. 10464–10479, Jul. 2022.
[9] W. Yang, R. F. Schaefer, and H. V. Poor, "Wiretap channels: Nonasymptotic fundamental limits," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4069–4093, Jul. 2019.
[10] C. Feng, H.-M. Wang, and H. V. Poor, "Reliable and secure short-packet communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1913–1926, Mar. 2022.
[11] T.-H. Vu et al., "Short-packet communications for UAV-based NOMA systems under imperfect CSI and SIC," *IEEE Trans. Cognit. Commun. Netw.*, vol. 9, no. 2, pp. 463–478, Apr. 2023.
[12] T.-H. Vu et al., "Rate-splitting multiple access-assisted THz-based short-packet communications," *IEEE Wireless Commun. Lett.*, vol. 12, no. 12, pp. 2218–2222, Dec. 2023.
[13] M. V. Katwe et al., "Spectrally-efficient beamforming design for STAR-RIS-aided URLLC NOMA systems," *IEEE Trans. Commun.*, vol. 72, no. 7, pp. 4414–4431, Jul. 2024.
[14] K.-T. Nguyen, T.-H. Vu, and S. Kim, "Performance analysis and deep learning design of short-packet communication in multi-RIS aided multi-antenna wireless systems," *IEEE Internet Things J.*, vol. 10, no. 19, pp. 17265–17281, Oct. 2023.
[15] W. Li, Y. Yang, and B. Jiao, "Permutation-based transmissions in finite blocklength regime: Efficient and effective resource utilization," *IEEE Trans. Commun.*, vol. 71, no. 6, pp. 3251–3262, Jun. 2023.

Biographies

Thai-Hoc Vu (Member, IEEE) (vuthaihoc1995@gmail.com) received the B.Sc. degree in electronics and telecommunications engineering from the Posts and Telecommunications Institute of Technology (PTIT) and the Ph.D. degree in electrical, electronic, and information systems engineering from the University of Ulsan (UoU), South Korea, in 2025. He is currently a Post-Doctoral Researcher in the short term at the School of Electronic Engineering, Kyonggi University, Republic of Korea. He is also a Lecturer at the Institute of Information Technology, Digital Transformation, Thu Dau Mot University, Ho Chi Minh, Vietnam. He has authored over 50 top-tier IEEE journal and flagship conference papers, accumulating more than 1000 citations on Google Scholar. During his Ph.D., he contributed over 20 top-tier IEEE publications as the first author. He specializes in performance analysis of wireless communication systems, with a focus on non-orthogonal multiple access, reconfigurable intelligent surfaces, short-packet transmission, and the application of convex optimization and machine learning. His accolades include the Vietnamese Young Scientists in Korea Award in 2022, five BK21 Best Research Awards from 2021 to 2025, the IEEE Communications Letters Top Reviewer Awards in 2023 and 2024, and the Best Paper Awards from IEEE ICCE from 2022 to 2024 and IEEE ATC in 2024.

Ming Zeng (Member, IEEE) (ming.zeng@gel.ulaval.ca) received the B.E. and master's degrees from the Beijing University of Posts and Telecommunications, Beijing, China, in 2013 and 2016, respectively, and the Ph.D. degree in telecommunications engineering from the Memorial University of Newfoundland, St John's, NL, Canada, in 2020. He is currently an Associate Professor and the Canada Research Chair with the Department of Electrical and Computer Engineering, Laval University, Quebec, QC, Canada. He has published more than 100 articles and conferences in first-tier IEEE journals and proceedings, and his work has been cited over 5,400 times per Google Scholar. His research interests include resource allocation for beyond 5G systems and machine learning-empowered optical communications. He serves as an Associate Editor for IEEE Transactions on Communications, IEEE Open Journal of the Communications Society, and IEEE Wireless Communications Letters.

Sunghwan Kim (Senior Member, IEEE) (skim@kyonggi.ac.kr) received the B.S., M.S., and Ph.D. degrees from Seoul National University, Seoul, South Korea, in 1999, 2001, and 2005, respectively. From 2005 to 2007, he was a Post-Doctoral Visitor with the Georgia Institute of Technology, Atlanta, GA, USA. From 2007 to 2011, he was a Senior Engineer with Samsung Electronics, Suwon, South Korea. From 2011 to 2024, he was a Professor with the Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Ulsan, South Korea. He is currently a Professor with the School of Electronic Engineering, Kyonggi University, Suwon. His research interests include 5G/6G communications, the IoT communications, multiple access, deep learning, transformer, error correction codes, and DNA-based storage.

H. Vincent Poor (Life Fellow, IEEE) (poor@princeton.edu) received the Ph.D. degree in EECS 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. He is currently the Michael Henry Strater University Professor at Princeton. During 2006 to 2016, he served as the dean of Princeton's School of Engineering and Applied Science, and he has also held visiting appointments at several other universities, including most recently at Berkeley and Caltech. His research interests include 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 book *Machine Learning and Wireless Communications* (Cambridge University Press, 2022). He is a member of the National Academy of Engineering and the National Academy of Sciences and is a foreign member of the Royal Society and other national and international academies. He received the IEEE Alexander Graham Bell Medal in 2017.

Quoc-Viet Pham (Senior Member, IEEE) (viet.pham@tcd.ie) received the B.Sc. and Ph.D. (Hons.) degrees in telecommunications engineering from the Hanoi University of Science and Technology in 2013 and the Ph.D. degree in telecommunications engineering from Inje University in 2017. He is currently an Assistant Professor in networks and distributed systems at the School of Computer Science and Statistics, Trinity College Dublin, Ireland. He has special research interests in the areas of wireless AI, edge computing, the Internet of Things, and distributed learning. He was a recipient of the IEEE TVT Top Reviewer Award in 2020, the Golden Globe Award in Science and Technology for Vietnam's Young Researchers in 2021, the IEEE ATC Best Paper Award in 2022, the IEEE MCE Best Paper Award in 2023, the IEEE M-COMSTD Exemplary Editor Award in 2024, and the Clarivate Highly Cited Researcher Award in 2024. He was honored with the IEEE ComSoc Best Young Researcher Award for EMEA 2023 in recognition of his research activities for the benefit of the Society. He was the Lead Guest Editor for the Special Issue on Aerial Computing for the Internet of Things of IEEE Internet of Things Journal. He currently serves as an Editor of IEEE Communications Letters, IEEE Communications Standards Magazine, IEEE Communications Surveys and Tutorials, *Journal of Network and Computer Applications*, IEEE Transactions on Mobile Computing, and IEEE Transactions on Network Science and Engineering.